Yang–Mills and the mass gap

Volume 1 Preparation

Kiichi, Shiori and Rin · Edition of 2026-11-14 · https://computoergosum.com/en/principia/yang-mills/

> This PDF is a print of the chapters of the web edition. Interactive figures are frozen in the state in which they were printed. Links between chapters lead to the web edition, which holds the latest text.

Chapter		Page
Part 0 Entrance — how to read this book		
0-01	How to read this book	5
0-02	One picture	21
0-03	Reading the problem statement word by word	34
0-04	Four dimensions in mathematics and four dimensions in physics	49
Part 1 Mathematical preparation		
1-01	Vector spaces and linear maps	63
1-02	What a group is	74
1-03	Lie groups and Lie algebras	87
1-04	SU(2) and SU(3)	100
1-05	Manifolds and tangent spaces	113
1-06	Differential forms and the exterior derivative	124
1-07	Vector bundles and connections	137
1-08	Probability and measure	150
1-09	Probability in infinite dimensions	162

Chapter		Page
1-10	Hilbert spaces and self-adjoint operators	175
1-11	Fourier analysis and distributions	187
1-12	Detour: what happens only in four dimensions	200
Part 2 Physical preparation		
2-01	Lagrangians and the action	217
2-02	The idea of a field	228
2-03	Electromagnetism is a gauge theory	239
2-04	Special relativity and spacetime	250
2-05	The bones of quantum mechanics	262
2-06	The harmonic oscillator and creation and annihilation	273
2-07	The idea of the path integral	284
2-08	Statistical mechanics and phase transitions	296
2-09	Detour: the gap and the correlation length in the Ising model	309
2-10	The relaxation time and the dynamical exponent z	320
Part 3 Yang–Mills theory (classical)		
3-01	The gauge principle	335
3-02	Non-abelian gauge fields	344
3-03	The action and the equations	357
3-04	Gauge theory as geometry	369
3-05	Instantons and topology	380
3-06	Detour: Donaldson and exotic R^4	392
3-07	Yang–Mills inside the Standard Model	403

Chapter		Page
Part 4 Quantisation		
4-01	Canonical quantisation and the Hamiltonian	415
4-02	The path integral and Euclideanisation	426
4-03	Perturbation theory and Feynman diagrams	438
4-04	Divergences and renormalisation	448
4-05	The algebra of divergences	459
4-06	Asymptotic freedom	471
4-07	Dimensional transmutation and Λ	481
4-08	Gauge fixing and Faddeev–Popov	493
4-09	The axioms of quantum field theory	505
4-10	The Osterwalder–Schrader axioms and reconstruction	518
4-11	The definition of the mass gap	530
4-12	Detour: the perturbation series diverges	543
Part 5 Lattice gauge theory		
5-01	Wilson's lattice	555
5-02	The strong-coupling expansion	566
5-03	Reflection positivity and the transfer matrix	578
5-04	The infinite-volume limit and the cluster expansion	590
5-05	Weak coupling and the continuum limit	602
5-06	How U(1) differs from the non-abelian case	613
5-07	The Monte Carlo method	627
5-08	Measuring glueballs and the string tension	639
5-09	What to hold fixed when taking the limit	652

Chapter		Page
5-10	Lattice gauge theory with finite groups	667
5-11	Detour: Bakry–Émery and one-link integration	678

How to read this book — read the strength of a claim from its label, choose an order from the map of prerequisites

> This chapter is a guide (how to read this book).

What this chapter shows — How the book is put together (three volumes, twenty-nine parts and the appendices, 264 chapters) and the form shared by its chapters.
What the five labels (Lean, Paper, Computed, Known, Physics) guarantee and do not guarantee. How they relate to the labels of Volumes 2 and 3.
Which reader starts where. The map of prerequisites, why detours may be skipped, and the reading roads.
What this book claims and what it does not.

Prerequisite chapters — None. The overall picture is in the next chapter, 0-02, and a reading of the problem statement is in 0-03.

Said first — This book does not claim to solve the problem of Yang–Mills existence and the mass gap. Existence was shown by the construction published on 2026-11-10 (the committee's assessment is that it is proved in Lean; Volume 2 reads it; 6-00). The mass gap is open (6-04, 16-02). This book's own results are statements at a fixed lattice spacing, and conditional statements that place assumptions (16-04, § 08). The numbers in § 05– § 07 were computed from the "Prerequisites" column of the contents of 2026-10-15 (166 chapters), and do not include the 98 chapters added after that (most of Volume 2, and Parts 19–22 and 28 of Volume 3, among others).

Order of this chapter

1. What the book is — the one sentence of the problem and the roles of the parts
2. Readers and prerequisite knowledge — from second-year knowledge
3. The key to the labels — what they guarantee and do not guarantee (Table 1)
4. The form of a chapter — the three opening lines, numbering, notation, moving figures
5. The map of prerequisites — how many chapters a given chapter needs (Figure 1, Table 2)

6. How to skip detours — skipping them leaves no prerequisite missing (Detour: how many orderings do not break the prerequisites)
7. Suggested reading orders — entrances for different readers, and roads built from the map of prerequisites (Figure 2, Table 3)
8. What this book claims, and what it does not — the published construction, the open questions, this book's results

---

01

What the book is — the one sentence of the problem and the roles of the parts

This book is for reading the Clay Mathematics Institute's prize problem "Yang–Mills existence and mass gap", from second-year university knowledge up to where research stands. The prize problems were announced on 24 May 2000 at the Collège de France in Paris Known , and the core of the official problem statement (the description by Arthur Jaffe and Edward Witten) is the following sentence Known .

> Prove that for any compact simple gauge group G, a non-trivial quantum Yang–Mills theory exists on $\mathbb{R}^4$ and has a mass gap $\Delta > 0$.

The mass gap means that the spectrum $\sigma(H)$ of the theory's Hamiltonian H (the energy operator) is empty just above the vacuum energy 0.

$$\sigma(H) \subset \{0\} \cup [\Delta, \infty), \qquad \Delta > 0 \quad (1)$$

Example. The way to (1) is how the volumes are divided. **Volume 1 "Preparation"** (Parts 0–5) gathers the tools. Parts 1 and 2 are the mathematical and physical tools, Part 3 is the classical theory, Part 4 is quantisation and the exact definition of (1) (4-11), and Part 5 is lattice gauge theory. **Volume 2 "Existence — reading the published construction"** (Parts 6–15) reads the construction of four-dimensional Yang–Mills existence published in 2026-11 and the committee's assessment. It starts from what was proved and what was not shown (Part 6, 6-00 onwards), and goes on to the background in constructive field theory (Part 7), objects and the order of limits (Part 8), one step of renormalisation (Part 9), matching the coupling and asymptotic freedom (Part 10), the multi-step estimates and large fields (Part 11), infinite volume and reconstruction (Part 12), the evidence that it is Yang–Mills (Part 13), and the comparison with this book's study (Part 14). Part 15, at the end of the volume, is the record of this book's study attempting ultraviolet control (15-00 onwards). **Volume 3 "Toward the mass gap"** (Parts 16–28 and the appendices) deals with the side that is

open: where things stand and the questions that remain (Part 16, 16-01 onwards), the view from the physical side (Part 17), the dictionary between physical and mathematical vocabulary (Part 18, 18-01 onwards), objects on the lattice and firm theorems (Parts 19–22, 19-00 onwards), the first road, with the diffusive-relaxation-time assumption as its axis (Parts 23 and 24, 23-00 onwards), the second road and the crossover band (Parts 25 and 26), conditional statements (Part 27), the tree of premises (Part 28, 28-00 onwards), and the appendices, such as the table of notation, the glossary and the guide to the literature (A-01 onwards). A word-by-word reading is in 0-03, and why four dimensions is special is in the existing articles What happens only in four dimensions and One sign is different, and in 0-04.

02

Readers and prerequisite knowledge — from second-year knowledge

The intended reader is a second-year student of mathematics or physics. What may be taken as known is multivariable calculus, the basics of linear algebra, complex numbers, ordinary differential equations, introductory mechanics and electromagnetism, and elementary probability. Other tools are prepared in Parts 1 and 2, and a tool not in the prerequisite chapters is defined, minimally, where it is used. If you can read the following three formulas, you can start from Part 1.

$$\int_{-\infty}^{\infty} e^{-x^2/2}\,dx = \sqrt{2\pi}, \qquad e^{i\theta} = \cos\theta + i\sin\theta, \qquad \det\begin{pmatrix} -\lambda & 1 \\ 1 & -\lambda \end{pmatrix} = \lambda^2 - 1 \tag{2}$$

Example. The Gaussian integral becomes the Gaussian measure that represents the free field (1-08), and $e^{i\theta}$, as an element of the group $U(1)$, becomes a gauge transformation of electromagnetism (2-03). The formula on the right finds the eigenvalues of the matrix $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, and the eigenvalues are ± 1. If this matrix is read as a Hamiltonian, the difference 2 between the two eigenvalues is the energy gap of a two-level system, the simplest example of a gap (2-05). We recommend that readers from mathematics read Part 2 carefully, and that readers from physics read the second half of Part 1 (1-07–1-11) carefully.

03

The key to the labels — what they guarantee and do not guarantee

This book shows the strength of each claim, sentence by sentence, with a **label**. It does not mix "proved", "conjectured" and "supported numerically" in one sentence, and puts a label on any sentence that might mix them.

Table 1 The five labels.

Label	Meaning	What it does not guarantee	Layer where it mostly appears
Lean	A statement machine-checked in Lean 4, given with its theorem name. The number of `sorry`, the marker for an omitted proof, is 0. `native_decide` (a shortcut that hands a computation to compiled code outside the checker) is not used. The axioms are only Lean's three standard ones (propositional extensionality `propext`, the axiom of choice `Classical.choice`, soundness of quotients `Quot.sound`). For details see A-04 and the Lean guide	That the premises are true	This book's study (5-11, 17-12, Part 15, Volume 3)
Paper	A proof exists on paper, but machine checking has not been done	That there are no errors	This book's study
Computed	The range checked by computation on this machine	That it holds outside the range, or that it is a theorem	All layers (including checks of known formulas)
Known	Known theorems, their restatements, checks against outside literature	That it is a result of this site	Foundations, existing knowledge
Physics	What is accepted in physics but is not a mathematical theorem (numerics, semiclassics, conjectures)	That there is a proof	Established knowledge (Part 17), the map

In Parts 16–28 of Volume 3 and in Part 15, these five labels are refined into five classes suited to the material (Proved here, Cited, Computer-assisted, Reading, Assumption) (23-00 § 01). Roughly, Paper → Proved here or Assumption, Computed → Computer-assisted (interval arithmetic) or Reading (floating point, Monte Carlo), Known → Cited, Physics → Reading or Assumption; Lean is a column outside the labels, in which the theorem name and the assumptions are written on the same row. The "View" label of Part 24 is at the same level as "Reading" (a physical reading of the situation, corresponding to "Physics" among the five labels). The home and the label of each tool are in Table 1 of 16-10.

The labels of Volume 2. Parts 6–14 of Volume 2 use three labels for reading the published construction: Source (the source of the published construction was read), Inferred (this book's inference) and Assessment (according to the committee's assessment). Their meaning, and what each does not guarantee, are in "Labels" in 6-00.

Besides these, the text of Volume 2 also uses the same five classes as Volume 3 (Proved here, Cited, Computer-assisted, Reading, Assumption) and the Lean label, with the same meaning. Volume 2 cites every result of the published construction at the level "the committee's assessment is that it is proved in Lean". This book has neither built nor run that code (6-00 § 02). The Lean label in Table 1 above is attached only to statements that this book's study has machine-checked itself.

The label most easily misunderstood is Lean. What machine checking guarantees is that the conclusion follows logically from the premises, that is, the arrow.

$$P_1 \wedge P_2 \wedge \cdots \wedge P_k \implies Q \tag{3}$$

The truth of the premises $P_1, \ldots, P_k$ is outside the label. If the premises contain measured values or a provisionally chosen form of a function, the conclusion is conditional too, and this book states those premises alongside (23-13). Conversely, even a value re-derived on this machine gets the Known label if the value itself is known (the one-loop coefficient; 4-06).

Example. In this site's paper (5-11), the labels divide stage by stage within a single paper. The chain from the lattice action to a certain lower bound (an inequality containing the constant 27) is Lean , and the theorem names are given in 5-11. The stage that goes from there to a mass gap on the lattice is Paper , and only a sketch at that. That the constant 27 cannot be made smaller than 21 has been checked by exact arithmetic on specific configurations Computed , and is not a Lean theorem. Terms such as what the lower bound means are defined in 5-11. Even within one paper, the label differs from sentence to sentence.

04

The form of a chapter — the three opening lines, numbering, notation, moving figures

Every chapter goes in this order: the three opening lines (what it shows, prerequisites, what is said first), the order of the chapter, numbered sections, and the two closing tables (what it can say, sources and reproduction). Within a section the order is words, formulas, examples. Formulas, tables and figures are numbered afresh in each chapter, and symbols are defined at first appearance and kept consistent with the table of notation (A-01). For example, the action is written this way in every chapter.

$$S[A] = \frac{1}{2g^2} \int_{\mathbb{R}^4} \operatorname{tr}\left(F_{\mu\nu} F_{\mu\nu}\right) d^4x \tag{4}$$

A is the gauge field, $F_{\mu\nu}$ its strength, g the coupling constant, and an index appearing twice is summed from 1 to 4 (3-03). (4) is the physicists' way of writing (the convention

of writing A and F as Hermitian matrices). It is related to the curvature F_{geom} of the geometers' way of writing by $F_{\text{geom}} = -iF$, and written that way the sign of $\operatorname{tr} F^2$ changes (the two ways of writing are in A-01). From Part 3 on we take $\hbar = c = 1$, and say so at the top of the chapter.

Example. Two meanings of the same letter are also stated at the top. β is the inverse coupling $\beta = 2N/g^2$ on the lattice (5-01), and the inverse temperature $\beta = 1/k_B T$ in statistical mechanics (2-08). Moving figures are still at first and move with "Play"; without JavaScript they can be read from the table below the figure. Terms are in A-02.

05

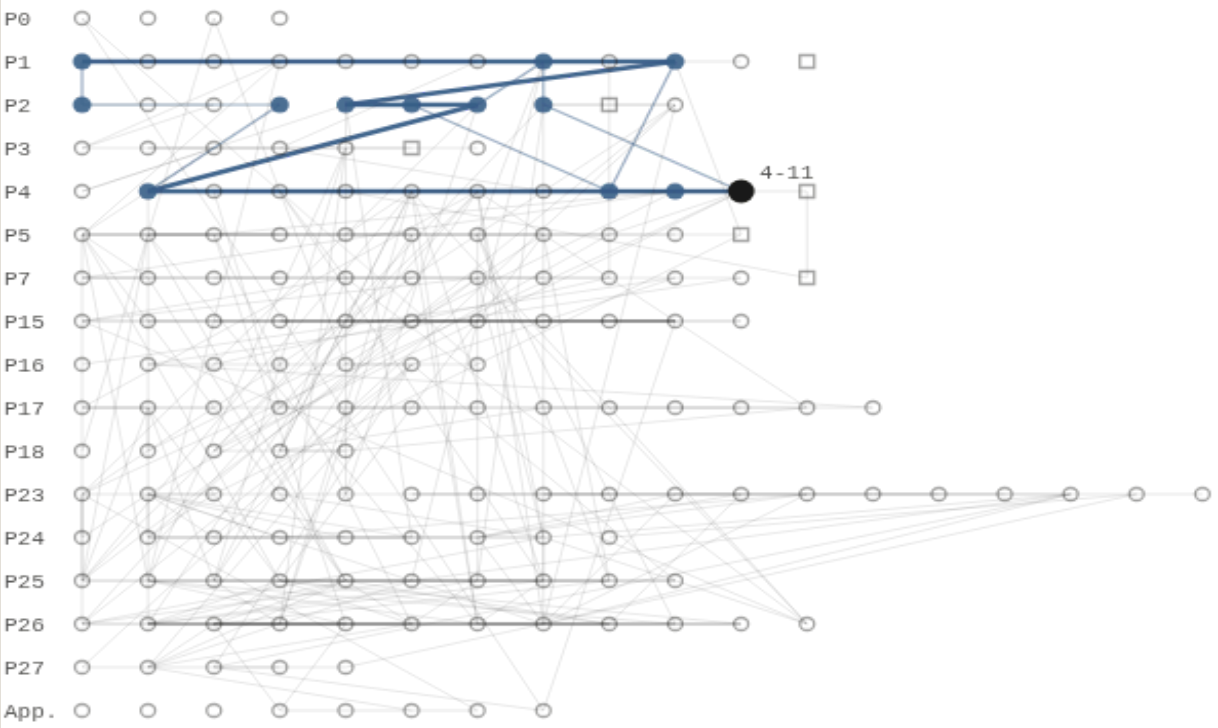
The map of prerequisites — how many chapters a given chapter needs

The "Prerequisites" column of each chapter in the contents lists the chapters to read first. Prerequisites have prerequisites of their own, so the whole set of chapters a chapter needs is the set collected by following the arrows backwards. We call this the **transitive prerequisites**. If the direct prerequisites of chapter c are $\mathrm{Pre}(c)$, the transitive prerequisites $\mathrm{Anc}(c)$ and the number of chapters in the longest chain $\ell(c)$ are determined by the following recursion (for a chapter with no prerequisites, $\varnothing$ and 1).

$$\mathrm{Anc}(c) = \bigcup_{p \in \mathrm{Pre}(c)} \big(\{p\} \cup \mathrm{Anc}(p) \big), \qquad \ell(c) = 1 + \max_{p \in \mathrm{Pre}(c)} \ell(p) \tag{5}$$

If the arrows form no cycle, this recursion terminates. Computed This map is a picture of the 166 chapters of the contents of 2026-10-15 (Volume 1, Part 7, 15-00 to 15-10, 7 chapters of Part 16, Parts 17 and 18, Parts 23–27, and the appendices). Among those 166 chapters there are 336 prerequisite arrows, and no cycle. There are 14 chapters with no prerequisites. The order of the contents of 2026-10-15 (the parts, in today's numbers, in the order 0–5, 7, 17, 18, 16, 23–27, 15, appendices) was arranged to respect the direction of the arrows: apart from 7 arrows in Parts 25–27 (26-01 → 25-06, 25-07; 26-02 → 25-08; 27-02 → 27-01; and the 3 that take the appendices A-04 and A-06 as prerequisites), for every chapter its prerequisite chapters came earlier in the contents. In today's order of the contents, renumbered so that the parts are in reading order, a further 2 arrows (17-13 → 16-05 and 26-11 → 15-00) have the prerequisite coming later Computed . So if you read straight through in today's order of the contents, no prerequisite is missing, provided you turn ahead once to the far end of those 9 arrows. This map does not include the 98 chapters added after that (Parts 6 and 8–14 of Volume 2, 15-11 to 15-14, 16-02 to 16-04 and 16-11, Parts 19–22, 26-12, Part 28, and the correspondence table of notation 24-09). Their prerequisites are in "Prerequisite chapters" at the top of each chapter.

Figure 1 The map of prerequisites of the 166 chapters of the contents of 2026-10-15 (the range is as in the text). Choosing a chapter with the slider colours its transitive prerequisites and draws the longest chain in a thick line. Rows are parts (the sixteen parts in this map, in the order of today's numbers; the last row is the appendices), columns are the order of the contents within a part, and squares are detour chapters. "Play" steps through the chapters in turn




```
4-11 The definition of the mass gap | direct prerequisites 4-10, 2-08 | transitive prerequisites
12 chapters | longest chain 7 chapters: 1-01 → 1-10 → 2-05 → 2-07 → 4-02 → 4-10 → 4-11 | chapters
that have this chapter as a prerequisite 67
```

Computed The numbers in this figure were computed on this machine. The figure follows (5) on the spot from the table of the "Prerequisites" column, and agrees with Table 2 (contents of 2026-10-15).

Example. Computed Table 2 is (5) computed for some chapters. The longest chain is 29 chapters, and leads to "A map to the mass gap" in the appendices (A-08) (the 28-chapter chain up to "What cannot be said, the index of questions" in Part 15, 15-09, with A-08 added). It passes through Parts 1, 2, 4 and 5 into Part 23, goes from the integrated table (23-12) and what remains (23-15) via 24-07 to the summary of ultraviolet control (26-11), and runs through Part 15 in order from 15-00 to 15-06. Up to Part 27 the longest is 26 chapters, leading to 26-10 and 27-01; up to Part 24 it is 19 chapters, leading to 24-07. The longest within Part 23 is 18 chapters, leading to 23-15 and 23-14; in Part 16 it is 14 chapters, leading to the list of open tasks (16-07).

Computed **Table 2** Prerequisites of representative chapters.

Chapter	Direct prerequisites	Transitive prerequisites	Longest chain
4-11 The definition of the gap	4-10, 2-08	12	7
5-09 What to hold fixed	5-05, 4-11	26	10

Chapter	Direct prerequisites	Transitive prerequisites	Longest chain
7-10 Where four dimensions stops	7-07	16	10
17-10 Back to four dimensions	17-09	13	11
16-07 The list of open tasks	16-05	25	14
23-12 The integrated table and the dependency graph	23-06, 23-07, 23-09, 23-11	38	17
23-15 What remains and its cost	23-12	39	18
24-07 Because it is four dimensions	24-01, 24-02, 23-03, 23-15, 4-07, 7-07	49	19
26-10 Candidate new viewpoints	26-06, 26-08, 26-09, 25-03, 26-01	86	26
27-01 The conditional statement	16-09, 27-00, 25-04, 25-05, 26-02, 26-05, A-06, 27-02	88	26
15-09 What cannot be said, the index of questions, the dependency diagram	15-00, 15-01, 15-02, 15-03, 15-04, 15-05, 15-06, 15-07, 15-08	60	28
15-10 Systematic attempts	15-05, 15-08	59	27

Counting in reverse shows the foundational chapters. Computed 1-01 (linear algebra) is a transitive prerequisite of 145 chapters, 2-01 (the action) of 132, and 1-10 (Hilbert spaces) of 119. On the other hand, the Prerequisites column lists only chapters needed as tools. Part 3 (classical Yang–Mills) is not among the 16 transitive prerequisites of 7-10. The Prerequisites column is a lower bound on the reading order, not every chapter worth reading.

06

How to skip detours — skipping them leaves no prerequisite missing

There are two kinds of detour. The 6 chapters whose titles carry "Detour" in the contents (1-12, 2-09, 3-06, 4-12, 5-11, 7-12), and subsections inside a chapter whose headings begin with "Detour:". Both are written on the understanding that you can skip them and still go on. A detour subsection inside a chapter appears with its "Detour:" title in the "Order of this chapter" at the top of that chapter, so you can recognise it there and skip it. If the set of detour chapters is Y, then Computed the main-line chapters whose transitive prerequisites contain a detour are only 2-10, which has

the two-dimensional Ising model (2-09) as a prerequisite, the chapters of Parts 23–27 that have 2-10 as a prerequisite, and 7-12.

$$\mathrm{Anc}(c) \cap Y \subseteq \{\text{2-09}, \text{4-12}, \text{5-11}\}, \qquad \text{2-09} \in \mathrm{Anc}(c) \iff \text{2-10} \in \mathrm{Anc}(c) \cup \{c\}, \quad \text{4-12} \in \mathrm{An}$$

Example. 1-12, 3-06 and 7-12 are prerequisites of no chapter. 4-12 is a direct prerequisite only of 7-12, and 5-11 only of 23-02, so you read 5-11 first only when going on to 23-02. 2-09 is a prerequisite of 2-10 (the relaxation time and the dynamical exponent), and many chapters of Parts 23 to 27 have 2-10 as a prerequisite, so when going on to Part 23 read 2-09 first as well.

Detour: how many orderings do not break the prerequisites

A subsection that may be skipped.

An ordering that respects the prerequisite arrows is called a **linear extension** of the partial order. The order of the contents is one of them. The number of ways to merge two chains of lengths a and b while keeping their orders is $\binom{a+b}{a}$, given by which of the $a+b$ seats go to the first chain Known . Looking at Part 1 with only the arrows inside the part, branches come out of 1-01 (the chain 1-02 → 1-04 and 1-05 → 1-06 join at 1-07, and separately 1-10 → 1-11), and apart from this there are 1-08 → 1-09 and 1-12. Merging the three lumps (9, 2 and 1 chapters), merging the two chains before 1-07, and merging 1-10 → 1-11 into those 6 chapters gives

$$e = \frac{12!}{9!\,2!\,1!} \cdot \binom{5}{2} \cdot \binom{8}{2} = 660 \cdot 10 \cdot 28 = 184800 \quad (7)$$

ways. **Example.** Computed Counting by dynamic programming over subsets, Part 1 agrees with (7), about 3.9×10^{-4} of 12!. Part 2 has 7560 ways and Part 3 has 20. In Parts 2 and 3 every chapter has at most one prerequisite within the part, so each is a **forest**, and the forest formula $e = n!/\prod_c h(c)$ ($h(c)$ is the number of c together with the chapters in the part that have c as a transitive prerequisite) Known can be used. The values of this formula also agreed with the dynamic programming. Part 1 is not a forest, because 1-07 has two prerequisites, and so it is counted with (7). The $\binom{a+b}{a}$ for two chains is the case of the forest formula where h becomes $1, \ldots, a$ and $1, \ldots, b$. For the 21 chapters of Parts 1 and 2 there are 44437771778400 ways, about 8.7×10^{-7} of 21!.

07

Suggested reading orders — entrances for different readers, and roads built from the map of prerequisites

Entrances for different readers. This book is arranged to be read through the three volumes in order, but the entrance can be changed according to purpose.

- **Readers who only want to know the problem and where things stand** — 0-02 → 0-03 → 6-00 → 6-01 → 6-04 → 16-02 → 16-04. This goes as far as what has been shown, what is open, and where this book stands.
- **Readers who want to read the published construction as mathematics (mathematics students)** — First read 1-08 to 1-11, 4-09 to 4-11, 5-01, 5-03 and 5-05 of Volume 1, and 7-01, 7-05 and 7-07 on constructive field theory; after that, read Volume 2 in order from 6-00.
- **Readers who want to enter from the physical side (physics students)** — From Parts 3 and 4 of Volume 1 and 5-01 and 5-08, to the view from the physical side (17-01 onwards) and the dictionary (18-01 onwards). After that, 6-01 and 13-01 onwards (the evidence that it is Yang–Mills).
- **Readers who want to read this book's results toward the mass gap** — After 16-01 to 16-04, the road "Reading from the firm places" below, and then the conditional statements (27-00 onwards).
- **Readers who want to follow the roads of this book's study** — The first road (23-00 onwards), the second road (25-00 onwards), the crossover band (26-00 onwards). The attempt at ultraviolet control is Part 15 (15-00 onwards), and the comparison with the published construction is Part 14 (14-01 onwards).

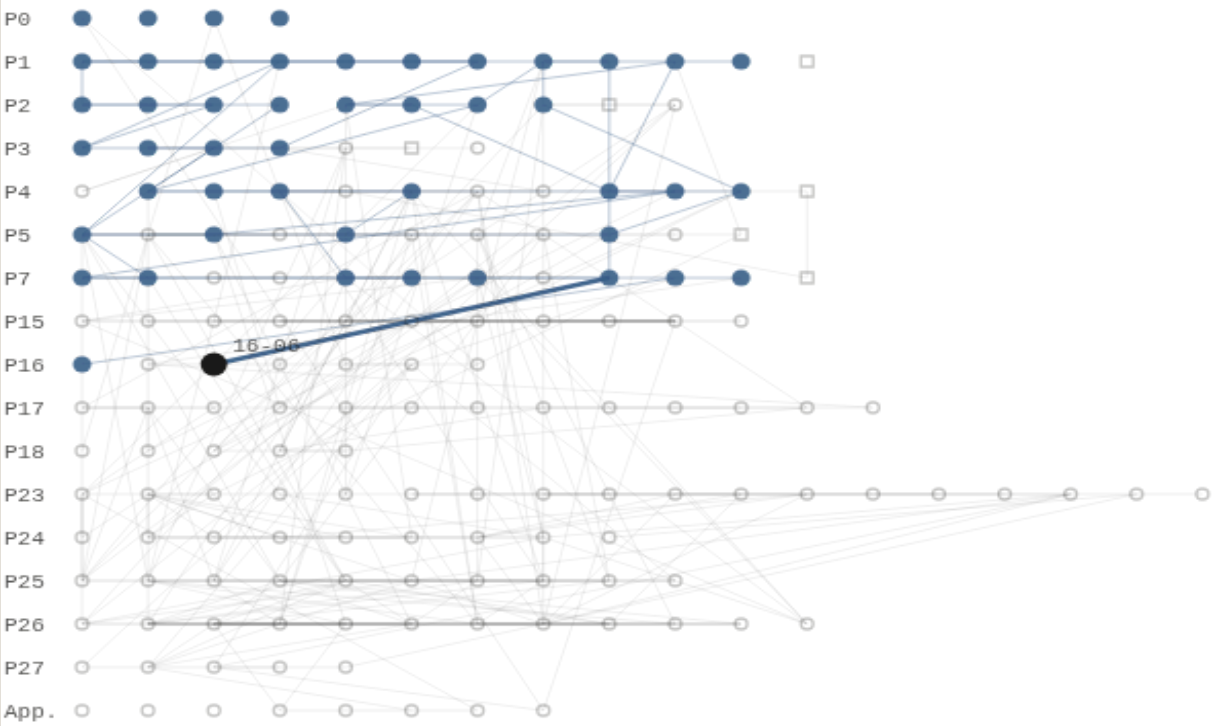
The roads below were built from the map of prerequisites of § 05 (the 166 chapters of 2026-10-15).

To read straight through, the order of the contents is enough. If you are in a hurry, add the transitive prerequisites to the set G of chapters you aim for, and arrange them in the order of the contents. If this sequence contains none of the 9 arrows of § 05 (those whose prerequisite comes later in the contents), no prerequisite is missing.

$$R(G) = G \cup \bigcup_{g \in G} \mathrm{Anc}(g) \qquad \text{(arranged in the order of the contents)} \tag{8}$$

There are five suggestions (Figure 2 shows the first three). The three roads into Parts 25–27 (following Part 25, following the band, from the first road to the band) and the three threads of Part 15 (ultraviolet control) are in "Reading orders" in the contents, and the latter are also in the last three rows of Table 3. **Entrance only** is 7 chapters chosen from the chapters whose prerequisites are empty or only Part 0 (16 chapters including the appendices Computed). **From the mathematical side** goes through the axioms, the lattice and constructive field theory to where four dimensions stops, and **from the physical side** goes through asymptotic freedom, confinement and holography to where physics parts from mathematics. Every road is the sequence built by (8) from the G of Table 3.

Figure 2 Choose a road with the buttons, and advance one chapter at a time with the slider or "Play". Chapters read are coloured, and the prerequisite arrows of the chapter being read are thickened. If a chapter comes before its prerequisite, it is counted as a "missing prerequisite"




```
From the mathematical side 48/48: 16-06 Candidates for new concepts | prerequisite 7-09 (chapter 44) | missing prerequisites on this road 0
```

Computed The numbers in this figure were computed on this machine. The figure builds (8) on the spot from the G of Table 3, and agrees with the sequences from python3.

Example. Computed None of the three sequences has a missing prerequisite. From the mathematical side is 48 chapters, from the physical side 39, and 24 are shared. The two roads meet at 16-05 (where the two ways of digging part), and whichever one you come from, you cannot go on without adding chapters of the other road. 16-05 has 24 transitive prerequisites, so it is not on the entrance-only road.

Computed **Table 3** Reading orders. On every road the 4 chapters of Part 0 are added to G. Figure 2 shows the top three. Rows four to eight are the four roads chosen by the role of the chapters (layers) and the road from the tools, and are not in Figure 2. The last five rows are roads into Part 23, Part 24 and Part 15, and are not in Figure 2. The number of chapters includes the transitive prerequisites of the target chapters.

Road	Target chapters G	Chapters	Chapters to add when going on to 16-05
Entrance only	1-12, 16-08, 23-16	7	— (next, to one of the two roads)
From the mathematical side	1-11, 3-04, 4-11, 5-03, 5-09, 7-02, 7-06, 7-09, 7-11, 16-01, 16-06	48	17-04, 17-05, 17-08, 17-09, 17-10, 17-13
From the physical side	3-05, 3-07, 4-07, 5-08, 17-02, 17-03, 17-10, 17-11, 17-13	39	2-06, 4-09, 4-10, 7-01, 7-05, 7-07, 7-10, 7-11

Road	Target chapters G	Chapters	Chapters to add when going on to 16-05
Layer: Foundations	0-02, 0-03, 1-01, 1-02, 1-04, 1-08, 1-10, 2-05, 2-07, 2-08, 3-01, 3-02, 4-02, 4-10, 4-11	24	—
Layer: Existing knowledge	4-06, 4-07, 5-03, 5-04, 5-06, 7-05, 7-07, 7-10, 17-02, 16-01, 24-07, 15-01, 15-02 (sections: 23-08 § 01, 23-08 § 03, 24-05 § 04, 24-06 § 02, 26-04 § 01, 26-08 § 03; not counted)	65	—
Layer: This book's study	5-11, 17-12, 16-09, 16-10, 23-00, 23-01, 23-05, 23-15, 24-04, 25-00, 25-04, 25-05, 26-00, 26-02, 27-01, 15-00, 15-03	101	—
Layer: Map	0-02, 16-05, 16-07, 16-09, 23-15, 25-09, 26-08, 27-02, 27-03, 27-04, 15-08, 15-09	104	—
Starting from the tools (16-10 § 02)	1-10 → 5-03 → 5-11 → 7-04 → 7-05 → 18-04 → 23-08 § 01–§ 03 → A-04 → 24-06 → 15-04 → 15-05	72	—
The road that follows the study in Part 23	23-00 → 23-01 → 23-03 → 23-04 → 23-05 → 23-12 → 23-15 → 23-16 (and their transitive prerequisites)	45	—
The road that follows Part 24	24-00 → 24-01 → 24-03 → 24-04 → 24-05 → 24-08 (and their transitive prerequisites)	49	—
Ultraviolet control (shortest)	26-11 → 15-00 → 15-03 → 15-09	65	— (the transitive prerequisites of 15-09 include every chapter of Part 15)
Ultraviolet control (the computational thread)	15-04 → 15-05 → 15-06	62	—
Ultraviolet control (the thread of the wall)	15-02 → 15-07 → 15-08 → 15-08 § 06 (how the small-field contribution to the coupling is split)	61	—

Adding the transitive prerequisites of 16-07 and 23-15 (what remains and its cost) to the union of the two roads gives 78 chapters, and in the order of the contents of 2026-10-15 nothing is missing. In today's order of the contents, read 17-13, a prerequisite of 16-05, first. In the four rows of Table 3 that enter Part 15 (the road that enters from the tools, and the three ultraviolet rows), in today's order read 26-11 before 15-00 as well

Computed . Chapters you learned in your degree course may be skimmed on any road. **The road that follows the study in Part 23** takes the relaxation-time assumption as its axis, splits the claim of the first step into conditions, and follows in order up to what remains and what is not claimed (as in § 08, it is not a record of having shown anything).

Reading from the firm places — 19-00 → 20-00 → 21-00 → 22-00 → 28-00 (the entrances of Parts 19–22 and Part 28). In order: the objects defined on the lattice, the machine-checked theorems, and the list of premises. The results at one scale (one step of the free field, and linearisation) are in the second half of Part 15 (15-11 onwards). This road is not counted in Table 3 or Figure 2.

08

What this book claims, and what it does not — the published construction, the open questions, this book's results

This book does not claim to solve the problem of Yang–Mills existence and the mass gap. The present situation is written in three parts.

- **Existence.** On 2026-11-10, a Lean proof showing, for every simple compact group, that a relativistic quantum field theory satisfying the Wightman axioms can be made as the scaling limit of four-dimensional Wilson lattice gauge theory was published, together with the assessment of a committee that checked it. The committee assesses this construction as proved. This book has neither built nor run that code, and Volume 2 reads the result at the level "the committee's assessment is that it is proved in Lean" (6-00, 6-01).
- **What is not shown.** According to the same assessment, the following are not shown: the convergence and uniqueness of the infinite-volume limit (the limit is taken along a subsequence), the probability measure of the continuum theory, a uniform mass gap (exponential decay of correlations and a gap in the spectrum), Wilson loops and the area law, and one-particle states. The mass gap is open (6-04, 16-02).
- **This book's results.** The firm theorems of Volume 3 (Parts 20–22) are statements at a fixed lattice spacing: inside the strong-coupling window, the infinite-volume limit, a gap in lattice units and the area law; the vortex floor, which holds at every coupling; and the rarity of large plaquettes on the weak-coupling side. All of them are formalisations of results of known types. The strong-coupling results lie at the opposite end (strong coupling, lattice spacing fixed) from the remaining question (the gap in the continuum limit), and the rarity of large plaquettes is also a statement at one scale before renormalisation; none of them answers that question

(16-04). The roads of Parts 23–27 are conditional statements that place assumptions, and the premises placed are listed in the tree of premises (Part 28).

One point that is easy to mistake. In lattice gauge theory (Part 5), much has been shown with the lattice spacing a fixed, on the side where the inverse coupling β is small (strong coupling). What the problem needs is the opposite side.

$$\underbrace{\beta \text{ small (strong coupling)}}_{\text{the side shown on the lattice}} \quad \longleftrightarrow \quad \underbrace{\beta \to \infty, \;\; a \to 0}_{\text{the side the problem needs}} \qquad (9)$$

Here is one sentence of this book. **A proof of the mass gap is a proof of a lower bound.** Trial states and finite measurements give only upper bounds on the gap; a lower bound needs an argument about all states (the box "The asymmetry of upper and lower bounds" in 1-10).

To show the mass gap on the road that enters from the lattice, what has to be shown is that $\Delta = m_{\text{lat}}/a$, the mass in lattice units m_{lat} divided by a, stays positive as $a \to 0$. That the limiting theory satisfies the axioms and is non-trivial was shown by the construction published in 2026-11 (at the level of the committee's assessment; 6-01). That construction, however, says nothing about the gap (5-09, 23-01, 0-03).

Example. On the strong-coupling lattice, the area law for Wilson loops has been shown (Osterwalder–Seiler 1978) Known . A Wilson loop is an observable made by multiplying the parallel transports of the gauge field along a closed path, and the area law means that its expectation decreases exponentially in the area enclosed by the path. It is taken as a sign of quark confinement Physics (5-02). But the area law holds at strong coupling in every lattice gauge theory, including those with the abelian group $U(1)$, so on its own it is not evidence of confinement in the non-abelian theory (5-06) Known . What has been shown on the strong-coupling side about the infinite-volume limit is treated in 5-04. The paper in the example of section 03 is also a result on the strong-coupling side, and claims nothing about the continuum limit.

The part of study with the relaxation-time assumption as its axis (23-00 onwards) sets out the range checked by computation, computer assistance and Lean, together with the assumption rows, and also contains ideas not yet verified (23-17), distinguished by labels. The second half of the first road (Part 24, 24-00 onwards) gives views other than the relaxation-time assumption and new views that take the relaxation-time assumption as a premise. What is not claimed is in 23-16; how to read Lean is in the Lean guide and the Lean record; the papers are in the list of papers.

What this chapter can say

What is said	What is not said (where it is treated)
Computed What was counted from the "Prerequisites" column of the contents of 2026-10-15 (166 chapters): 336 arrows, no cycles, the order of the contents of 2026-10-15 respects the arrows (except 7 in Parts 25–27; 9 in today's order), Table 2, (6), the number of linear extensions, no missing prerequisites in the sequences of Table 3	The Prerequisites column records dependencies at the design stage and does not necessarily capture the full content of the chapters (the numbers change if the contents are revised). That the five roads are the best (they are suggestions, not an optimisation). That detours inside chapters can be skipped is a commitment, and has not been computed
Known The one sentence of the problem, and the date and place of the announcement. The area law on the strong-coupling lattice, and that it is a result on the opposite side from the continuum limit	The proof of the mass gap (unsolved; 6-04, 16-02). The construction of the continuum limit was shown by the construction published in 2026-11 (at the level of the committee's assessment; 6-00). The infinite-volume limit on the strong-coupling side (5-04)
The meaning of the labels, and that the Lean label does not guarantee the truth of the premises (a commitment of this book)	The premises and theorem names of individual Lean statements (23-13, A-04; for the paper, 5-11)

Sources and reproduction

Item	Kind	Source / tool
The one sentence of the problem	Known	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory” (the official problem statement PDF of the Clay Mathematics Institute). The sentence in § 4 was checked against the original. The translation is in 0-03
Date and place of the announcement	Known	J. Carlson, A. Jaffe, A. Wiles (eds.), *The Millennium Prize Problems*, CMI / AMS, preface (p. viii): 'The prize was announced at a meeting on May 24, 2000, at the Collège de France.'. Checked via the source table of 0-03 (the PDF of the book published by the Institute)
How the labels divide in the paper	Lean Paper Computed (by stage)	This site's paper (5-11). The labels by stage were checked against the manuscript of the paper in the review of this book's material
The area law on the strong-coupling lattice	Known	K. Osterwalder, E. Seiler, “Gauge field theories on a lattice”, Ann. Phys. 110 (1978) 440–471. Bibliographic data only (attributing the area law to this paper rests on Chatterjee § 4 in the next row)

Item	Kind	Source / tool
That the area law holds at strong coupling in every lattice gauge theory, including those with $U(1)$	Known	S. Chatterjee, “Yang–Mills for probabilists”, arXiv:1803.01950 (2018) § 4. The survey of this book's material checked the text of § 4: 'the area law holds for any lattice gauge theory at sufficiently small β. A rigorous proof was given by Osterwalder and Seiler ... this also implied that the area law at small β cannot be evidence for quark confinement'
The forest formula, $\binom{a+b}{a}$	Known	D. E. Knuth, *The Art of Computer Programming*, Vol. 3, Addison–Wesley, 1973, p. 70. Bibliographic data only (attributing the formula for linear extensions of forests to this page rests on the text of S. C. Billey, J. P. Swanson, arXiv:2010.12701 § 1.3 and § 2.3). $\binom{a+b}{a}$ is the case of two chains
Table 2, Table 3, (6), (7), the numbers in the text, the figures	Computed	Computed with python3 from the "Prerequisites" column of the contents of 2026-10-15 (separators ",", "、", "・"; the table of notation for Part 24, 24-09, is not counted as a chapter). Linear extensions by dynamic programming over subsets, cross-checked with the forest formula and (7). The figures are traced on the spot by JavaScript in the browser from the same table

Next chapter: 0-02 One picture.

One picture — the tools go furthest at short distances, and let go at long distances

In this chapter, § 01– § 05 are established knowledge, § 06 and § 07 are a map of open questions and § 08 is guide material.

What this chapter shows — The two jobs the problem asks for, and why the gap shows up neither in classical theory nor in perturbation theory.
In one picture that lines up the scales of length, how far the known tools go, and where mathematics and physics let go of each other.
Which part of this book treats which part of the picture.

Prerequisite chapters — None. Formulas are defined where they appear. A word-by-word reading of the problem statement is in 0-03, and the two four-dimensional worlds in 0-04.

Said first — This is a guide-map chapter. This book does not claim to solve the problem of Yang–Mills existence and the mass gap. Existence was shown by the construction published in 2026-11 (the committee's assessment is that it is proved in Lean; 6-00), and the mass gap is open (6-04). Figure 1 is a schematic based on a one-loop approximate formula (at the level of physics), and the boundaries and widths of the bands are choices made for explanation. Figure 2 is an example of a field without interaction.

Order of this chapter

1. The question
2. A mass invisible in classical theory (Table 1)
3. One picture: the ladder of scales — weak at short distances, strong at long distances (Figure 1, Table 2)
4. A finite box and a uniform gap (Figure 2, Table 3)
5. How far the world has come (Table 4)
6. What there is besides divergence (Table 5, Detour: if infinity were given a symbol)
7. Two ways of digging and seven seams (Table 6)
8. The map of this book

01

The question — what to construct, and what to show

On 24 May 2000, in Paris, the Clay Mathematics Institute announced seven prize problems. The official problem statement for this problem was written by Jaffe and Witten. Yang–Mills theory is a gauge theory that uses a non-abelian group, and is the framework of the description of the strong and weak forces (3-02). The core sentence is this. Known

> Prove that for any compact simple gauge group G, a non-trivial quantum Yang–Mills theory exists on $\mathbb{R}^4$ and has a mass gap $\Delta > 0$ (§ 4).

"Non-trivial" means that the theory does not reduce to a free field (a field without interaction; its probability weight is **Gaussian**, the normal distribution extended to infinitely many variables) (0-03).

There are two jobs. **Constructing** — "existence" includes establishing properties at least as strong as the axioms of quantum field theory (the Wightman axioms, the Osterwalder–Schrader axioms). The axioms are a list, in the language of mathematics, of the conditions a quantum field theory should satisfy. Wightman's are written on the side of spacetime (Minkowski space), and Osterwalder–Schrader's on the side where time is made imaginary (Euclidean space) (4-09, 4-10). **Showing** — for the energy operator H (the Hamiltonian),

$$\sigma(H) \subset \{0\} \cup [\Delta, \infty), \qquad \Delta > 0 \quad (1)$$

$\sigma(H)$ is the spectrum, which you may think of as the set of values the energy can take (1-10). 0 is the energy of the vacuum, and formula (1) says that every excitation has energy at least Δ.

Example. The energy of the harmonic oscillator, measured from the vacuum, is $0, \hbar\omega, 2\hbar\omega, \ldots$, with a gap $\hbar\omega$; the $p^2/2m$ of a free particle on a line comes arbitrarily close to 0, so there is no gap. In relativity the energy of a particle at rest is its mass ($E = mc^2$), so the gap corresponds to the mass of the lightest excitation. Known To be exact, the gap in the problem statement is defined by the spectrum, not by particles. The object is pure Yang–Mills theory without quarks, and confinement (that things carrying colour charge cannot be taken out on their own) is an extension of the problem (0-03).

02

A mass invisible in classical theory — a scale that shows up in no power

According to the problem statement, the classical Yang–Mills equations describe massless waves travelling at the speed of light (§ 1), and in four dimensions the mass gap is not visible classically (§ 6.6). Known The only constant in the action is the dimensionless coupling g, and a mass cannot be made from the classical theory. Even so, physics holds that in the quantum theory one scale is born (dimensional transmutation). In the one-loop approximation

$$\Lambda = \mu \, \exp\!\Big(-\frac{1}{2 b_0 \, g^2(\mu)}\Big), \qquad b_0 = \frac{11N}{3} \cdot \frac{1}{16\pi^2} \quad (2)$$

(μ is a reference scale, $g^2(\mu)$ the coupling there, and N is N for $SU(N)$). The gap is expected to be $\Delta = c\Lambda$ (c a number), but this is not a theorem. Physics

The function $e^{-c/x}$ ($c > 0$) vanishes faster than any power as $x \to 0$.

$$\lim_{x \to 0+} \frac{e^{-c/x}}{x^n} = 0 \qquad (n = 0, 1, 2, \dots) \quad (3)$$

So all its Taylor coefficients at $x = 0$ are 0, and perturbation theory, which expands in power series in g^2, sees Λ as 0 at every order. Known What is missing is not an infinite value but a quantity that does not appear in the series (Table 1).

Computed Table 1 Pure $SU(3)$, formula (2). $1/(2b_0) = 24\pi^2/33 \approx 7.178$

g^2	Λ/μ	$\div (g^2)^{10}$	$\div (g^2)^{20}$
0.5	5.83e−7	5.97e−4	0.611
0.3	4.06e−11	6.88e−6	1.17
0.1	6.71e−32	6.71e−22	6.71e−12

The ratio divided by the 20th power exceeds 1 at $g^2 = 0.3$, but tends to 0 as $g^2 \to 0$. Continued in 4-07, 7-11 and 4-12.

03

One picture: the ladder of scales

Put the length ℓ (logarithmic) on the horizontal axis, and the strength of the coupling at that length on the vertical axis. In four-dimensional Yang–Mills the coupling becomes weaker at shorter distances (asymptotic freedom; Gross–Wilczek, Politzer, 1973). Known At one loop, for $\alpha = g^2/4\pi$,

$$\alpha(\ell) = \frac{1}{2 b_0' \ln\!\big(1/(\ell\Lambda)\big)}, \qquad b_0' = 4\pi b_0 = \frac{11N}{12\pi}, \qquad \ell\Lambda < 1 \quad (4)$$

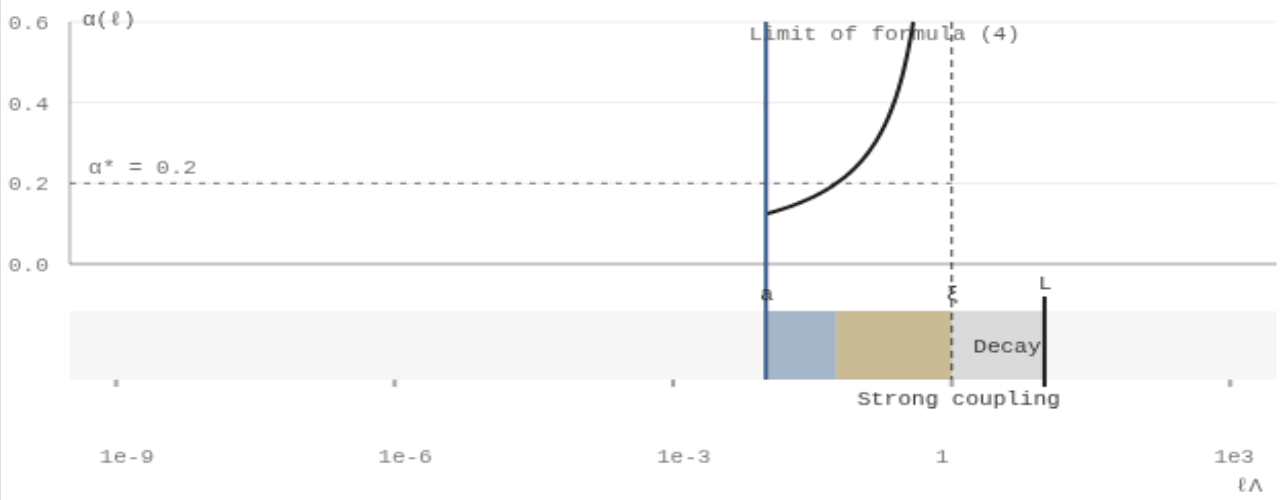
(written with g^2, this is formula (2)). Physics As $\ell \to 0$, α vanishes only at a logarithmic rate. In two and three dimensions g^2 has the dimension of mass to the power $4 - d$, and the dimensionless strength $g^2 \ell^{4-d}$ vanishes at a power rate, so four dimensions is the boundary (0-04).

A figure that sets the same $4 - d$ beside other quantities of the book and moves them with a single slider is 1-12 §10 Figure 4 (opens at d = 4).

The marker on the long side is the **correlation length** ξ. It is the length when the correlation between two distant points decreases at the rate $e^{-r/\xi}$ in the distance r, and it is related to the gap by $\xi = 1/\Delta$ (in units of $\hbar = c = 1$; 4-11).

To construct the theory, cut the short side off with a lattice spacing a and the long side with a box of side L, and take $a \to 0$ (the ultraviolet limit) and $L \to \infty$ (the infinite-volume limit). Doubling the length is counted as one step (one step of the renormalisation group; 7-05).

Figure 1 Making a smaller and L larger with the sliders lengthens the weak-coupling band on the left and the decay band on the right. The strong-coupling band in the middle does not depend on a. "Play" takes the two limits at the same time



log₁₀(aΛ) = −2.0, α(a) = 0.124, steps: weak 2.5 / strong 4.1 / ξ–L 3.3

Computed The numbers in this figure were computed on this machine (formula (4), $N = 3$). The guides $\alpha_* = 0.2$ and $\xi\Lambda = 1$ are choices made for explanation. If JavaScript is disabled, see Table 2.

Computed Table 2 Number of steps for formula (4), $N = 3$, $\alpha_* = 0.2$

$\log_{10}(a\Lambda)$	$\alpha(a)$	Weak coupling	Strong coupling
−1	0.248	0	3.3
−2	0.124	2.5	4.1
−3	0.083	5.8	4.1
−6	0.041	15.8	4.1

$\log_{10}(a\Lambda)$	$\alpha(a)$	Weak coupling	Strong coupling
−9	0.028	25.8	4.1

Each time a is divided by 10 the number of steps grows by about 3.3, and all the added steps are weak-coupling steps. The strong-coupling steps do not depend on a and are about 4.1 (with the guides set to 0.1 and 0.3 they are 8.2 and 2.7, and again do not depend on a). Making L larger adds $\log_2(L\Lambda)$ steps on the side longer than ξ. Computed

The width of the bands changes with how ξ and α_* are placed. In physics numerics, the mass m of the lightest glueball of pure $SU(3)$ (a bound state made only of gluons, that is, particles of the Yang–Mills field) is estimated as $r_0 m \approx 4.21$ (Morningstar–Peardon 1999), and Λ of the same theory, defined as $\overline{\text{MS}}$, as $r_0\Lambda \approx 0.602$ (Capitani–Lüscher–Sommer–Wittig 1999). r_0 is a ruler for length fixed from the force between static quarks. Setting $\xi = 1/m$ gives $\xi\Lambda \approx 0.14$, and ξ is expected to fall on the side shorter than $1/\Lambda$. Physics Its definition differs from the Λ of formula (4), so this is a guide, but with this value the strong-coupling band is about 1.3 steps at $\alpha_* = 0.2$ and almost 0 steps at $\alpha_* = 0.3$. Computed The numbers for the width of the bands are not physical quantities. What is robust in this picture is rather that the width does not depend on a.

The weak-coupling band is the side where there is a foothold for expansion, and it is also where mathematics has gone furthest (Bałaban; not complete; § 06, 7-07). The strong-coupling band has no small parameter and does not vanish even as $a \to 0$. In the band longer than ξ, the key is to show exponential decay of correlations in physical units, uniformly in the box (§ 04). The body of the problem is to join the three bands by estimates that do not break down in either limit. To be exact, formula (4) is a formula of perturbation theory, and this picture is a picture at the level of physics. Physics

04

A finite box and a uniform gap

What the problem statement asks for is the gap of the theory on $\mathbb{R}^4$. As a road towards it, the problem's authors consider finite-volume approximations, and name, as places where present methods do not show the way, "a mass gap uniform in the volume" and "the infinite-volume limit $T^4 \to \mathbb{R}^4$" (§ 6; T^4 is a four-dimensional box with opposite faces glued). They write that the former needs new ideas, and that if it were obtained the study of $T^4 \to \mathbb{R}^4$ would become possible (end of § 5, § 6.5). Known

With the lattice spacing a and the side of the box L fixed, that the lattice gap is positive is the easy side (the detour in 17-13). What the problem's authors see as the key is that

the gap of the continuum theory (after $a \to 0$, or uniformly in a as well) is bounded from below independently of the size of the box,

$$\inf_{L} \ \Delta_{\mathrm{phys}}(L) > 0 \quad (5)$$

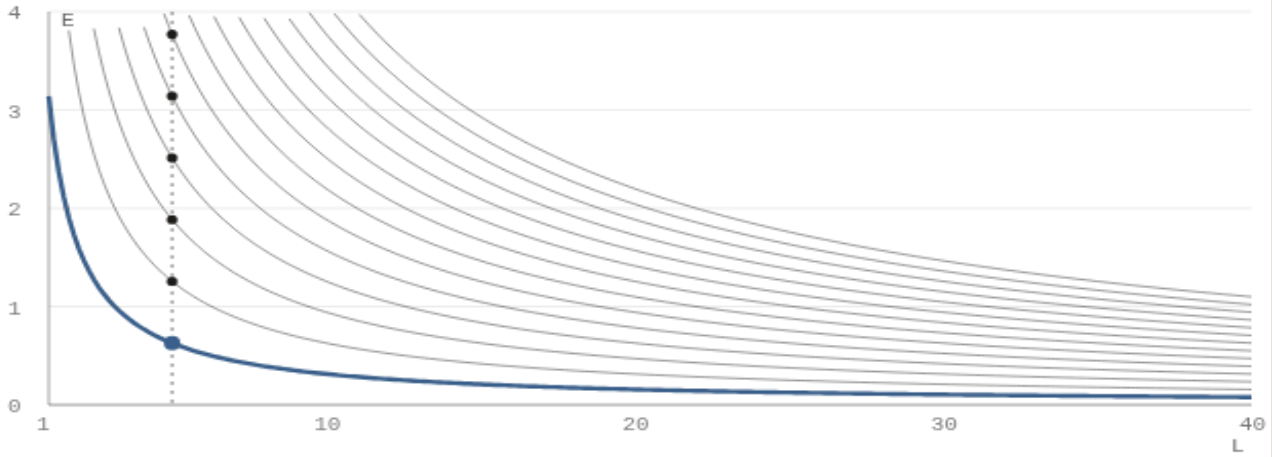
(§ 5, § 6.5). $\Delta_{\mathrm{phys}}(L)$ is the gap in a box of side L measured on a physical scale such as Λ, not measured in lattice units $1/a$. This distinction matters. If a is fixed and one measures in lattice units, exponential decay in infinite volume has already been shown on the strong-coupling side (Table 4). What is needed is that the gap measured in physical units is uniform in both $a \to 0$ and $L \to \infty$ (5-09).

Let us see the difference with a field without interaction (a free field). The one-particle levels in a box of length L with the field 0 at both ends are

$$\omega_n(L) = \sqrt{(n\pi/L)^2 + m^2}, \qquad n = 1, 2, 3, \ldots \quad (6)$$

and the gap is ω_1 ($\hbar = c = 1$). If $m = 0$, the gap π/L exists in every box but collapses as $L \to \infty$; if $m > 0$, it is always larger than m and satisfies a uniform lower bound of the form of formula (5). Known

Figure 2 Moving the box length L with the slider moves the levels (dots) along the curves ω_n(L). The buttons switch the mass. At mass 0 the gap ω_1 collapses; at mass 1 it does not come below the dashed line E = 1



m = 0, L = 5.0, gap ω_1 = 0.6283, levels with E < 1: 1

Computed The numbers in this figure were computed on this machine (formula (6)). If JavaScript is disabled, see Table 3.

Computed Table 3 The gap $\omega_1(L)$ of formula (6)

L	$m = 0$: π/L	$m = 1$	**Number of levels with $E < 1$ at $m = 0$**
1	3.1416	3.2969	0
10	0.3142	1.0482	3
100	0.0314	1.0005	31
1000	0.0031	1.0000	318

At $m = 1$ the levels with $L^2(\omega_1 - 1) \to \pi^2/2$, and at $m = 0$ those with $E < 1$, increase at the rate L/π. Computed To be exact, this m was put in by hand, and the Yang–Mills gap must be born from the interaction. The other face of the gap is the decay of correlations: formula (2) in § 5 of the problem statement says that for local quantities with vacuum expectation 0, for any positive C smaller than Δ, correlations at sufficiently large distances are bounded by e^{-Cr} (0-04, 2-09).

05

How far the world has come

The job of constructing can be divided into six stages (7-01). ① regularisation (cutting off short distances, with a lattice or otherwise), ② the finite-volume measure $\mu_{a,L}$ (the probability weight on field configurations; 1-08), ③ $a \to 0$, ④ $L \to \infty$, ⑤ the axioms, ⑥ the gap. ①–④ construct

$$\mu = \lim_{L\to\infty} \lim_{a\to 0} \mu_{a,L} \quad (7)$$

and ⑤⑥ are properties of μ.

Known Table 4 What has been reached, theory by theory

Theory	What has been reached	Literature
YM, two dimensions	The continuum measure exists (plane, closed surfaces, torus), by an exactly solvable structure. The axioms are outside the scope	Driver 1989, Gross–King–Sengupta 1989, Lévy, Chevyrev–Shen
YM, three dimensions	Up to ultraviolet estimates on the lattice uniform in the spacing. The continuum measure is not constructed even on T^3	Bałaban 1985, Chevyrev 2022
YM, four dimensions	Ultraviolet stability in finite volume (Bałaban; with assumptions, at the level of convergent subsequences), and a sketch of a continuum construction that keeps an infrared cutoff (MRS). The construction published in 2026-11 reaches ③ and ⑤, and ④ at the level of subsequences (the committee's assessment is that it is proved in Lean). The uniqueness in ④, and ⑥, are not done	Bałaban 1984–89, Magnen–Rivasseau–Sénéor (MRS) 1993, the published construction and the committee's assessment 2026 (6-00)
Lattice YM, strong coupling	With a fixed (in lattice units), infinite volume and exponential decay. The side opposite to the continuum limit	Osterwalder–Seiler 1978, Shen–Zhu–Zhu 2023
φ^4	In two and three dimensions non-trivial theories satisfying the axioms exist. In four dimensions, the scaling limit near the critical point of the single-component φ^4 with a lattice ultraviolet cutoff is Gaussian (a free field)	Problem statement § 6.2, Aizenman–Duminil-Copin 2021

Terms in Table 4: "ultraviolet stability" means that quantities do not run wild as the lattice spacing is made small, and convergent subsequences can be taken (uniqueness of the limit is not included). "Gaussian" is the type of the free field in §01. "Axioms" are the Wightman and Osterwalder–Schrader axioms of §01.

How to read Table 4. In the row for four-dimensional Yang–Mills, the construction published in 2026-11 passes ③ and ⑤, and ④ has come as far as the level of subsequences of volumes (6-01). The gap of ⑥ is untouched (6-04). The strong-coupling row shows something like ④ and ⑥, but it is about a lattice with a fixed and does not pass through ③, so it is a different proposition from formula (5) of § 04. The φ^4 row shows that there are theories whose limit becomes a free field in four dimensions. For four-dimensional Yang–Mills, the published construction made a non-trivial limit (at the level of the committee's assessment; 6-01).

The problem statement of 2000 says that not a single non-trivial relativistic field theory satisfying the axioms is known in four dimensions. Douglas's report of 2004 places what has been reached at Bałaban and MRS, and Chatterjee's 2018 survey says that a construction of a continuum limit above two dimensions in which Wilson loops (the quantity made by multiplying the rotations by the gauge field along a closed path and taking the trace; 5-01) behave non-trivially does not yet exist. Known The construction published in 2026-11 made a non-trivial theory satisfying the Wightman axioms in four dimensions (the committee's assessment is that it is proved in Lean; 6-00). Wilson loops, however, are outside that construction, and a continuum limit in Chatterjee's sense has still not been made (6-04). "Two and three dimensions can be done" is inaccurate for non-abelian Yang–Mills: in three dimensions the continuum measure has not been constructed even in finite volume (T^3) (7-02, 7-06, 16-01).

06

What there is besides divergence

"It is hard because mathematics does not allow the handling of infinities" is one side of it. Handling divergences is a matter of the left band in Figure 1, and it is the most advanced (Bałaban's programme is unfinished: its main theorem assumes that the effective coupling is small, and the theorem removing that assumption is unpublished; 7-07. The construction published in 2026-11 controlled the ultraviolet by a differently assembled renormalisation; Parts 9–11). Where the problem's authors asked for new ideas is the place in § 04 that is not about divergence. Known

Known Table 5 Obstacles that cannot be reduced to divergence (7-11)

1	No small parameter in the infrared	There is no easy solvable model to put at the centre of an expansion (MRS). The lattice strong coupling that can be handled rigorously is on the side opposite to the continuum limit
2	Invisible to perturbation theory	The scale of the gap is of the type in Table 1

3	Non-triviality	φ^4_4 (the four-dimensional φ^4; not asymptotically free) has a Gaussian limit near the critical point with a lattice cutoff (7-08). For four-dimensional Yang–Mills, the construction published in 2026-11 made a non-trivial limit (at the level of the committee's assessment; 6-01)
4	The state space is curved	The orbit space is not linear, and one cannot choose one field from each orbit continuously (a global gauge fixing) (on closed spaces such as S^4; Singer 1978; the setting is in 3-04)
5	Few regularisations preserve reflection positivity	The lattice preserves it but breaks rotational symmetry. Many continuum regularisations break it (§ 6.5)
6	The structure of infinitely many degrees of freedom	Under standard conditions, the interaction picture does not exist (Haag's theorem)

Terms in Table 5: the "orbit space" is the space made by regarding fields related by gauge transformations as the same. "Reflection positivity" is the positivity condition for recovering the Hilbert space of quantum mechanics from a measure on Euclidean space (4-10). The "interaction picture" is the way of writing time evolution by taking the state space of the free field as it is and adding the interaction to it (4-09).

How to read Table 5. 1 and 2 concern the middle and right bands of Figure 1, that is, the side where the coupling is not small. 3 is the job of showing that the limit constructed has not collapsed to a free field. 4–6 are structures carried by the framework itself of gauge theory and infinitely many degrees of freedom. Each is a different problem from "how to handle infinities", and handling divergences alone does not settle them. Known

"Because it is non-abelian" also, to be exact: for continuous non-abelian groups in three or more dimensions, exponential decay in infinite volume has been shown only on the strong-coupling side, and weak-coupling results are limited to finite groups (Adhikari–Cao; 5-06, 5-10).

Detour: if infinity were given a symbol

A subsection that may be skipped.

"Would anything change if $1/0$ were made a symbol?" As far as known cases go, it is thought likely that nothing would change. Handling divergences inside perturbation theory is already rigorous (BPHZ, Epstein–Glaser, Connes–Kreimer; renormalisability by 't Hooft–Veltman 1972 and others), and the $0/0$ that appears has been handled not by assigning a value but by rates of approach or by changing the geometry. The scale of the gap is of the type e^{-1/g^2}. This is a sorting of cases, not an impossibility theorem (7-12).

07

Two ways of digging and seven seams

The mathematical side starts from "the theory does not yet exist" and constructs in the order existence → axioms → gap (Parts 5 and 7). The physical side starts from "the theory exists" and measures glueball (§ 03) masses and the string tension (Part 17). As a metaphor, it is a tunnel dug from both ends. With the construction published in 2026-11, the mathematical side has dug as far as existence and the axioms (6-00), but at the gap it has not yet broken through. To be exact, the numbers of physics are checks of consistency on the assumption of existence (17-01, 16-05).

Table 6 The seven seams

1	The premise of existence	That numbers appear to converge is not a proof of existence	17-13
2	One scale	Physics holds that the gap and the string tension are both constant multiples of Λ. Not a theorem	17-12
3	Different shores	Shores with a foothold: perturbation theory (formal series), lattice strong coupling (theorems), holography ($N \to \infty$, $\lambda \to \infty$, at the level of physics). The theory of the problem statement is on none of the shores	17-13
4	Finite and infinite	Numbers are extrapolations. What is needed is uniform estimates	5-09
5	Spacetime and space	Construction is on the Euclidean side. The way back to spacetime is reflection positivity	0-04
6	The word "particle"	Physics counts particles; the problem statement asks for a gap in the spectrum	4-11
7	Extra dimensions	The gap turns into geometry, but the order of limits for returning to four dimensions is not worked out	17-10

Terms in Table 6: "shore" is a metaphor; exactly, it is a range of parameters where theorems or controlled approximations are available. $\lambda = g^2 N$ is the 't Hooft coupling. "Turns into geometry" is also a metaphor; exactly, in holography the gap is computed as an eigenvalue of a wave equation on a geometry that includes extra dimensions (the leading term in $1/N$; at the level of physics; 17-09).

How to read Table 6. All seven are places where a step taken for granted in physics becomes, in mathematics, a proposition that needs a proof. 1 and 4 concern the relation between numerical computation and limits, 2 and 6 what is called a "mass", 3 and 7 the distance between computable models and the theory of the problem statement, and 5 the translation between spacetime and Euclidean space. The rightmost column gives the chapter that treats each in detail.

This book's study has set to work around 2: it cut out "one scale" as the diffusive-relaxation-time assumption, but has not proved it (23-01; the correspondence with the equivalence principle is an analogy; 17-12). The remaining six are only named.

08

The map of this book

This book is in three volumes. Volume 1 "Preparation" (Parts 0–5) gathers the tools for reading the pictures of this chapter. The remaining two chapters of Part 0 read the words of the problem statement (0-03) and the two four-dimensionalities (0-04). Parts 1 and 2 are the mathematical and physical preparation, and provide the tools (spectrum, measure, path integral, statistical mechanics) for reading formulas (1) and (7) and § 04. Part 3, classical Yang–Mills theory, deals with § 02; Part 4, quantisation, with formulas (2) and (4) and the axioms; and Part 5, lattice gauge theory, with a (the lattice spacing) in Figure 1 and the strong-coupling side.

Volume 2 "Existence — reading the published construction" (Parts 6–15) reads the construction of 2026-11 that rewrote the four-dimensional row of Table 4. In order: what was proved and what was not shown (Part 6), the literature of Tables 4 and 5 against the background of constructive field theory (Part 7), the objects from the lattice to the continuum and the order of limits (Part 8), one step of renormalisation (Part 9), matching the coupling and asymptotic freedom, which is the left band of Figure 1 (Part 10), the multi-step estimates and large fields (Part 11), infinite volume and reconstruction (Part 12), the evidence that it is Yang–Mills (Part 13), and the comparison with this book's study (Part 14). Part 15, at the end of the volume, is the record of this book's study attempting ultraviolet control (15-00 onwards).

Volume 3 "Toward the mass gap" (Parts 16–28 and the appendices) deals with the side that is still open: the middle and right bands of Figure 1, and formula (5). In order: where things stand and the questions that remain (Part 16; the continuation of § 05 and Table 6), the view from the physical side (Part 17), the dictionary between physical and mathematical vocabulary (Part 18), objects on the lattice and firm theorems (Parts 19–22), the two roads of this book's study and the crossover band (Parts 23–26; including the area around 2 of Table 6), and conditional statements and the tree of premises (Parts 27 and 28). The appendices are the table of notation, the glossary, the guide to the literature, the guide to Lean, and so on. Chapters and sections marked "Detour" may be skipped. Suggested reading orders are in 0-01.

What this chapter can say

What is said	What is not said (where it is treated)
Known The relevant passages of the problem statement (checked against the PDF)	The word-by-word reading (0-03)

What is said	What is not said (where it is treated)
Known Formula (3). Computed Table 1	$\Delta = c\Lambda$ is the expectation of physics (4-07)
Computed Figure 1, Table 2 (the number of steps within formula (4))	Formulas (2) and (4) are not theorems. The boundaries and widths of the bands are choices of drawing. $\xi\Lambda \approx 0.14$ is a guide via Λ, which is defined differently (Physics)
Computed Figure 2, Table 3 (free field)	Shows nothing about the Yang–Mills gap
Known Tables 4 and 5	Not a complete survey of the literature (16-01)
—	This book shows neither the continuum limit nor the gap (the continuum limit was shown by the published construction, at the level of the committee's assessment; 6-00). The diffusive-relaxation-time assumption remains an assumption

Sources and reproduction

Item	Kind	Source / tool
The one sentence, § 1– § 6.6	Known	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory” (the official problem statement PDF of the Clay Mathematics Institute; the relevant passages were read). As a book, included in J. Carlson, A. Jaffe, A. Wiles (eds.), *The Millennium Prize Problems*, Clay Mathematics Institute / American Mathematical Society (2006) pp. 129–152 (contents checked in the PDF published by CMI, year of publication on the AMS book page)
Date and place of the announcement	Known	The Millennium Prize Problems page of the Clay Mathematics Institute website (read; 24 May 2000, Collège de France, Paris). The preface of the same book gives the same date and place (checked in the PDF published by CMI)
Asymptotic freedom	Known	D. J. Gross, F. Wilczek, Phys. Rev. Lett. 30 (1973) 1343 / H. D. Politzer, ibid. 1346. Abstracts checked (the former: a wide class of non-abelian gauge theories has the asymptotic form of a free field up to logarithmic corrections; the latter: perturbation theory becomes arbitrarily good in the deep Euclidean region)
The guide for $\xi\Lambda$ in § 03	Physics	C. Morningstar, M. Peardon, Phys. Rev. D 60 (1999) 034509 Table VII (bibliographic data only; the survey of the material checked the text) / S. Capitani, M. Lüscher, R. Sommer, H. Wittig, Nucl. Phys. B 544 (1999) 669 eq. (6.2) (the text was read). Ratios and numbers of steps by python3
b_0	Physics	Particle Data Group, Phys. Rev. D 110 (2024) 030001, QCD chapter § 9.1.1. Bibliographic data only (checked by the survey of the material)

Item	Kind	Source / tool
Table 4	Known	B. K. Driver, CMP 123 (1989) 575 / L. Gross, C. King, A. Sengupta, Ann. Phys. 194 (1989) 65 / T. Lévy, Mem. AMS 166 no. 790 (2003) / I. Chevyrev, H. Shen, arXiv:2302.12160 / T. Bałaban, CMP 102 (1985) 255, CMP 122 (1989) 355 / J. Magnen, V. Rivasseau, R. Sénéor, CMP 155 (1993) 325 / I. Chevyrev, arXiv:2202.13359 / K. Osterwalder, E. Seiler, Ann. Phys. 110 (1978) 440 / H. Shen, R. Zhu, X. Zhu, CMP 400 (2023) 805 / M. Aizenman, H. Duminil-Copin, Ann. of Math. 194 (2021) 163 / M. R. Douglas, report for the Clay Mathematics Institute (2004) / S. Chatterjee, arXiv:1803.01950. Bibliographic data only (the survey of the material checked them against the originals; CMP = Comm. Math. Phys.) / The four-dimensional row of 2026-11: yang-mills-group/OpenAI-YM-20261028 (the Lean source) and the committee's assessment, docs/assessment-1.pdf in yang-mills-group/YM (2026-11-10). How to read them, and at what level, is in 6-00
Table 5, § 06	Known	I. M. Singer, CMP 60 (1978) 7 / R. Haag, Mat.-Fys. Medd. Danske Vid. Selsk. 29 (1955) no. 12 (via secondary sources) / A. Adhikari, S. Cao, Ann. Probab. 53 (2025) 140 / G. 't Hooft, M. Veltman, Nucl. Phys. B 44 (1972) 189. Bibliographic data only
Tables 1–3, Figures 1 and 2	Computed	python3 (formulas (2), (4), (6)). The figures are drawn by the browser from the same formulas
Restatements of terms (axioms, reflection positivity, orbit space, interaction picture)	Known	Detailed definitions and sources in 4-09, 4-10 and 3-04 of this book

Next chapter: 0-03 Reading the problem statement word by word.

Reading the problem statement word by word — each word closes a loophole to an easy answer

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — Where the core sentence of the problem statement, the paragraphs that fix its words, and the two footnotes are, and what they say.
What each of the words "any compact simple group", "$\mathbb{R}^4$", "exists", "non-trivial", "mass gap" and "short distances" demands.
A list of the "easy answers" each word directly shuts out (Table 5).

Prerequisite chapters — None. Each word is made into a tool in later chapters (the destination is given at the end of each section). The overall picture is in the previous chapter, 0-02, and the difference between the two four-dimensional worlds is in the next, 0-04.

Said first — This chapter only reads the problem statement, and shows nothing about the problem. This book does not claim to solve this problem. The existence side was shown by the construction published in 2026-11 (the committee's assessment is that it is proved in Lean; 6-01 sets the words of this chapter against the construction), and the mass gap is open. The English quotations are the original text, checked against the official problem statement PDF of the Clay Mathematics Institute, and the restatements are this chapter's. Table 5 is this chapter's reading, not something the problem's authors stated in that form.

Order of this chapter

1. The core sentence — where it is written, and how it is built
2. Any compact simple group — all of them, down to the exceptional groups (Table 1; Detour: for non-compact groups)
3. On $\mathbb{R}^4$ — neither a box nor a lattice, but an infinitely wide continuous four dimensions
4. Exists — constructing something that satisfies the axioms (Table 2)
5. Non-trivial — not collapsed to a Gaussian

6. Mass gap — the gap in the spectrum and $m < \infty$ (Figure 1, Table 3)
7. Agreement at short distances — asymptotic freedom names "Yang–Mills"
8. The two footnotes — quantum operators, and existence by compactness alone (Figure 2, Table 4)
9. What is not written — extensions of the problem, QCD, formulations on the lattice
10. What each word shuts out — a word-by-word list (Table 5)

01

The core sentence — where it is written, and how it is built

On 24 May 2000, at a meeting at the Collège de France in Paris, the Clay Mathematics Institute announced seven prize problems (the Institute's website, and the preface of the book that contains the problem statements). The official problem statement for one of them is Arthur Jaffe and Edward Witten's description "Quantum Yang–Mills Theory" (14 pages) Known . Below, § is a section of the problem statement, and "section 06" is a section of this chapter. The problem is the one sentence in § 4.

> *Yang–Mills Existence and Mass Gap.* Prove that for any compact simple gauge group G, a non-trivial quantum Yang–Mills theory exists on $\mathbb{R}^4$ and has a mass gap $\Delta > 0$. Existence includes establishing axiomatic properties at least as strong as those cited in [45, 35].

Restated: **For any compact simple gauge group G, prove that a non-trivial quantum Yang–Mills theory exists on $\mathbb{R}^4$ and has a mass gap $\Delta > 0$.** Existence includes establishing properties at least as strong as the axiomatic properties cited in reference [45] (the book of Streater–Wightman) and [35] (the two papers of Osterwalder–Schrader).

Writing the skeleton as a logical formula makes the order of the quantifiers visible ($\sigma(H)$ is the spectrum of the Hamiltonian H; section 06).

$$\forall G \; \exists \mathcal{T}_G : \quad \mathcal{T}_G \text{ satisfies the axioms, is non-trivial, agrees at short distances, and has } \sigma(H) \cap$$

One theory $\mathcal{T}_G$ is needed for each group, and the gap Δ_G may differ from group to group. § 4 also imposes one more thing: that the supremum of the gap m_G is finite ($m_G < \infty$; section 06). The meaning of the words is fixed by the surrounding paragraphs. Quantum fields and the axioms are in § 3, local fields and the short-distance condition

in the first paragraph of § 4, the mass gap in the second paragraph of § 4, clustering and the extensions of the problem in § 5, and finite-volume estimates and the footnote on weak existence (footnote 2 of Jaffe–Witten) in § 6 and § 6.5.

Example. For $G = SU(3)$, what is asked for is a theory of only the 8 kinds of gauge field (gluons), with no quarks (8 is the dimension of $SU(3)$; Table 1). QCD with quarks added is not the object (section 09).

02

Any compact simple group — all of them, down to the exceptional groups

The **gauge group** is the group of symmetries by which the values of the field may be rotated independently at each point (3-01). **Compact** means bounded and closed as a group of matrices. **Simple** is meant in the sense of Lie groups: the Lie algebra (1-03) is not commutative and has no non-trivial ideal (a subspace that bracketing with any element does not take outside). $SU(2)$ has the centre $\{\pm 1\}$, so as an abstract group it is not simple. In the problem statement, § 1 names $G = SU(3)$ as the gauge group of the strong force (it has the centre $\mathbb{Z}_3$, so it is not simple as an abstract group), and § 6.6 also treats $SU(N)$, so we read it in the sense of Lie groups.

A gauge field has as many components as the dimension of the group. Choosing a basis T^a of the Lie algebra, the bracket (commutator) can be expanded in the basis as $[T^a, T^b] = i f^{abc} T^c$. These coefficients f^{abc} are the **structure constants**, the multiplication table of the Lie algebra (1-03). With them,

$$A_\mu = \sum_{a=1}^{\dim G} A_\mu^a \, T^a, \qquad \sum_{c,d} f^{acd} f^{bcd} = C_A \, \delta^{ab} \tag{2}$$

C_A is 0 for a commutative group and positive for a compact simple group Known . Its value changes with how the size of the basis is taken, so we fix a normalisation. The basis T^a is taken orthonormal for an inner product invariant under the group action, and that inner product is chosen so that the squared length of the longest root (1-04) is 1. For $SU(N)$, the standard normalisation of physics, $\operatorname{tr} T^a T^b = \tfrac{1}{2}\delta^{ab}$ in the fundamental representation, corresponds to this. In this normalisation C_A equals the **dual Coxeter number** $h^\vee$, a positive integer determined by each group, and for $SU(N)$ it is N Known . In the normalisation common in mathematics, "squared length 2", it becomes $C_A = 2h^\vee$ (checked from the structure constants for $SU(2), SU(3), SU(4)$, $SO(5), SO(7), SO(8)$ Computed). The compact simple Lie algebras are exhausted by the series A_n, B_n, C_n, D_n and the exceptional types G_2, F_4, E_6, E_7, E_8 (the Killing–Cartan classification; 1-04) Known , and "any" means all of these.

Computed **Table 1** The dimension and the positive integer $h^\vee$ determined by each group (list of values in 1-04). Every row agrees with the standard values. The procedure of computation is in the source table. b_0 is the one-loop coefficient of section 07.

Group	Type	Dimension	$C_A = h^\vee$	$b_0 = 11C_A/12\pi$
$SU(2)$	A_1	3	2	0.5836
$SU(3)$	A_2	8	3	0.8754
$Sp(3)$	C_3	21	4	1.1671
$SO(8)$	D_4	28	6	1.7507
G_2	G_2	14	4	1.1671
F_4	F_4	52	9	2.6261
E_7	E_7	133	18	5.2521
E_8	E_8	248	30	8.7535
$U(1)$ (not simple)	Commutative	1	0	0

The general form of the one-loop coefficient is $b_0 = (11C_A - 4n_f T_R)/(12\pi)$ (n_f is the number of kinds of quark, and T_R a positive number determined by the representation of the quarks; PDG 2024) Known . In a theory without quarks, $n_f = 0$ gives $b_0 = 11C_A/12\pi$, and from $C_A > 0$ it is $b_0 > 0$ for every simple group. At the level of one-loop perturbation theory, the coupling becomes weaker at shorter distances (section 07). The numerical value of b_0 depends on the normalisation of the coupling g (the normalisation of C_A above), so comparing its size between groups has no meaning. The point is that it is positive. For $U(1)$ it is $b_0 = 0$, and that theory is the free field of the massless photon. There is no gap. For a product of two simple groups the action splits into a sum, giving two theories that do not act on each other.

Detour: for non-compact groups

A subsection that may be skipped.

For the action $\frac{1}{4g^2}\int \mathrm{Tr}\, F \wedge *F$ (formula (1) of the problem statement writes it as the Lagrangian L; its content is the integrated action), the problem statement says only that Tr is "an invariant quadratic form on the Lie algebra". Requiring it to be definite is this chapter's argument, and the reason is that it is needed for the classical energy to be positive. The elements of $\mathfrak{su}(2)$ are anti-Hermitian, so $\mathrm{tr}\, X^2 \le 0$. In the Lie algebra of $SL(2,\mathbb{R})$,

$$X = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} : \ \mathrm{tr}\, X^2 = 2, \qquad Y = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} : \ \mathrm{tr}\, Y^2 = -2$$

both signs appear, and the invariant quadratic form (limited to constant multiples of $\operatorname{tr} XY$) is indefinite. Real Lie algebras with a positive-definite invariant inner product are limited to the Lie algebras of compact groups Known .

03

On $\mathbb{R}^4$ — neither a box nor a lattice, but an infinitely wide continuous four dimensions

$\mathbb{R}^4$ is the set of all quadruples of real numbers; it is not a finite box, not a periodic box (the four-dimensional torus T^4), and not a lattice. On the approximations, symmetry is lost. The lattice breaks continuous translations and rotations, and T^4 keeps translations but breaks rotations.

The sentence in § 4 does not state the signature of $\mathbb{R}^4$. § 3 introduces the Wightman axioms as a theory on "$\mathbb{R}^4$ with Minkowski signature", and § 6, in a parenthesis describing the state of constructive field theory, writes "in the Euclidean formulation, replace Euclidean space-time $\mathbb{R}^4$ by $\mathbb{R}^2$ or $\mathbb{R}^3$", from which one can read that the stage of the Euclidean version is $\mathbb{R}^4$ Known . There are two ways of measuring length.

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 \ \text{(Euclidean)}, \qquad -t^2 + x^2 + y^2 + z^2 \ \text{(Minkowski)} \ (3)$$

The two are connected by the theory of Osterwalder and Schrader (§ 3; formally, the substitution $t = -i\tau$). How this works is treated in 0-04 and 2-04 and in the existing article One sign is different.

Example. § 6 says of T^4 that "the construction is not complete, but there is good indication that known methods can be extended", and that methods for obtaining the $T^4 \to \mathbb{R}^4$ limit are not suggested by present techniques Known . This gap goes straight into the problem. Why four is special is in 1-12 and in the existing article What happens only in four dimensions.

Symmetry is not the only thing lost in approximation. On reflection positivity § 6.5 writes 'few methods exist to recover reflection positivity in case it is lost through regularization—such as with dimensional regularization, ...', giving dimensional regularisation and Pauli–Villars as examples Known . The lattice keeps reflection positivity and breaks rotations, and continuum regularisations do the opposite (4-04, 4-10, 7-11).

04

Exists — constructing something that satisfies the axioms

"Exists" does not mean that some limit exists. § 4 says to **define** a quantum field theory in the sense of § 3, with local fields corresponding to **gauge-invariant local polynomials** in the curvature F and its covariant derivatives Known . The curvature is written $F = dA + A \wedge A$ in § 1, and the local polynomial § 4 gives as an example is $\mathrm{Tr}\, F_{ij}F_{kl}(x)$.

$$F = dA + A \wedge A \ \ (\S 1), \qquad \mathrm{Tr}\, F_{ij}F_{kl}(x) \ \ (\S 4)\ (4)$$

The gauge field A itself is not asked for. A changes under gauge transformations and is not an observable. A "quantum field" is defined in § 3 as "an operator-valued distribution on space-time". The value at a point $\mathcal{O}(x)$ has no meaning; $\mathcal{O}(f) = \int \mathcal{O}(x) f(x)\, d^4x$, smeared with a test function f, is an operator (1-11).

Table 2 The properties included in "existence" (a summary of § 3 of the problem statement; the exact form is in references [45] and [35]; in this book, 4-09 and 4-10).

Property	Content
The space of states	A Hilbert space $\mathcal{H}$ carries a representation of the Poincaré group (the group combining translations and Lorentz transformations of spacetime; 2-04). H and P are the generators of time and space translations
Energy	$0 \le H$
vacuum	A vector Ω invariant under the Poincaré group, unique up to phase
Fields	Gauge-invariant functions of the field act as operators on $\mathcal{H}$ and transform covariantly
Locality	Fields in regions that cannot be joined by a light signal are independent (the operators commute)
Euclidean version	Conversion back and forth with a Euclidean-invariant theory. The key is reflection positivity

Example. The free Klein–Gordon field was the field that served as a model for the first axiom system (§ 6.2), but it is excluded by "non-trivial", next. § 6.5 states plainly that "establishing the Hilbert space of quantum mechanics is part of the solution of this Millennium problem" Known .

05

Non-trivial — not collapsed to a Gaussian

The problem statement does not define "non-trivial". It is usually read as "not a free field (a field without interaction)". § 6.2 says that the **Gaussian measure** with mean 0 and covariance $(-\nabla^2 + m_0^2)^{-1}$ (∇^2 the Laplacian) gives the free field of mass m_0, and that non-linear fields need non-Gaussian measures Known . Whether something is

Gaussian can be told from higher correlations. For a Gaussian random variable X with mean 0,

$$\frac{\mathbb{E}[X^4]}{(\mathbb{E}[X^2])^2} = 3 \quad (5)$$

and in several variables too, four-point expectations can be written as sums of products of two-point expectations (Isserlis' formula; 1-08) Known .

Example. Computed For the density $\propto e^{-x^4}$ the ratio (5) is 2.18844 (agreeing with $\Gamma(\frac{1}{4})\Gamma(\frac{5}{4})/\Gamma(\frac{3}{4})^2$). For the density $\propto e^{-x^2/2-\lambda x^4}$, with $\lambda = 1,\ 0.1,\ 0.01$ it is 2.319, 2.537, 2.843, returning to 3 as $\lambda \to 0$. **Triviality** of a field theory is the phenomenon in which each approximation is non-Gaussian, yet in the limit ratios of this kind return to the Gaussian value.

To be exact, the one-variable ratio is a guide. For four-dimensional $\lambda\varphi^4$, it is a theorem that if the ultraviolet is cut off by a lattice and the scaling limit near the critical point is taken with lattice spacing 0 and infinite volume, the limit is Gaussian (Aizenman–Duminil-Copin 2021; 1-08) Known . Yang–Mills differs in that its coupling becomes weak at short distances, and a non-trivial limit has been expected Physics . In four dimensions, the construction published in 2026-11 made a non-trivial limit (the committee's assessment is that it is proved in Lean; 6-01). A free field with mass has a gap (Table 3), so the gap alone cannot shut out free fields (7-08). § 6 says that at present not a single non-trivial relativistic field theory satisfying the Wightman (or other reasonable) axioms is known in four dimensions (as written in 2000) Known . The committee assesses that the construction published in 2026-11 made such a theory for four-dimensional Yang–Mills (6-00).

06

Mass gap — the gap in the spectrum and $m < \infty$

The definition in § 4 is written in terms of the spectrum $\sigma(H)$ of the Hamiltonian H (what corresponds to the set of eigenvalues for a matrix; 1-10). The vacuum is an eigenvector with energy 0, and the energy is never negative. On top of that, having an empty interval just above 0 is called a gap.

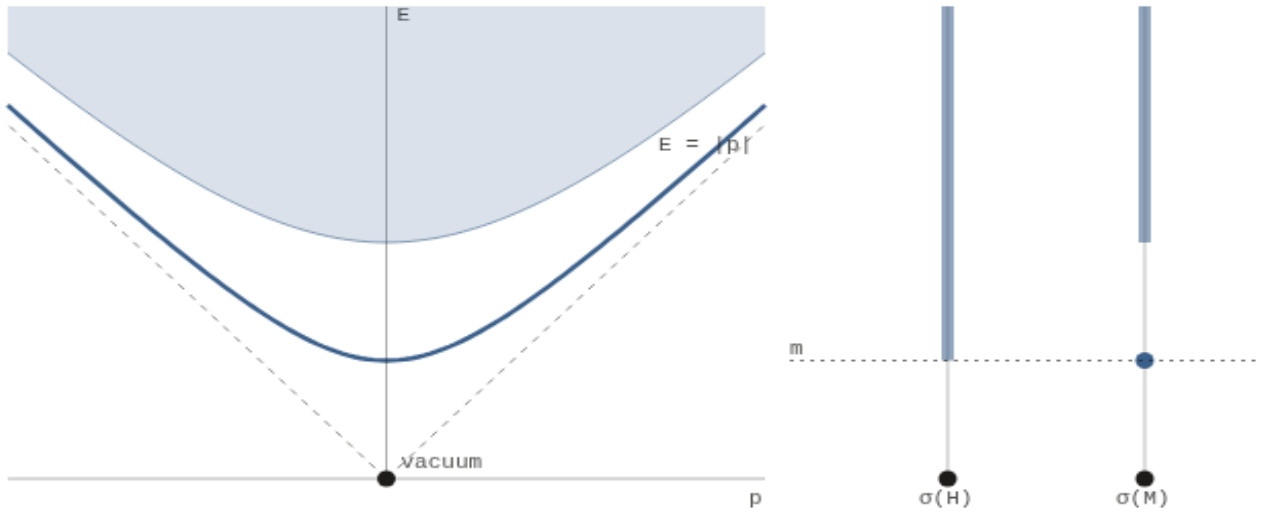
$$H\Omega = 0, \qquad \sigma(H) \subset [0, \infty), \qquad \sigma(H) \cap (0, \Delta) = \varnothing, \qquad m = \sup \Delta < \infty \quad (6)$$

$\sup \Delta$ is the supremum of all Δ satisfying the condition, and we call it the **mass** m Known . The problem statement does not give the reason for imposing $m < \infty$. Reading it from the formula shows what it excludes. If $\sigma(H) = \{0\}$, every Δ satisfies the condition, and $m = \infty$. **Self-adjoint** (1-10) is the property of Hermitian matrices

(matrices whose eigenvalues are all real) extended to infinite-dimensional operators. For self-adjoint H, $\sigma(H) = \{0\}$ is the same as $H = 0$, so $m < \infty$ means "there is some state of positive energy above the vacuum".

The reason for calling it "mass" lies in the relation between energy and momentum ($\hbar = c = 1$; 2-04). A particle of mass μ has $E = \sqrt{|p|^2 + \mu^2} \geq \mu$. For the **mass operator** $M = \sqrt{H^2 - P^2}$ of § 6.2, $H \geq M$, and at momentum 0, $H = M$. A state with $M > 0$ can have its momentum brought to 0 by a Lorentz transformation (using the Lorentz covariance of the axioms), so the gap of H and the gap of M have the same width Known . If there is a state with $M = 0$ besides the vacuum, both gaps are 0, and the conclusion is the same. This equivalence is not written in the problem statement itself.

Figure 1 Moving m with the slider raises and lowers the hyperbola E = √(p² + m²) in the energy–momentum plane on the left, and opens and closes the gaps of σ(H) and σ(M) on the right. Switching the isolated one-particle state on and off with the button does not change σ(H)



m = 1.00, σ(H) = {0} ∪ [1.00, ∞), σ(M) = {0} ∪ {1.00} ∪ [2.00, ∞)

Computed The figure is drawn on the spot from the formula $E = \sqrt{p^2 + m^2}$ (momentum in one direction only; the shading is continuous spectrum).

Table 3 The spectra of four examples.

Example	$\sigma(M)$	$\sigma(H)$	Verdict
One kind of particle of mass m (no bound states)	$\{0\} \cup \{m\} \cup [2m, \infty)$	$\{0\} \cup [m, \infty)$	Gap
Only continuous spectrum above the gap	$\{0\} \cup [m, \infty)$	$\{0\} \cup [m, \infty)$	Gap
The massless free field (photon)	$[0, \infty)$	$[0, \infty)$	No gap
The vacuum only	$\{0\}$	$\{0\}$	Excluded by $m = \infty$

Rows 1 and 2 have the same $\sigma(H)$. What the problem statement asks for is only the gap of $\sigma(H)$; an isolated one-particle state is an extension of the problem (section 09). § 5 gives **clustering** as a consequence of the gap. For local fields $\mathcal{O}$ and $0 < C < \Delta$ with $\langle \Omega, \mathcal{O}\Omega \rangle = 0$, if the two points are far enough apart,

$$|\langle \Omega, \mathcal{O}(x)\, \mathcal{O}(y)\, \Omega \rangle| \leq e^{-C|x-y|} \quad (7)$$

Known . The gap determines how fast fluctuations at distant places approach independence. The body of the definition is in 4-11, and how to read it for finite systems is in 2-05.

07

Agreement at short distances — asymptotic freedom names "Yang–Mills"

Quantum Yang–Mills theory has no definition before it is constructed. What ties the constructed object to "the quantum Yang–Mills theory of G" is the condition in the first paragraph of § 4. The correlation functions of the quantum fields must agree **at short distances** with the predictions of asymptotic freedom and perturbative renormalisation theory found in textbooks (original: 'should agree at short distances with the predictions of asymptotic freedom and perturbative renormalization theory, as described in textbooks'). The predictions include the existence of a stress tensor and an operator product expansion, whose local singularities have the form predicted by asymptotic freedom Known .

Asymptotic freedom is the property that the strength of the coupling changes with scale and becomes weak at short distances (large momentum μ) (Gross–Wilczek and Politzer, 1973; references [21] and [38] of the problem statement). In a theory without quarks, $\alpha = g^2/4\pi$ runs in the following way (formula (9.3) of PDG 2024 with $n_f = 0$; on the right, the one-loop solution) Known .

$$\mu^2 \frac{d\alpha}{d\mu^2} = -b_0 \alpha^2 - b_1 \alpha^3 - \cdots, \qquad b_0 = \frac{11 C_A}{12\pi}, \qquad \alpha(\mu^2) \approx \frac{1}{b_0 \ln(\mu^2/\Lambda^2)} \quad (8)$$

Λ is a constant of integration, the scale at which the coupling defined in perturbation theory diverges.

Example. Computed For $SU(3)$, $b_0 = 0.8754$, and when $\mu/\Lambda = 10,\ 100,\ 1000$, $\alpha = 0.248,\ 0.124,\ 0.083$. Each time the momentum grows tenfold, $1/\alpha$ grows by the same amount $2b_0 \ln 10 \approx 4.03$. The weakening is logarithmic, and slow.

Solving (8) gives $\Lambda = \mu\, e^{-1/(2b_0\alpha)}$, and $e^{-1/x}$ approaches 0 faster than any x^k as $x \to 0^+$, so all coefficients of the power series in α are 0 Known . The classical action has no constant with the dimension of mass, and physics holds that the mass of the quantum

theory takes the form $\Delta = c\Lambda$ Physics . This is consistent with "the mass gap is not visible classically" in § 6.6. The problem statement does not specify the value of the coupling constant g (4-06, 4-07).

08

The two footnotes — quantum operators, and existence by compactness alone

The footnote on local fields (footnote 1 of Jaffe–Witten; § 4) says that, because of the subtleties of renormalisation, there is no natural one-to-one correspondence between classical "differential polynomials" and quantum operators, and that the spaces of those of dimension at most d are expected to correspond Known . The dimension is the mass dimension: A has 1, F has 2, and $\operatorname{Tr} F^2$ has 4.

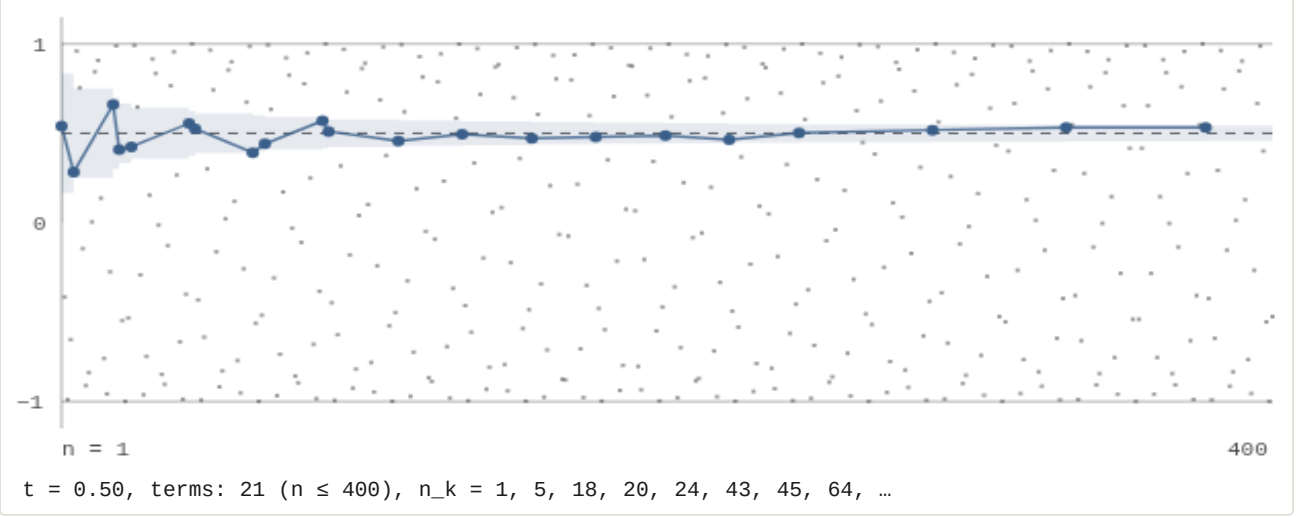
The footnote on weak existence (end of § 6.5) says that **weak existence (compactness)** is specifically excluded as a solution of the existence part, unless the properties of the limit (the mass gap and the axioms) are established by other methods Known . The text to which the footnote is attached says, in effect, "showing a mass gap uniform in the volume needs new ideas; with that, one could study the $T^4 \to \mathbb{R}^4$ limit". Uniform means that the right-hand side holds for the gap Δ_L for each box size L.

$$\Delta_L > 0 \ \text{(each } L\text{)} \qquad \text{rather than} \qquad \inf_L \Delta_L > 0 \ (9)$$

The left-hand side alone holds even for the massless free field. With the field set to 0 on the walls of a box of side L, the energy of the lowest vibration is $\sqrt{3}\,\pi/L$, and the gap closes as $L \to \infty$ (2-05).

A compactness argument is an argument of the type "a bounded sequence has a convergent subsequence" (Bolzano–Weierstrass). For a sequence of probability measures too, given a uniform estimate (tightness), a weakly convergent subsequence can be taken (Prokhorov's theorem) Known . **Tight** means that every measure in the sequence puts almost all of its probability (at least $1 - \varepsilon$) on a common compact set (for $\mathbb{R}^n$, a bounded closed set), and **weak convergence** means that the expectations of all bounded continuous functions converge. What is weak is that the limit can change with the choice of subsequence. Whether the limit is unique, and which of the gap or the axioms it satisfies, cannot be known from compactness alone (some properties, like reflection positivity, are preserved under weak convergence; 5-03). Figure 2 uses the sequence $a_n = \cos n$ to show how the limit is not unique.

Figure 2 Moving the target t with the slider extracts, from the same sequence a_n = cos n (dots), a subsequence converging to t (coloured dots and line). The band is the error allowed for the next term. "Play" moves t back and forth



Computed The numbers in this figure were computed on this machine. Rule: n_1 is the smallest n with $|\cos n - t| < 1/2$, and n_{k+1} is the smallest n with $n > n_k$ and $|\cos n - t| < 1/(k+2)$. The figure computes the same rule on the spot, and agrees with the python3 results (Table 4).

Computed **Table 4** Subsequences extracted with $n \le 400$.

t	First 8 terms n_k	Terms	$\cos n_k$ of the last term
1.0	1, 6, 7, 12, 13, 19, 25, 31	40	0.9873
0.5	1, 5, 18, 20, 24, 43, 45, 64	21	0.5328
−1.0	3, 9, 10, 16, 22, 28, 35, 41	41	−0.9998

The remainders of n divided by 2π are uniformly distributed (Weyl's equidistribution theorem), so a subsequence converging to any $t \in [-1, 1]$ can be taken Known . The difference between the empirical distribution of $\cos n$ for $n \le 100000$ and the arcsine distribution was below 0.0001 in all 10 intervals Computed . Boundedness alone does not make the limit unique.

To be exact, in the lattice method there is a measure μ_a for each lattice spacing a, and one wants to take $a \to 0$. Even if a subsequential limit is obtained from uniform estimates, whether it is unique, and whether it satisfies the axioms and the gap, are separate questions, and the footnote on weak existence says that these must be shown. It has been stated that what can be expected from the known results on the ultraviolet side in four dimensions is also a limit by a compactness argument using subsequences, and it is also written that the limit need not be unique (the introduction of Magnen–Rivasseau–Sénéor 1993; 7-07) Known .

In the construction published in 2026-11 too, the infinite-volume limit is taken along a subsequence by compactness, and uniqueness is not shown. The committee's assessment notes that, in this footnote, the problem statement regards subsequential limits alone as not enough, and then states that the reconstruction and the theorems on local fields give information beyond compactness (6-01; at the level "the committee's assessment is that it is proved in Lean"). The mass gap is not shown (6-04).

09

What is not written — extensions of the problem, QCD, formulations on the lattice

§ 5 lists the following as **natural extensions**, separate from the body of the problem Known . The existence of isolated one-particle states (an "upper gap" in addition to the mass gap). A proof of confinement. The existence of other four-dimensional gauge theories with added fields that preserve asymptotic freedom, and their dynamics (gap, confinement, chiral symmetry breaking). Extending the existence theorem to arbitrary four-dimensional manifolds. Written as a formula, the upper gap means that the spectrum of M has the following form (this chapter's way of writing; it corresponds to "m is an isolated eigenvalue of M" in § 6.2).

$$\sigma(M) = \{0\} \cup \{m\} \cup [m', \infty), \qquad m < m' \tag{10}$$

In a theory with only one kind of particle of mass m and no bound states, the two-particle continuous spectrum begins at $m' = 2m$ (row 1 of Table 3).

§ 1 names three properties that QCD should have at the quantum level: the mass gap, quark confinement and chiral symmetry breaking. Only the first entered the body of the problem, and the object, too, is pure Yang–Mills without quarks. The problem statement does not ask for the **value** of the gap either. Glueball masses and their ratio to the string tension are objects of physical measurement (5-08).

Example. Chatterjee's (2018) lattice correlation-length problem (Chatterjee's Problem 5.1) asks that on the lattice (β is the inverse of the lattice coupling; the larger it is, the weaker the coupling; 5-01) the correlations of plaquette variables decay exponentially (the correlation length $\xi(\beta)$ is finite), and that $\xi(\beta) \to \infty$ as $\beta \to \infty$. It is a formulation on the lattice, and a different thing from Clay's requirements, which include the continuum limit, $\mathbb{R}^4$, the axioms and non-triviality. Even if it were shown, that alone would not solve the Clay problem (5-09, 16-08).

10

What each word shuts out — a word-by-word list

We take the conditions of (1) word by word, together with the $m_G < \infty$ mentioned in section 01, and list what each word directly shuts out.

$$\forall G \; \exists \mathcal{T}_G \; \big[\, \text{axioms} \wedge \text{non-trivial} \wedge \text{short distances} \wedge \Delta_G > 0 \wedge m_G < \infty \, \big] \tag{11}$$

Table 5 A word-by-word list (this chapter's reading).

Word	What the word directly shuts out	Section
Any	A construction for only one convenient group (for example $SU(2)$)	02
Compact	Groups whose invariant quadratic form is indefinite, so that the classical energy is not positive	02
Simple	$G = U(1)$. The free electromagnetic field has no gap, so including it would make the claim false	02
$\mathbb{R}^4$	Theories on T^4 or on a lattice. On a finite lattice the integrals are finite-dimensional (§ 6.5) and always exist	03
Existence (axioms)	Something for which merely some limit exists, with neither a Hilbert space nor a Hamiltonian	04
Non-trivial	Limits collapsed to a Gaussian. The free field with mass satisfies the axioms and also has a gap	05
$\Delta > 0$	The massless free field. It satisfies the axioms but has no gap	06
$m < \infty$	A theory with only the vacuum ($\sigma(H) = \{0\}$)	06
Short distances	Another theory, unrelated to Yang–Mills, that has a gap	07
The footnote on weak existence	Compactness limits taken along subsequences (with neither uniqueness nor the gap and the axioms shown)	08

The words of Table 5 fall into two kinds according to what happens when they are removed. One kind makes the claim **false** when removed. "Simple" is one: as in the example below, a theory without a gap gets mixed in. The other kind makes the claim **easy** when removed. Removing "$\mathbb{R}^4$" leaves only finite-dimensional integrals on a finite lattice; removing "existence (axioms)" lets through something for which merely a limit exists; removing "the footnote on weak existence" makes a subsequential limit enough. Removing "$m < \infty$" lets a theory with only the vacuum satisfy the gap condition in form. In this chapter's reading, most of the words of the problem statement are of the latter kind, and close the loopholes to easy answers one by one.

The shutting-out overlaps. The free field with mass is caught by both "non-trivial" and "short distances". Even when one word is removed, another word may shut out the same thing. Table 5 is a guide to the role of each word.

Example. Removing "simple" means showing a gap for $G = U(1)$ as well, but that theory is the free electromagnetic field, which, as row 3 of Table 3 shows, has no gap. The claim with the word removed is false by this one example.

What this chapter can say

What is said	What is not said (where it is treated)
Known Where the sentence, the definitions, the footnotes and the extensions of the problem are in the problem statement, and what they say (checked against the Clay PDF)	The meaning of words the problem statement does not define ("non-trivial" and others) is this chapter's reading
Computed The dimensions and $h^\vee$ of Table 1 obtained from root systems, agreeing with the standard values. That $C_A = h^\vee$ in the normalisation of formula (2), checked for six groups	That the classification is exhausted by these, and a proof of $C_A = h^\vee$ for general groups (not proved in this book; list of values in 1-04)
Known That $m < \infty$ and $\sigma(H) \neq \{0\}$ are the same. The examples of Table 3	That the H of four-dimensional quantum Yang–Mills exists at all (the construction published in 2026-11 made it by reconstruction, at the level of the committee's assessment; 6-00; its gap is not shown)
Computed Table 4, the ratio (5), the numbers of (8), the empirical distribution of $\cos n$	A proof of Prokhorov's theorem (not treated in this book)
Physics $\Delta = c\Lambda$, the expectation of non-triviality	Neither has a proof (4-07, 17-12)
—	Table 5 is a reading, not a statement of the problem's authors. This chapter shows nothing about the problem

Sources and reproduction

Item	Kind	Source / tool
The sentence, § 1–§ 6.6, footnotes	Known	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory” (the official problem statement of the Clay Mathematics Institute, PDF on the Institute's website; checked in full). In: J. Carlson, A. Jaffe, A. Wiles (eds.), *The Millennium Prize Problems*, CMI / AMS, pp. 129–152 (year of publication not checked)
Date and place of the announcement	Known	The Clay Mathematics Institute website (the Millennium Prize Problems page) / the preface of the book above (Introduction, p. viii): 'The prize was announced at a meeting on May 24, 2000, at the Collège de France.' (checked in the PDF of the book published by the Institute)
(8), the general form $(11C_A - 4nfT_R)/(12\pi)$ of b_0, Λ	Known	S. Navas et al. (Particle Data Group), Phys. Rev. D 110 (2024) 030001, QCD chapter § 9.1.1, formula (9.3). The relevant passage was read (the same passage as in the source table of 4-06)
Asymptotic freedom	Known	D. J. Gross, F. Wilczek, Phys. Rev. Lett. 30 (1973) 1343 / H. D. Politzer, ibid. 30 (1973) 1346. Bibliographic data only

Item	Kind	Source / tool
Axiom systems	Known	R. Streater, A. Wightman, *PCT, Spin and Statistics and All That*, Benjamin, 1964 / K. Osterwalder, R. Schrader, Comm. Math. Phys. 31 (1973) 83–112, 42 (1975) 281–305. Bibliographic data only (content as in § 3 of the problem statement)
Triviality of φ^4_4	Known	M. Aizenman, H. Duminil-Copin, Ann. of Math. 194 (2021) 163–235, arXiv:1912.07973. Abstract checked
What is expected for the ultraviolet limit	Known	J. Magnen, V. Rivasseau, R. Sénéor, Comm. Math. Phys. 155 (1993) 325–383. Bibliographic data only (the relevant passage of the introduction is in the source table of 7-07)
Lattice correlation-length problem (Chatterjee's Problem 5.1)	Known	S. Chatterjee, "Yang–Mills for probabilists", arXiv:1803.01950 (2018). Abstract checked. The text of the lattice correlation-length problem is in the source table of 5-09
Definitions of $C_A = h^\vee$ (in the normalisation where the squared length of the longest root is 1) and $h^\vee$	Known	V. G. Kac, *Infinite Dimensional Lie Algebras*, 3rd edition (1990). Bibliographic data only (the same as in the source table of 1-04)
Classification, $C_A > 0$, invariant inner products, Isserlis, Weyl, Prokhorov	Known	Standard facts (no bibliography given)
Tables 1 and 4, numbers in the text, figures	Computed	python3. Table 1 by exact rational computation: generate all positive roots from the Cartan matrix; dimension = rank + $2\times$ (number of positive roots), $h^\vee = (\theta, \theta+2\rho)/(\theta,\theta)$ (θ is the highest root, 2ρ the sum of the positive roots; as a ratio it does not depend on the normalisation of the inner product). The normalisation of formula (2): structure constants built from the generators of the defining representation, checked numerically for $SU(2), SU(3), SU(4), SO(5), SO(7), SO(8)$ to give $C_A = h^\vee$ with the inner product where the squared length of the longest root is 1, and $2h^\vee$ with the one where it is 2. Numerical integration by SciPy. Figures by JavaScript in the browser (Figure 1 from the formula, Figure 2 by the same rule as Table 4)

Next chapter: 0-04 Four dimensions in mathematics and four dimensions in physics. For digging deeper into each word, go to the chapters named at the end of each section.

Four dimensions in mathematics and four dimensions in physics — the same R⁴ has two rulers, and the problem asks for a theory that works with both

In this chapter, § 01– § 08 are established knowledge and § 09 is a map of open questions.

What this chapter shows — That the $\mathbb{R}^4$ of the problem statement can be read in two ways: as Euclidean four dimensions (the mathematical side) and as the four dimensions of spacetime (the physical side). That what joins the two is imaginary time and reflection positivity, and that the mass gap too has two faces, "the gap of the Hamiltonian" and "exponential decay of correlations". That the "four" that matters in the problem is the four at which the dimension of the coupling constant vanishes.

Prerequisite chapters — None. The vocabulary of quantum mechanics is set out minimally in section 04, and reflection positivity in section 06 (for details, 2-05 and 4-10). The examples of sections 06 and 07 only use the modified Bessel function $I_k(J) = \frac{1}{2\pi}\int_0^{2\pi} e^{J\cos\theta}\cos k\theta\, d\theta$ $(k = 0, 1)$ as numbers. Having read the articles What happens only in four dimensions and One sign is different makes the connections easier to see.

Said first — This is a chapter that guides you along the road. The substitution of imaginary time is a heuristic, and the rigorous bridge (the Osterwalder–Schrader reconstruction) is treated in 4-10. The computation of section 07 is an example in the quantum mechanics of a single particle. For four-dimensional Yang–Mills, the Hilbert space at the end of the way back was made by reconstruction in the construction published in 2026-11 (at the level of the committee's assessment; section 09), but a gap on it has not been shown. This chapter shows nothing about the mass gap.

Order of this chapter

1. The R^4 of the problem statement — the signature is not written, and two axiom systems are specified
2. The same points, two rulers — a rotation by an imaginary angle becomes a boost
3. Where the cases of the two articles matter in the problem (Table 1)

4. Rotating time — oscillation turns into decay because the energy is positive (Figure 1, Table 2)
5. A dictionary between the two four dimensions (Table 3, Detour: self-duality lives only in imaginary time)
6. Reflection positivity — recovering a Hilbert space from probability
7. The two faces of the mass gap — counting with a particle on a circle (Figure 2, Table 4)
8. The "four" that matters in the problem — what is counted is the Euclidean d
9. Three barriers on the way back — mismatches peculiar to Yang–Mills

---

01

The R⁴ of the problem statement — the signature is not written

The problem statement (the Clay Mathematics Institute's description by Jaffe–Witten) asks to show that for any compact simple group G a non-trivial quantum Yang–Mills theory exists "on $\mathbb{R}^4$" and has a mass gap $\Delta > 0$ (a word-by-word reading is in 0-03). For this $\mathbb{R}^4$, the way of measuring length is not written.

The way of measuring appears in the surrounding sections. § 3 introduces the axioms as 'Gårding and Wightman gave mathematically precise axioms for quantum field theories on R4 with a Minkowski signature', and § 6 adds '(Equivalently in the Euclidean formulation one replaces Euclidean space-time R4 by R2 or R3.)'. Existence includes 'axiomatic properties at least as strong as those cited in [45, 35]', where [45] is Streater–Wightman (the axioms on the side of spacetime) and [35] is Osterwalder–Schrader (the axioms on the Euclidean side) Known . One can say that the same $\mathbb{R}^4$ has two rulers. To be exact, the set of points is the same, and what differs is the quadratic form that measures squared length ($c = 1$).

$$Q_{\mathrm{E}}(x) = x_1^2 + x_2^2 + x_3^2 + x_4^2$$
$$Q_{\mathrm{M}}(t, x) = -t^2 + x_1^2 + x_2^2 + x_3^2 \quad (1)$$

Example. Reading the last coordinate of the point $(1, 0, 0, 1)$ as time, $Q_{\mathrm{E}} = 2$ but $Q_{\mathrm{M}} = 0$, and it is joined to the origin exactly by light. $(0, 0, 0, 1)$ has $Q_{\mathrm{E}} = 1$, $Q_{\mathrm{M}} = -1$ (timelike). There is also the convention of writing the signature as $(+, -, -, -)$; the content is the same.

02

The same points, two rulers — a rotation by an imaginary angle becomes a boost

Fixing a ruler fixes the group of transformations that preserve it. $SO(4)$ is the group of 4×4 real matrices of determinant 1 that leave Q_{E} unchanged, and $SO(3,1)$ the one that leaves Q_{M} unchanged. The former is the rotation group of four dimensions and the latter the Lorentz group; together with translations they are the Euclidean group and the Poincaré group. Both have one direction for each of the $\binom{4}{2} = 6$ coordinate planes (What happens only in four dimensions). The difference is in the 3 planes that contain time, where rotations become boosts in spacetime (2-04).

$$\begin{aligned} \tau' &= \tau\cos\alpha - x\sin\alpha \\ x' &= \tau\sin\alpha + x\cos\alpha \\ &\downarrow\quad \tau = it,\ \ \alpha = i\varphi \\ t' &= t\cosh\varphi - x\sinh\varphi \\ x' &= x\cosh\varphi - t\sinh\varphi \end{aligned} \tag{2}$$

This is because $\cos(i\varphi) = \cosh\varphi$ and $\sin(i\varphi) = i\sinh\varphi$ Known . This φ is called the rapidity. A point with $x' = 0$ in the lower formula of (2) moves along $x = t\tanh\varphi$, so the velocity of the boost is $v = \tanh\varphi$. Example. Putting $\varphi = 0.7$, $(t, x) = (0.4,\ 1.3)$ into the left gives $t' = -0.484091,\ x' = 1.328286$, agreeing with the right Computed . The angle of a rotation goes once round in 2π, but the rapidity has no upper bound (as $\varphi \to \infty$, v only approaches 1). Restated in terms of quadratic forms, $SO(4)$ can move a point only on a sphere where Q_{E} is constant, but $SO(3,1)$ can carry it arbitrarily far on the hyperboloid $Q_{\mathrm{M}} = -1$. That $SO(4)$ is compact and $SO(3,1)$ non-compact is an expression of this difference (One sign is different § 01).

03

Where the cases of the two articles matter in the problem

The article What happens only in four dimensions looked at Euclidean $\mathbb{R}^4$ and its topology, and One sign is different at spacetime. Not every case matters for the problem. What matters most on the Euclidean side is that the 6 planes form 3 pairs of "perpendicular planes", so that the field strength F splits in two (1-12). The field strength is the generalisation to Yang–Mills (3-02) of $F_{\mu\nu}$, which combines the electric and magnetic fields of electromagnetism (2-03), and it has one component for each coordinate plane. In electromagnetism, the 3 components of the electric field correspond to the 3 planes containing time, and the 3 components of the magnetic field to the 3 purely spatial planes. Writing the operation that moves the component of

each plane to the perpendicular plane as the Hodge star $*$, the way the planes are divided becomes exactly the way F is divided.

$$\binom{4}{2} = 6 = 3 + 3$$
$$F = F^{+} + F^{-}, \qquad F^{\pm} = \tfrac{1}{2}\left(F \pm *F\right) \quad (3)$$

As examples, we list how the cases show up in the problem.

Table 1 The cases of the two articles and the problem (E = Euclidean, M = spacetime)

Case		Relation to the problem	This book
The axis disappears from rotations	E	Splits into $F^{\pm}$. The classification of instantons	3-05
24-cell	E	Finite subgroups of SU(2). Lattice models with finite groups	5-10
Exotic $\mathbb{R}^4$	E	Classical equations tell topologies apart (Donaldson). Does not appear in the problem statement	3-06
Rotation stops being rotation; introducing imaginary numbers	E, M	A heuristic joining the two four dimensions	Section 04
The light cone and causality	M	The axiom that fields in regions that cannot be joined by light commute	4-09
Gravity, tails of waves	M	The problem is only about flat spacetime; does not appear	—

How to read the table. The top three rows (E) matter as tools and models in Part 3 (classical Yang–Mills) and Part 5 (the lattice). The fourth row (E, M) is used in sections 04–07 of this chapter to join the two four dimensions. Of the M rows, the light cone enters the axioms of Part 4 (4-09), and the gravity row is outside the problem.

The "four" that matters in the problem is not a phenomenon peculiar to four dimensions like exotic structures, but the four at which the dimension of the coupling constant vanishes (section 08). However, the proof of the existence of exotic $\mathbb{R}^4$ uses, together with Freedman's topological theorem, the same Yang–Mills equations (the space of their anti-self-dual solutions) (1-12, 3-06). Classical theory uses the solutions of the equations as a tool for seeing topology; quantum theory uses the action as a probability weight.

04

Rotating time — oscillation turns into decay because the energy is positive

The first tool for going back and forth between the two four dimensions is to replace time t by an imaginary number (Wick rotation; the name is said to come from Wick

1954; for the form in field actions, 4-02 § 02 is the authoritative place). Putting $t = -i\tau$ into (1) turns Q_{M} into Q_{E}. What changes can be seen in time evolution in quantum mechanics.

A state is a vector of length 1 in a Hilbert space (a complex vector space with an inner product, closed under limits; 1-10), and we assume it can be expanded in the eigenvectors ψ_n (eigenvalues E_n) of the energy operator H (the Hamiltonian). The component with energy E_n turns with phase $e^{-iE_n t}$ ($\hbar = 1$). Taking the lowest state as the vacuum Ω (energy 0) and writing $O\Omega = \sum_n c_n \psi_n$ for the vector obtained by applying an observable O, the correlation across a time t is

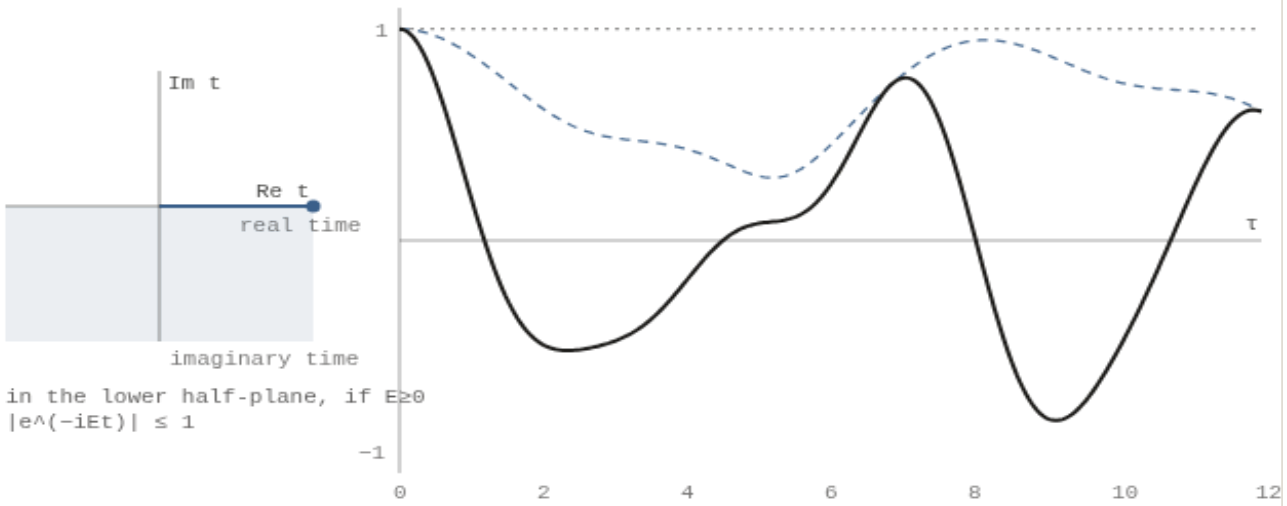
$$C(t) = \langle O\Omega,\ e^{-iHt}\, O\Omega \rangle = \sum_n |c_n|^2 e^{-iE_n t}$$

$$t = \tau e^{-i\vartheta} \implies \left| e^{-iE_n t} \right| = e^{-E_n \tau \sin\vartheta} \qquad (4)$$

ϑ is the angle by which time is rotated in the complex plane: $\vartheta = 0$ is real time and $\vartheta = 90^\circ$ imaginary time. If every E_n is at least 0, then rotating t into the lower half-plane keeps each term no greater than 1, and at $\vartheta = 90^\circ$ oscillation turns into decay $e^{-E_n\tau}$. If there were a negative E_n, that term would grow and the sum would lose its meaning. § 3 of the problem statement writes 'Wightman used positive energy to establish analytic continuation of expectations of Minkowski field theories to Euclidean space.' Known .

Example. We draw the connected part C_{conn}, with the vacuum term removed, for a model with $E_n = 1,\ 1.7,\ 2.6$ and weights $0.6,\ 0.3,\ 0.1$. The gap is $\Delta = 1$.

Figure 1 Moving the angle ϑ from 0° to 90° turns the direction of time (left) towards the imaginary axis, and the oscillation of the real part of the correlation (solid line, right) turns into decay. The dashed line is |C| and the dotted line e^(−Δτ sinϑ). Adding "negative energy" turns it into growth



ϑ = 0°, rate of decay Δ·sinϑ = 0.000, rate of oscillation Δ·cosϑ = 1.000, |C(τ=8)| = 0.94469

Computed The numbers in this figure were computed on this machine (the sum (4) computed in the browser; Table 2 by python3). "Negative energy" is a term with $E = -0.3$ and weight 0.03.

Table 2 $|C_{\text{conn}}(\tau e^{-i\vartheta})|$ (the last column is the case with negative energy added)

ϑ	τ = 2	τ = 8	$e^{-8\sin\vartheta}$	With negative, τ = 8
0°	0.62270	0.94469	1	0.92401
30°	0.23920	0.01104	0.01832	0.08978
60°	0.11873	0.00059	0.00098	0.24004
90°	0.09176	0.00020	0.00034	0.33090

In real time it is still 0.94 at $\tau = 8$, and the correlation does not decrease. The more it is rotated, the faster it decreases; with negative energy added, the more it is rotated, the more it grows. The upper bound $|C_{\text{conn}}| \le e^{-\Delta\tau\sin\vartheta}$ follows from the triangle inequality applied to each term of (4), since the weights sum to 1 and $E_n \ge \Delta$ Known . It was also checked at all points of a fine grid in $0 \le \vartheta \le 90°$ and $0 \le \tau \le 12$ Computed .

05

A dictionary between the two four dimensions

Moving to imaginary time replaces the vocabulary of quantum theory with the vocabulary of probability. § 3 of the problem statement sums up that Symanzik read Euclidean expectations as 'a statistical mechanical ensemble of classical Markov fields [46], with a probability density proportional to exp(−S)', and that Osterwalder and Schrader gave a general theory of the equivalence of the two axiom systems Known (the exact direction and conditions of the theorem are in 4-10).

$$e^{iS[\phi]} \longrightarrow e^{-S_{\mathrm{E}}[\phi]}$$
$$e^{-iHt} \longrightarrow e^{-H\tau} \quad (5)$$

The e^{iS} on the left (S is the action and ϕ the field; 2-01, 2-02) oscillates with size 1 and is not a probability. The $e^{-S_{\mathrm{E}}}$ on the right (S_{E} is the action moved to imaginary time) is positive, and when normalised can be a probability weight (2-07, 4-02). This is why on the mathematical side Yang–Mills becomes "the problem of constructing a probability measure on the space of connections" (a connection is the gauge field A; for details, 1-07). As examples, we list the main rows of the correspondence.

Table 3 A dictionary between the two four dimensions (the rigorous form is in 4-10)

Four dimensions of physics (spacetime)	Four dimensions of mathematics (Euclidean)
The Lorentz group $SO(3,1)$	The rotation group $SO(4)$
Time evolution e^{-iHt}	$e^{-H\tau}$. On the lattice, the transfer matrix (5-03)
The weight of the path integral e^{iS}	The probability weight $e^{-S_{\mathrm{E}}}$ (2-08)

Four dimensions of physics (spacetime)	Four dimensions of mathematics (Euclidean)
Vacuum expectations of products of fields (Wightman functions)	Moments of a probability measure (Schwinger functions)
A Hilbert space and positivity of energy $H \geq 0$	Reflection positivity (section 06)
Fields in regions that cannot be joined by light commute	Schwinger functions do not change under permutation of their arguments (for bosonic fields)
Mass gap: the spectrum of H is absent from $(0, \Delta)$	Connected correlations decrease exponentially (section 07)

Every row finally passes through the single point "with reflection positivity, one can get back to a Hilbert space".

Detour: self-duality lives only in imaginary time

A subsection that may be skipped.

There are also rows the dictionary cannot translate. The sign of applying the Hodge star $*$ (the operation that moves the component of each plane to the perpendicular plane; 1-12) twice changes with the ruler.

$$
\begin{aligned}
** &= +1 \quad \text{(Euclidean)} \\
** &= -1 \quad \text{(spacetime)} \\
* &: (E, B) \mapsto (-B, E) \quad \text{(spacetime)}
\end{aligned}
\tag{6}
$$

In Euclidean space there are real fields satisfying $F = *F$ (self-dual), and instantons are its solutions (3-05). In spacetime the eigenvalues of $*$ are $\pm i$, so for a real field, $*F = \pm F$ implies $F = 0$. Example. Building $*$ as a six-component matrix, the eigenvalues were ± 1 in Euclidean space and $\pm i$ with signature $(-,+,+,+)$, three of each Computed (the sign on the right of (6) is by convention). Reading instantons as tunnelling in imaginary time is a semiclassical interpretation Physics .

06

Reflection positivity — recovering a Hilbert space from probability

On the Euclidean side there is only a probability measure (expectation $\mathbb{E}$), with neither a Hilbert space nor a Hamiltonian. To recover them an inner product is needed, and the material is the mirror of time $\Theta : \tau \mapsto -\tau$. Let X, Y be a quantity determined only by the fields on the side $\tau > 0$, and ΘX the result of reversing the sign of every time in X (a quantity on the side $\tau < 0$), and set

$$
\begin{aligned}
\langle X, Y \rangle &:= \mathbb{E}\big[\, \Theta X \; Y \,\big] \\
\langle X, X \rangle &\geq 0
\end{aligned}
\tag{7}
$$

The second half is reflection positivity (also called OS positivity). If it holds, $\langle \cdot,\cdot \rangle$ becomes an inner product (after identifying vectors of length 0), and the operation of shifting by s in the direction of τ acts as e^{-sH} $(H \geq 0)$ Known . § 6.5 of the problem statement writes 'Establishing a quantum mechanical Hilbert space is part of the solution to this Millennium problem.'

Example. Put a particle rotating on a circle (moment of inertia I, angle θ) on an imaginary-time lattice with step a. In this chapter θ is used for the angle of this particle (it is different from the angle of the θ term in 3-05 and elsewhere). Writing the angle at time ka as θ_k and approximating one step of the action $\int \frac{I}{2}\dot\theta^2 d\tau$ by $\frac{I}{a}\big(1-\cos(\theta_{k+1}-\theta_k)\big)$, the weight is

$$\prod_k e^{J\cos(\theta_{k+1}-\theta_k)}, \qquad J=\frac{I}{a}$$

$$e^{J\cos u} = \sum_{n\in\mathbb{Z}} I_n(J)\, e^{inu} \tag{8}$$

$$I_n(J) = \frac{1}{\pi}\int_0^{\pi} e^{J\cos u}\cos nu\, du$$

J is the strength with which the angles at neighbouring times are pulled into line, and in statistical mechanics this chain of weights is called the one-dimensional XY model (the vocabulary of partition functions and transfer matrices is in 2-08 and 2-09). I_n is a modified Bessel function, and since the coefficients of its power series are all positive, $I_n(J)>0$ Known . The transfer matrix T is the operator whose integral kernel is the weight of one step.

$$(Tf)(\theta) = \int_0^{2\pi} K(\theta,\theta')\, f(\theta')\, d\theta', \qquad K(\theta,\theta') = e^{J\cos(\theta-\theta')}$$

Substituting the expansion (8), T maps $e^{in\theta}$ to itself with eigenvalue $2\pi I_n(J)$ (agreeing to 12 digits with a numerical computation that divides θ into 256 points Computed). If the plane of reflection is placed between lattice points, the left-hand side of (7) becomes a quadratic form sandwiching T, and since the eigenvalues are positive it is at least 0. In this way $H_a = -\frac{1}{a}\log\big(T/2\pi I_0(J)\big) \geq 0$ is determined. Reflection positivity in Wilson's lattice gauge theory was shown by Osterwalder–Seiler 1978 (for the construction of the transfer matrix, also Lüscher 1977; 5-03).

07

The two faces of the mass gap — counting with a particle on a circle

The definition of the mass gap is on the side of spacetime; § 4 of the problem statement writes 'A quantum field theory has a mass gap Δ if H has no spectrum in the

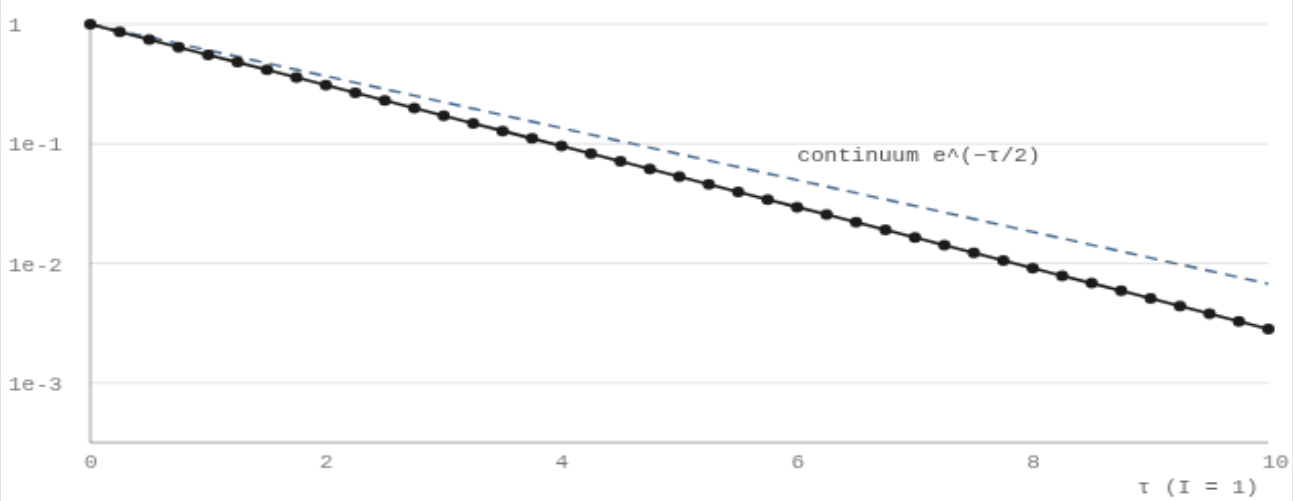
interval $(0,\ \Delta)$ for some $\Delta > 0$.' (H is the Hamiltonian; the spectrum $\operatorname{spec} H$, in finite dimensions, is the set of eigenvalues). On the Euclidean side the same thing appears as decay of correlations. We call this the two faces of the mass gap. To be exact, they are a condition on the spectrum of the Hamiltonian and exponential decay of connected correlations (Schwinger functions with the vacuum contribution subtracted; Table 3). Setting $\vartheta = 90^\circ$ in (4),

$$\begin{aligned} \operatorname{spec} H \subset \{0\} \cup [\Delta, \infty) \implies \\ C_{\mathrm{conn}}(\tau) = \sum_{E_n \geq \Delta} |c_n|^2 e^{-E_n \tau} \\ \leq C_{\mathrm{conn}}(0)\, e^{-\Delta \tau} \end{aligned} \tag{9}$$

If O overlaps with the lowest excitation, $-\frac{d}{d\tau} \log C_{\mathrm{conn}}$ approaches Δ; in the model of Figure 1 it was 1.178, 1.086, 1.021, 1.001 at $\tau = 1, 2, 4, 8$ Computed . Measuring the masses of glueballs (candidates for excitations corresponding to Yang–Mills particles) on the lattice uses this face (5-08). In Euclidean space time and space are carried into each other by rotations, so decay in the direction of τ is also decay in spatial directions, which leads to the clustering of § 5 of the problem statement Known . A local observable (a quantity determined only by the fields in a bounded region) rotated is again a local observable, so decay at the same rate holds in every direction. But this is a matter for after the reconstruction has given a Hilbert space and a Hamiltonian.

Example. Take $O = e^{i\theta}$ for the particle of section 06. The eigenvalues of $H = -\frac{1}{2I}\frac{d^2}{d\theta^2}$ are $\frac{n^2}{2I}$, the gap is $\Delta = \frac{1}{2I}$, and $O\Omega$ has only components with $n = 1$, so $C_{\mathrm{conn}}(\tau) = e^{-\Delta\tau}$. On the lattice, an infinitely long chain gives $\langle e^{i(\theta_k - \theta_0)} \rangle = (I_1/I_0)^k$, and the rate of decay read as $\tau = ka$ is $\Delta(a) = -\frac{1}{a} \log \frac{I_1(J)}{I_0(J)}$. This is the gap on the lattice measured in physical units (the smallest positive eigenvalue of H_a in section 06), and in lattice units it is $a\,\Delta(a)$ (the m_{lat} of Appendix A-01).

Figure 2 Making the step a = I/J finer brings the slope of the lattice correlation (solid line; dots too on coarse lattices) close to the continuum e^(−Δτ) (dashed line). The vertical axis is logarithmic. Clicking a row of Table 4 also changes J




```
J = I/a = 4, a = 0.25, Δ(a) = 0.586941, Δ(a)/Δ = 1.173882 (Δ = 0.5)
```

Computed The numbers in this figure were computed on this machine (I_1/I_0 and $\Delta(a)$ are the values of Table 4 computed with python3 and SciPy; the figure only joins the points $(ka,\ (I_1/I_0)^k)$).

Table 4 The lattice for a particle on a circle ($I = 1$, $\Delta = 0.5$, step $a = 1/J$). The last column is the ratio to the second level $E_2 = 2$

J	Δ(a)/Δ	1 + 1/(2J)	E₂(a)/E₂
1	1.613125	1.500000	1.116436
2	1.439436	1.250000	1.196582
4	1.173882	1.125000	1.130427
8	1.071311	1.062500	1.065295
16	1.033214	1.031250	1.032022
32	1.016093	1.015625	1.015824
64	1.007927	1.007813	1.007863

On coarse lattices ($J \le 2$) $\Delta(a)/\Delta$ is 1.4–1.6, a different number from the gap. Making the lattice finer brings it to 1 at almost the same rate as $1 + \frac{1}{2J}$, and the second level approaches too. The correlation computed from products of transfer matrices on a chain of period 400 also agreed with $(I_1/I_0)^k$ to 10 digits Computed . In this example the existence of a gap is clear from the start. What is needed for four-dimensional Yang–Mills is that $\Delta(a)$ stays positive in the limit where the decay seen in lattice units $a\,\Delta(a)$ goes to 0 (5-09). The Hilbert space of the limit was made by reconstruction in the construction published in 2026-11 (at the level of the committee's assessment; section 09).

```
08
```

The "four" that matters in the problem — what is counted is the Euclidean d

The counting of dimensions is done on the Euclidean side, over d equivalent coordinates. With $c = 1$, time too is measured as a length, so 3+1-dimensional spacetime is $d = 4$. Write the d-dimensional Euclidean action as $S_d[A] = \frac{1}{2g^2} \int_{\mathbb{R}^d} \operatorname{tr} F_{\mu\nu} F_{\mu\nu} \, d^d x$ (summed over μ, ν). This is the form of the action in Appendix A-01; here we take the convention that puts the coupling g in front of the action and does not include it in F. With $\hbar = c = 1$, writing the mass dimension as $[\cdot]$ (for the derivation, 2-04 and 1-12),

$$\begin{gathered} [g^2] = 4 - d \\ S_d[A_\lambda] = \lambda^{4-d} \, S_d[A] \\ \left(A_\lambda(x) = \lambda A(\lambda x)\right) \end{gathered} \tag{10}$$

and only in four dimensions does the classical action have no scale of length. A mass cannot be built from classical constants, so the gap is expected to be a constant multiple $\Delta = C\Lambda$ (C a dimensionless constant) of a scale Λ born from quantum effects Physics . This is what lies behind 'the mass gap is not classically visible' in § 6.6 of the problem statement (4-07). A figure that moves the dimension d continuously and sets it beside other quantities that switch at four is the dimension slider in 1-12 § 10.

Example. In three dimensions $[g^2] = 1$, so g^2 itself becomes a mass scale, and by this counting it falls on the easy side. Chevyrev (arXiv:2202.13359 § 1.1) also writes 'The cases d = 2, 3 are considered substantially simpler than d = 4', but in the same section the construction of the three-dimensional measure 'even on T3, remains open' Known . In two dimensions the continuum measure has been constructed; in four it has not been constructed as a measure (what the construction published in 2026-11 made is at the level of correlation functions; section 09. Where things stand by dimension is in 7-10, misconceptions in 16-08).

09

Three barriers on the way back — mismatches peculiar to Yang–Mills

For the particle on a circle the dictionary could be used as it is. In this chapter we call the way back the road that starts from a theory constructed on the Euclidean side and follows the dictionary in reverse to a theory on the side of spacetime. To be exact, the way back means obtaining, from a Euclidean probability measure by the OS reconstruction, a theory satisfying the Wightman axioms (one with a Hilbert space and $H \geq 0$). For Yang–Mills there are three barriers along it: observables, regularisation, volume. The whole road, set out, is as follows.

$$\underbrace{\text{lattice, regularisation}}_{\text{reflection positivity}} \xrightarrow{a\to 0} \text{continuum measure}$$
$$\xrightarrow{T^4\to\mathbb{R}^4} \text{Euclidean-invariant}$$
$$\xrightarrow{\text{reconstruction}} (\mathcal{H},\, H\geq 0) \xrightarrow{?} \Delta > 0 \qquad (11)$$

The regularisation at the left end puts a minimal length into the theory (the lattice step a or the like) so that infinities do not appear. T^4 is a four-dimensional box with each side joined periodically, to keep the volume finite. For four-dimensional non-abelian Yang–Mills, the construction published in 2026-11 went through everything up to just before the last arrow of (11). The continuum object, however, is not a measure but correlation functions, and the volume limit is along a subsequence (the committee's assessment is that it is proved in Lean; 6-00, 12-06, 12-08). The last arrow, the "?", the mass gap, has not been shown (6-04). § 6 of the problem statement of 2000 writes of the construction on T^4 'while the construction is not complete'. In two dimensions the road goes through from the lattice to the continuum measure on the plane $\mathbb{R}^2$ (7-02). The three barriers are as follows.

Observables. Only gauge-invariant quantities have meaning, that is, quantities whose values do not change under gauge transformations (the operation of choosing a group element at each point and replacing the field; 3-01). Local ones are composite quantities such as $\mathrm{Tr}\, F_{ij}F_{kl}(x)$ (§ 4 of the problem statement). Chatterjee (arXiv:1803.01950 § 1), citing Seiler, writes 'the key observables in these theories do not take values at points, but at curves', and Driver 1989 says that Wilson loops are 'quite singular in d = 3 and especially d = 4' Known . This barrier bears on every stage of (11).

Regularisation. The barrier at the left end of (11). Wilson's lattice preserves reflection positivity but reduces rotations to steps of 90°. § 6.5 of the problem statement writes that there are few methods for recovering reflection positivity lost through regularisation (quoted in 0-03 § 03), and Magnen–Rivasseau–Sénéor 1993, which uses a continuum regularisation, too stops at considering that OS positivity (reflection positivity) can be shown with additional work Known .

Volume. The barrier at the second arrow of (11). Even on T^4, if there is reflection positivity, a finite-volume Hilbert space and transfer matrix are obtained, but a theory with Poincaré symmetry on $\mathbb{R}^4$ needs $T^4 \to \mathbb{R}^4$. § 6 of the problem statement writes 'Nor do present methods suggest how to obtain the existence of the infinite volume limit T4 → R4.' Known . The construction published in 2026-11 took the infinite-volume limit along a subsequence by compactness (12-05). That the limit does not depend on the subsequence has not been shown (6-04).

Example. For non-abelian Yang–Mills in three and four dimensions, a unique non-trivial continuum limit has not yet been shown (in four dimensions, a non-trivial limit was made by the construction published in 2026-11, but its uniqueness has not been

shown; in three dimensions it is not constructed even on T^3; section 08). The final "?" of (11) is the body of the mass gap, but the gap is not a single arrow left over at the end. § 6.5 of the problem statement writes 'New ideas are needed to prove the existence of a mass gap that is uniform in the volume of space-time', and goes on to expect that if a gap independent of the volume (uniform however large the box is made) were shown, it would become possible to study the $T^4 \to \mathbb{R}^4$ limit Known . So the barriers of the gap and of volume are entangled (7-11, 16-05).

What this chapter can say

Content	Status
That the $\mathbb{R}^4$ of the problem statement has no specified signature, and that the axiom systems of both sides are specified as [45, 35]	Said Known (problem statement § 3, § 4, § 6)
(2), (6), (8); that with positive energy it becomes decay in imaginary time (4); that the gap gives an upper bound on decay (9)	Said Known (Figures 1–2, Tables 2 and 4 are Computed)
The dictionary (Table 3), and the procedure (7) for building a Hilbert space from reflection positivity	Known as a correspondence. The form of the theorem is treated in 4-10; this chapter only states it
That the gap is born from quantum effects in the form $\Delta = C\Lambda$	At the level of physics Physics (4-07)
That each arrow of (11) holds for four-dimensional Yang–Mills	The last arrow (the mass gap) is unsolved. The arrows before it were passed at the level of correlation functions by the construction published in 2026-11 (at the level of the committee's assessment; the volume limit along a subsequence; 6-00). This chapter shows nothing

Sources and reproduction

Item	Kind	Source / tool
The numbers of (2) and (6), Tables 2 and 4, Figures 1–2, the effective mass of section 07 and the check with chains	Computed	python3 (NumPy, SciPy). Figure 1 computes (4) in the browser; Figure 2 draws the values of Table 4

Item	Kind	Source / tool
Quotations from the problem statement	Known	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory”, the official problem statement of the Clay Mathematics Institute (problem announced in 2000). In: J. Carlson, A. Jaffe, A. Wiles (eds.), *The Millennium Prize Problems*, CMI / AMS, pp. 129–152 (checked by contents and pages; year of publication not checked). Text checked
The axioms of both sides, and reconstruction	Known	R. F. Streater, A. S. Wightman, *PCT, Spin and Statistics and All That* (1964). K. Osterwalder, R. Schrader, Comm. Math. Phys. 31 (1973) 83–112; 42 (1975) 281–305. Bibliographic data only
Reflection positivity and the transfer matrix on the lattice	Known	K. Osterwalder, E. Seiler, Ann. Phys. 110 (1978) 440–471. Checked from the abstract (physical positivity of the Schwinger functions of lattice gauge theory, a positive self-adjoint transfer matrix). M. Lüscher, “Construction of a selfadjoint, strictly positive transfer matrix for Euclidean lattice gauge theories”, Comm. Math. Phys. 54 (1977) 283–292. Checked from the abstract
Wick rotation	Known	G. C. Wick, Phys. Rev. 96 (1954) 1124–1134. Checked from the abstract (analytic continuation to purely imaginary values of the relative time, rotation of the integration contour). Attributing the name "Wick rotation" is a convention
Observables on curves / where two and three dimensions stand	Known	S. Chatterjee, arXiv:1803.01950 § 1 / I. Chevyrev, arXiv:2202.13359 § 1.1. Text checked
Singularity of Wilson loops	Known	B. K. Driver, Comm. Math. Phys. 123 (1989) 575–616. The relevant sentence of the introduction checked
Construction in four dimensions with an infrared cutoff	Known	J. Magnen, V. Rivasseau, R. Sénéor, Comm. Math. Phys. 155 (1993) 325–383. The passage on OS positivity in § I checked
Modified Bessel functions	Known	NIST DLMF 10.32.3 (integral representation) checked. Power series in § 10.25

Next chapter: 1-01 Vector spaces and linear maps. This chapter continues in 2-04, 4-02, 4-10 and 4-11.

Vector spaces and linear maps — reading a matrix as "a rule that moves vectors", and lining up eigenvalues, the exponential, Hermitian and unitary

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — Reading a matrix as "a rule that moves vectors". That an eigenvalue is the stretching factor of "a direction whose orientation does not change". That Hermitian matrices (real eigenvalues) and unitary matrices (preserving length) are joined by $U = e^{iH}$.

Prerequisite chapters — None. Products and determinants of matrices, complex numbers, and ordinary differential equations with constant coefficients are taken as known.

Said first — We restrict to finite dimensions. The mass gap is a problem about operators in infinite dimensions, and conclusions in finite dimensions do not necessarily extend as they are (1-10). All the theorems are known linear algebra.

Order of this chapter

1. Why start from linear algebra — three places where matrices appear
2. Vector spaces — a real vector space with complex entries
3. Linear maps and matrices — the result changes with the order of multiplication
4. Eigenvalues and eigenvectors — directions whose orientation does not change (Figure 1) / Detour: eigenvectors of rotations
5. Inner products and adjoints — length with complex numbers
6. Hermitian matrices — real eigenvalues and the gap (Figure 2) / Detour: levels avoid each other
7. Unitary matrices — transformations that preserve length
8. The matrix exponential — series and differential equations (Figure 3) / Detour: an example where summing the series fails

9. Joining Hermitian and unitary — $U = e^{iH}$

01

Why start from linear algebra

In Yang–Mills theory, matrices go where numbers would be expected. They have three roles.

Transformations	Fields	Energy
A gauge transformation $g(x)$ is a unitary matrix multiplying by a matrix that preserves length at each point (§ 07)	The value of the gauge field $A_\mu(x)$ is a matrix μ is the number of a direction (time, space). The value changes with the order of multiplication = non-commutative (§ 03, § 09)	The Hamiltonian H is Hermitian the mass gap corresponds to the gap between the lowest value and the next (§ 06). To be exact, the bottom of the spectrum with the vacuum (the lowest state) removed (1-10, 4-11)

If the size of the matrices used for gauge transformations is N ($N \times N$ matrices), electromagnetism is the case $N = 1$, and a gauge transformation just multiplies by a number $e^{i\alpha}$ of absolute value 1 (2-03). Products of numbers do not depend on order, so in the absence of matter the equations of the gauge field itself (Maxwell's equations) remain linear. For $N \geq 2$ order matters, and products of gauge fields with each other appear in the equations even without matter (3-02). This chapter lays that foundation in finite dimensions.

02

Vector spaces

A vector space is a collection in which addition and scalar multiplication can be done, and the same rules as for numbers hold, such as the associative, commutative and distributive laws. Taking scalars from $\mathbb{R}$ gives a real vector space, from $\mathbb{C}$ a complex vector space. When every vector can be written uniquely as $v = c_1 v_1 + \cdots + c_n v_n$, $v_1, \ldots, v_n$ is called a **basis** and the number n the **dimension**. Examples are $\mathbb{R}^n$, $\mathbb{C}^n$, all $n \times n$ matrices (dimension n^2), and all continuous functions on an interval (infinite-dimensional, 1-10).

An example this book returns to again and again is the set of all traceless 2×2 Hermitian matrices. Hermitian means $H^\dagger = H$, where † is the operation of transposing

and taking the complex conjugate of each entry (§ 05, § 06). Sums and real multiples are Hermitian, but multiplying by i gives $(iH)^\dagger = -iH$, which falls outside. So it is **a real vector space even though its entries are complex**, and its dimension is 3.

$$\begin{pmatrix} x & y - iz \\ y + iz & -x \end{pmatrix} = x \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + y \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + z \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad x, y, z \in \mathbb{R}$$

The three on the right reappear in 1-02 § 05 and 1-04 as the Pauli matrices $\sigma 3, \sigma 1, \sigma 2$, and the values of the $\mathrm{SU}(2)$ gauge field lie in this space (times i).

03 Linear maps and matrices

A linear map is a map that preserves sums and scalar multiples ($L(u + v) = Lu + Lv$, $L(cv) = c\,Lv$). Once a basis is fixed it can be written as a matrix, and **column j is where the basis vector e_j goes**. Composition is the matrix product $(AB)v = A(Bv)$, and unlike numbers, swapping the order generally changes it.

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},\ B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}: \qquad AB = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad BA = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

The difference $[A, B] = AB - BA$ is called the **commutator**. In this example it is $[A, B] = \mathrm{diag}(1, -1)$, the first of the basis of § 02. How commutators become the skeleton of Lie algebras is treated in 1-03.

The **determinant** $\det A$ is the signed scaling factor of area (volume), and the **trace** $\mathrm{tr}\,A$ is the sum of the diagonal entries. From $\det(AB) = \det A \det B$ and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, neither changes under a change of basis $A \mapsto P^{-1}AP$; they are properties of the map itself. Example: the rotation $R(\theta) = \left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right)$ has $\det R = 1$, $\mathrm{tr}\,R = 2\cos\theta$.

04 Eigenvalues and eigenvectors

A map changes the orientation of most vectors, but it may have **directions that it only stretches or shrinks without changing their orientation**. A vector in such a direction is an eigenvector, and the stretching factor is the eigenvalue.

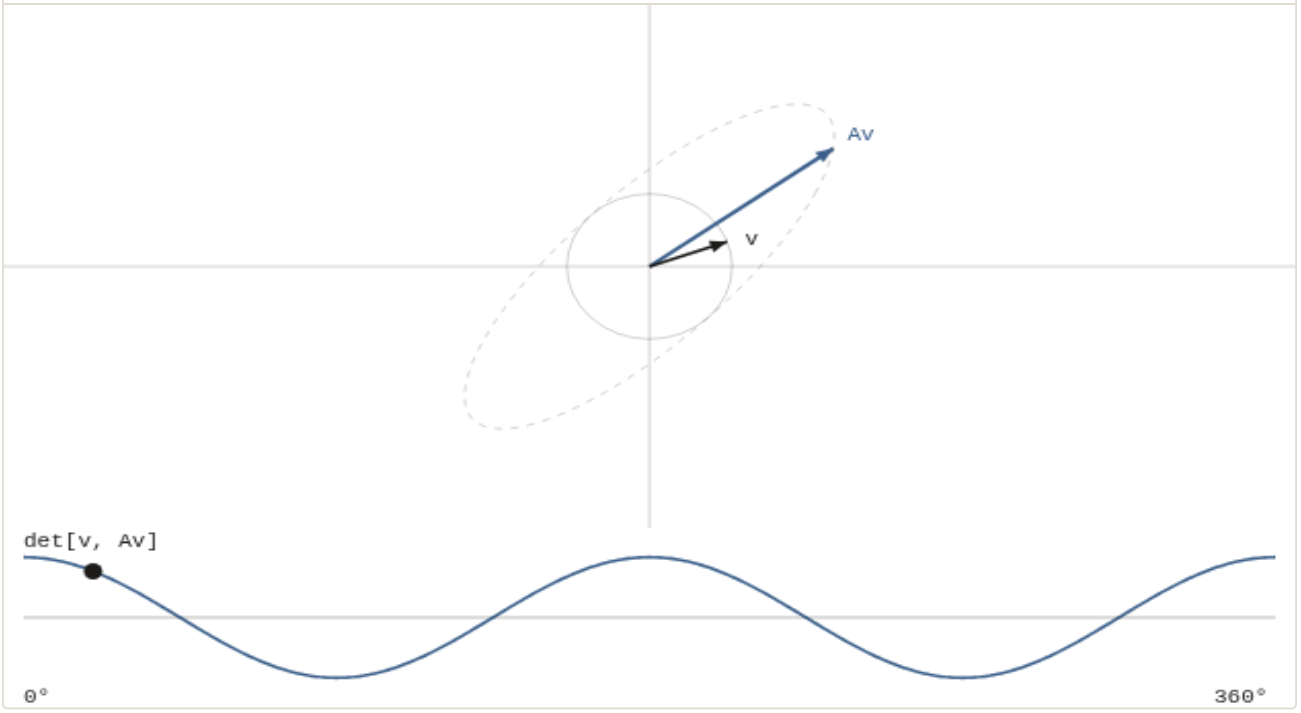
$$Av = \lambda v \ \ (v \neq 0) \qquad \Longleftrightarrow \qquad \det(A - \lambda I) = 0$$

For 2×2, $\lambda^2 - (\mathrm{tr}\,A)\lambda + \det A = 0$. For a real 2×2 matrix, if the discriminant $D = (\mathrm{tr}\,A)^2 - 4\det A$ is positive there are two real eigendirections, if negative there are none (the

eigenvalues are a complex pair), and if 0 there is a double root (with a double root there is only one eigendirection, or, for a scalar matrix λI, every direction). With P, made by lining up n independent eigenvectors, $P^{-1}AP$ becomes a diagonal matrix (diagonalisation).

That v and Av lie on the same line is the same as the determinant $\det[v, Av]$ of the two side by side being 0. This can be checked in Figure 1.

Figure 1 Turning the unit vector v with the slider moves Av (blue). The angles at which they lie on the same line are the eigendirections, and the curve $\det[v, Av]$ below becomes 0 there. The dotted line is the image of the unit circle



Computed The numbers in this figure were computed on this machine (the table by NumPy's eigenvalue computation). Matrices separate rows with ";". Clicking a row of the table below puts that matrix into the figure.

Matrix	D	Eigenvalues	Real eigendirections (from the x axis)
Symmetric (2 1 ; 1 2)	**4**	**3, 1**	**45°, 135° (orthogonal)**
Not symmetric (1.5 1 ; 0.5 1)	2.25	2, 0.5	26.565°, 135°
Shear (1 1 ; 0 1)	0	1 (double root)	Only one, at 0°
60° rotation	−3	$0.5 \pm 0.866025\,i$	None

For a symmetric matrix the eigendirections are orthogonal (the reason is in § 06). The shear has only one eigendirection and cannot be diagonalised. The rotation has no real eigendirection, and $\det[v, Av]$ is $\sin 60° \approx 0.866$ in every direction.

Detour: rotations have no real eigenvectors

A subsection that may be skipped.

The eigenvalues of $R(\theta)$ are $e^{\pm i\theta}$, and the eigenvectors $(1, \mp i)$. As a metaphor, a rotation of the plane splits, in the world of complex numbers, into two "multiplications by a phase". To be exact, changing the basis of $\mathbb{C}^2$ to $(1,-i)$, $(1,i)$ gives $R(\theta) = \operatorname{diag}(e^{i\theta}, e^{-i\theta})$. Over the complex numbers the characteristic equation always has a root (the fundamental theorem of algebra), so every square matrix has eigenvalues. The two angles of a rotation in four dimensions (the article on four dimensions) are the two pairs of eigenvalues $e^{\pm i\alpha}$, $e^{\pm i\beta}$.

05

Inner products and adjoints

The inner product measures lengths and angles. If we take $\sum_k u_k v_k$ directly with complex entries, the "squared length" of $v = (1, i)$ becomes $1 + i^2 = 0$. So we put a complex conjugate on one side ($z\,z = |z|^2 \geq 0$).

$$\langle u, v\rangle = \sum_k u_k\, v_k, \qquad \|v\| = \sqrt{\langle v, v\rangle}, \qquad \langle u, v\rangle = \langle v, u\rangle$$

This book follows the physicists' convention and takes it linear in the second argument ($\langle cu, v\rangle = c\langle u, v\rangle$). Many mathematics textbooks use the opposite convention (for example Axler's definition 6.2), and the conclusions do not change. If $\langle u, v\rangle = 0$ the vectors are orthogonal, and a basis of mutually orthogonal vectors of length 1 is an orthonormal basis.

The **adjoint** $A^\dagger$ of A is the matrix obtained by transposing and taking the complex conjugate of the entries (in mathematics A^*). It moves A to the other side inside the inner product.

$$(A^\dagger)_{ij} = A_{ji}, \qquad \langle u, Av\rangle = \langle A^\dagger u, v\rangle, \qquad (AB)^\dagger = B^\dagger A^\dagger$$

06

Hermitian matrices

A matrix with $H^\dagger = H$ is a **Hermitian matrix**; with real entries it is a symmetric matrix. In quantum mechanics measured values (energy and so on) are represented by Hermitian operators (2-05). The key is two properties.

The eigenvalues are real: if $Hv = \lambda v$ and $\|v\| = 1$, then $\lambda = \langle v, Hv\rangle = \langle Hv, v\rangle = \lambda$.
Eigenvectors of different eigenvalues are orthogonal: if $Hu = \mu u$, then $\lambda\langle u, v\rangle = \langle u, Hv\rangle = \langle Hu, v\rangle = \mu\langle u, v\rangle = \mu\langle u, v\rangle$ (μ is real), and since $\mu \neq \lambda$, $\langle u, v\rangle = 0$.

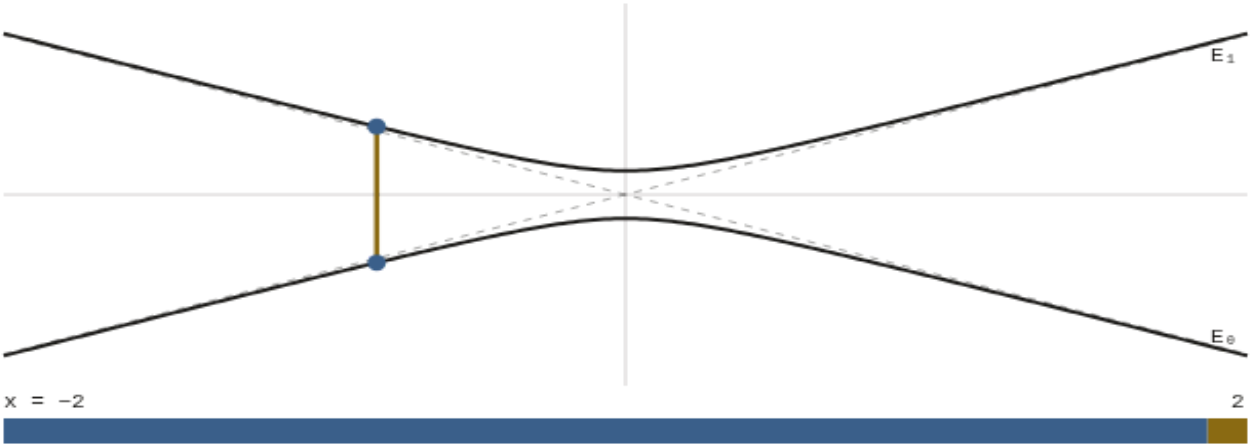
Pushing this further gives the **spectral theorem** Known . A Hermitian matrix has an orthonormal basis of eigenvectors, and the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$ are real. With

the unitary matrix U (§ 07) that lines up the basis, $U^\dagger H U$ becomes real diagonal.

The difference between the smallest eigenvalue and the next is called the **gap**. In quantum mechanics the smallest eigenvalue E_0 is the energy of the ground state, and $E_1 - E_0$ the energy needed for the first excitation. In finite dimensions the gap can always be computed. The problem is in infinite dimensions: whether the gap stays positive after the limit (1-10, 4-11).

Example: the matrix of § 02 has $\operatorname{tr} H = 0$ and $\det H = -(x^2 + y^2 + z^2)$, so its eigenvalues are $\pm\sqrt{x^2 + y^2 + z^2}$, and the gap is 0 only for $x = y = z = 0$. Figure 2 is the case $y = c$, $z = 0$.

Figure 2 The two eigenvalues of $H(x) = \left(\begin{smallmatrix} x & c \\ c & -x \end{smallmatrix}\right)$. Raising the off-diagonal entry c opens the crossing and leaves a gap; moving x swaps the content of the lower eigenvector (the band; blue is the weight of e_1)



Computed The curves are drawn on the spot from $\pm\sqrt{x^2 + c^2}$, and the table gives the values of NumPy's Hermitian eigenvalue computation (the two agree to the digits shown). Taking x to $-x$ swaps the weights. Clicking a row of the table puts that x and $c = 0.3$ into the figure.

x $(c = 0.3)$	$E_0 = -E_1$	Gap	Lower eigenvector $e_1 : e_2$
0.0	−0.300000	0.600000	0.5000 : 0.5000
0.2	−0.360555	0.721110	0.2226 : 0.7774
0.5	−0.583095	1.166190	0.0713 : 0.9287
1.0	−1.044031	2.088061	0.0211 : 0.9789

At $c = 0$ the lines $\pm x$ cross. If $c \neq 0$ the curves only come close, a minimal gap $2c$ remains, and past that point the content of the lower eigenvector swaps.

Detour: levels avoid each other

A subsection that may be skipped.

For the eigenvalues of a 2×2 matrix to coincide, three conditions $x = y = z = 0$ are needed if it is Hermitian, and two if it is real symmetric. A family moving with one parameter passes through that point only in special cases, so in a general family without symmetry the eigenvalue curves usually do not cross Known . When there is a symmetry and the matrix splits into blocks, levels in different blocks can cross. The counting for general $n \times n$ is said to go back to von Neumann and Wigner (1929). As a metaphor, the levels are said to "repel". To be exact, no force acts; the matrices where they coincide form a thin set whose dimension is lower by the number of conditions.

07

Unitary matrices

A matrix that preserves the inner product is a **unitary matrix**. It is a transformation that changes neither lengths nor angles; with real entries it is an orthogonal matrix (rotations and reflections).

$$U^\dagger U = I \qquad \Longleftrightarrow \qquad \langle Uu, Uv\rangle = \langle u, v\rangle \ \ (\text{for all } u, v)$$

Left to right follows from $\langle Uu, Uv\rangle = \langle u, U^\dagger U v\rangle$. It is the same as the columns forming an orthonormal basis, and $U^{-1} = U^\dagger$. The eigenvalues are $|\lambda| = 1$, $|\det U| = 1$. Products and inverses are unitary too, so the whole set $\mathrm{U}(n)$ and the part $\mathrm{SU}(n)$ with $\det U = 1$ form "groups" (1-02).

Example: $\mathrm{U}(1) = \{e^{i\alpha}\}$ is the unit circle. $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$, made from complex numbers with $|a|^2 + |b|^2 = 1$, is an element of $\mathrm{SU}(2)$ (that every element has this form is in 1-02 § 05).

Unitary matrices can also be diagonalised. The basis is the complex spectral theorem Known . A matrix satisfying $AA^\dagger = A^\dagger A$ (a **normal matrix**) can be diagonalised by a unitary matrix. A unitary matrix has $U^\dagger U = UU^\dagger = I$, so it is one (Hermitian matrices are too), and it can be written $U = V\,\mathrm{diag}(e^{i\theta 1}, \dots, e^{i\theta n})V^\dagger$ (V unitary, θ_k real).

Unitary matrices appear in gauge theory because matter fields take values in $\mathbb{C}^N$ and the squared length $\psi^\dagger\psi$ appears in the action. Gauge transformations are chosen to preserve it (3-01). On the lattice, the gauge field itself is a unitary matrix on each edge (5-01).

08

The matrix exponential

The series of the exponential uses only multiplication and addition, so it can be used for matrices as it is.

$$e^X = I + X + \frac{X^2}{2!} + \frac{X^3}{3!} + \cdots = \sum_{k \ge 0} \frac{X^k}{k!}$$

Convergence. The operator norm $\|X\| = \max_{\|v\|=1} \|Xv\|$ (the largest factor by which it multiplies a vector of length 1) satisfies $\|XY\| \le \|X\|\|Y\|$, so each term is at most $\|X\|^k/k!$ and the sum at most $e^{\|X\|}$. The error of the partial sum $S_N(X)$ up to $k = N$ is as follows.

$$\left\| e^X - S_N(X) \right\| \le \frac{\|X\|^{N+1}}{(N+1)!}\, e^{\|X\|}$$

Properties. (a) $\frac{d}{dt} e^{tX} = X e^{tX}$. Hence $x(t) = e^{tX}x_0$ is the solution of $x' = Xx$, $x(0) = x_0$. (b) $e^{(s+t)X} = e^{sX}e^{tX}$, in particular $(e^X)^{-1} = e^{-X}$. (c) $P e^X P^{-1} = e^{PXP^{-1}}$. (d) $\det e^X = e^{\operatorname{tr} X}$. If diagonalisable, $\prod_k e^{\lambda_k} = e^{\sum_k \lambda_k}$ follows from (c); in general it can be shown by triangularisation Known . Every complex square matrix X can be brought to upper triangular form $V^\dagger X V$ by a unitary matrix V (Schur's theorem), and the exponential of an upper triangular matrix is upper triangular with diagonal entries e^{λ_k}.

Caution. If $XY = YX$, then $e^{X+Y} = e^X e^Y$, but not in general. The A, B of § 03 have $A^2 = B^2 = 0$, so $e^A = I + A$, $e^B = I + B$, and

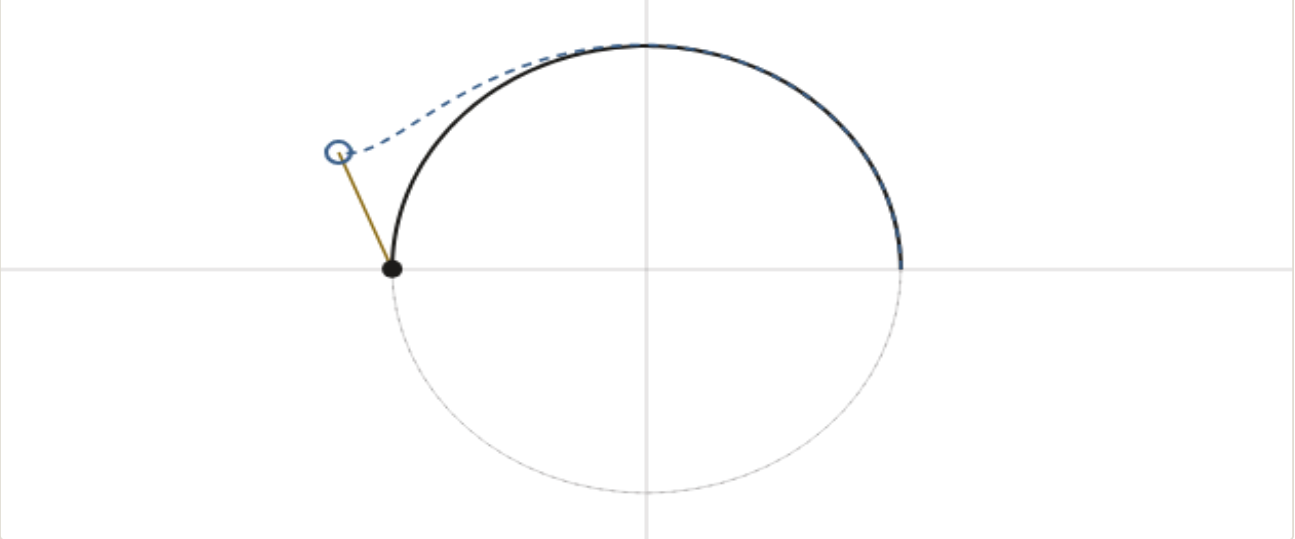
$$e^A e^B = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \qquad e^B e^A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix},$$

$$e^{A+B} = \begin{pmatrix} \cosh 1 & \sinh 1 \\ \sinh 1 & \cosh 1 \end{pmatrix} \approx \begin{pmatrix} 1.5431 & 1.1752 \\ 1.1752 & 1.5431 \end{pmatrix}.$$

Computed All three differ. A formula that writes the difference in terms of commutators is the subject of 1-03. (d) held to within a relative error of 5.7×10^{-14} for 1000 random 4×4 complex matrices.

Example: $J = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ has $J^2 = -I$, so splitting into even and odd orders gives $e^{tJ} = \cos t\, I + \sin t\, J = R(t)$, that is, a rotation. In general, if $X^2 = -\omega^2 I$ then $e^{tX} = \cos(\omega t)\, I + \frac{\sin(\omega t)}{\omega} X$.

Figure 3 $X = \left(\begin{smallmatrix} a & -1 \\ 1 & a \end{smallmatrix}\right)$, $x_0 = (1, 0)$. The solid line is $e^{tX}x_0$, and the blue dashed line the partial sum $S_N(tX)x_0$. Increasing N makes the dashed line lie on the solid line, and lengthening t increases the N needed. With $a \neq 0$ the circle becomes a spiral (the dotted line has radius e^{at})



Computed The partial sums in the figure add up the series on the spot. The table gives the difference, measured in the operator norm, between the double-precision partial sums and the rotation matrix e^{tJ} ($a = 0$). Clicking a number cell of the table puts that N and t ($a = 0$) into the figure.

N	$t = 1$	$t = \pi$	$t = 2\pi$	$t = 4\pi$
4	8.3×10⁻³	2.3	57	1.0×10³
8	2.7×10⁻⁶	7.9×10⁻²	36	1.3×10⁴
16	2.8×10⁻¹⁵	7.8×10⁻⁷	9.9×10⁻²	1.1×10⁴
30	1.1×10⁻¹⁶	2.5×10⁻¹⁶	6.6×10⁻¹⁰	1.4
Smallest N with error < 10^{-6}	9	16	26	43

When t is large the terms $t^k/k!$ swell for a while, so the N needed grows with t. At $t = \pi$, $N = 16$, the actual error is 7.8×10^{-7} against the error bound 1.8×10^{-5} above.

Detour: an example where summing the series as it is fails

A subsection that may be skipped.

A formula correct as a definition is not necessarily good as a method of computation. The example in the survey of Moler and Van Loan (1978; reprinted with an addendum in 2003) Known is $A = \begin{pmatrix} -49 & 24 \\ -64 & 31 \end{pmatrix}$ (eigenvalues −1 and −17). The entries of e^A are of order 1, yet the intermediate terms of the series swell to about 9.3×10^6 and cancel each other. Summing on this machine until the value stops changing Computed , in single precision (about 7 digits) the lower-right entry, which should be 1.1036, came out as −0.8392 after 61 terms, and in double precision the error was 5.7×10^{-9} after 74 terms. The paper says that the problem lies not in truncating the series but in truncating the arithmetic. Practical programs (SciPy's `expm` and others) use scaling and squaring: divide the matrix by a power of 2, approximate, and then square repeatedly to get back.

```
09
```

Joining Hermitian and unitary

Hermitian matrices and unitary matrices are connected by the exponential. As a metaphor, a Hermitian matrix is "how to turn", and a unitary matrix is "the result of turning". To be exact, it is as follows.

$$H^\dagger = H \implies U = e^{iH} \text{ is unitary}, \qquad \det e^{iH} = e^{i \operatorname{tr} H}$$

Taking the adjoint term by term gives $(e^{iH})^\dagger = e^{-iH}$. iH and $-iH$ commute, so $U^\dagger U = e^0 = I$ follows from § 08 (b), and the determinant from (d). If $\operatorname{tr} H = 0$, then $\det U = 1$ and $U \in \mathrm{SU}(n)$. Conversely, every unitary matrix can be written as e^{iH} by setting $H = V \operatorname{diag}(\theta_k) V^\dagger$ from the diagonalisation of § 07 Known (θ_k can be chosen only up to integer multiples of 2π, so H is not unique).

Example: $H = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ has $H^2 = I$, so $e^{i\theta H} = \cos\theta\, I + i \sin\theta\, H$, with determinant $\cos^2\theta + \sin^2\theta = 1$. Computed For 1000 random 3×3 Hermitian matrices, $\|U^\dagger U - I\| \le 1.6 \times 10^{-15}$, and with trace 0, $|\det U - 1| \le 1.1 \times 10^{-15}$.

This correspondence is used in two places. Time evolution in quantum mechanics $e^{-itH/\hbar}$ is unitary and preserves total probability (2-05). In gauge theory, sending the space of § 02 widened to $N \times N$ (traceless $N \times N$ Hermitian matrices, times i) through the exponential lands in $\mathrm{SU}(N)$; the former is the space of values of the gauge field, and the latter the space of gauge transformations (1-03, 1-04). Restricted to real entries, it takes the form "the e^K of an antisymmetric matrix K is a rotation".

Attaching "times i" to the values of the gauge field is the geometers' way of writing (1-07). The physicists' way writes the gauge field as a Hermitian matrix as it is, and the two are carried into each other by $A_{\text{geom}} = -iA_{\text{phys}}$ (the form in chapters that write the coupling constant separately is in the table in 3-02).

What this chapter can say

What is said	What is not said
Known A finite-dimensional Hermitian matrix has real eigenvalues and an orthonormal basis of eigenvectors. Unitary matrices can likewise be diagonalised, with eigenvalues of absolute value 1.	Whether the same holds in infinite dimensions. Some operators have no eigenvectors (1-10).
Known e^X always converges, and $\det e^X = e^{\operatorname{tr} X}$. If H is Hermitian, e^{iH} is unitary, and every unitary matrix has this form.	A formula that expresses the difference between e^{X+Y} and $e^X e^Y$ in terms of commutators (1-03).

What is said	What is not said
Computed The numbers in the figures and tables, and the checks of identities with random matrices.	Checks with random numbers are not proofs.
Known In finite dimensions, the gap between the smallest eigenvalue and the next can always be computed.	Whether, for operators in infinite dimensions, the gap stays positive after the limit. The mass gap problem has this form (4-11), and this chapter says nothing about it.

Sources and reproduction

Item	Kind	Source / tool
Convention for inner products, the spectral theorem (real, complex, normal matrices), triangularisation (Schur's theorem)	Known	S. Axler, *Linear Algebra Done Right*, 4th ed., Springer (Undergraduate Texts in Mathematics), 2024. Definitions 6.2, 7.29, 7.31, 6.38. Text checked in the open-access version.
An example of computation by series failing; scaling and squaring	Known	C. Moler, C. Van Loan, SIAM Review 45 (2003) 3–49 (Methods 1 and 3, text checked). First published in SIAM Review 20 (1978) 801–836 (bibliographic data only). The method of SciPy's `expm` is as described in the SciPy 1.11 documentation (A. H. Al-Mohy, N. J. Higham, SIAM J. Matrix Anal. Appl. 31 (2009) 970–989, bibliographic data only).
The number of conditions needed for eigenvalues to cross	Known	J. von Neumann, E. Wigner, Physikalische Zeitschrift 30 (1929) 467–470 (bibliographic data only).
Tables of Figures 1 and 2	Computed	`eig` and `eigh` of Python 3 and NumPy 1.26 on this machine. For 1000 traceless 2×2 Hermitian matrices, the difference from the formula $\pm\sqrt{x^2+y^2+z^2}$ was within 2.2×10^{-15}.
Table of Figure 3, recomputation of the detour in § 08, identities of § 08 and § 09	Computed	Partial sums computed in double and single precision (half precision cannot compute, as intermediate terms exceed its upper limit). Exact values are the rotation matrix and $P\,\mathrm{diag}(e^{-1}, e^{-17})P^{-1}$ ($P = \left(\begin{smallmatrix}1 & 3\\ 2 & 4\end{smallmatrix}\right)$). SciPy 1.11's `expm` and random matrices (fixed seed, 1000 each).

Next chapter: 1-02 What a group is — it starts from the fact that the set of unitary matrices forms a "group", and goes on to the relation between $\mathrm{SO}(3)$ and $\mathrm{SU}(2)$.

What a group is — treating symmetries as "things that can be multiplied"

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — The rules of a group, and examples. SO(3), the group of rotations in three dimensions, and SU(2), which covers it twice (it does not return at 360° but returns at 720°). "Representations", which write a group as matrices. How to read "compact simple gauge group" in the problem statement.

Prerequisite chapters — 1-01 Vector spaces and linear maps (determinants, unitary matrices, the matrix exponential). Lie algebras are treated in 1-03, and the Pauli matrices and SU(3) in 1-04.

Said first — This chapter is preparation, and shows nothing about the mass gap. Theorems cited without proof carry the "Known" label.

Order of this chapter

1. From symmetry to groups — the six operations of an equilateral triangle and their multiplication table (a figure with choices)
2. The definition of a group, and examples — commutative and non-commutative, groups of matrices, compactness
3. Homomorphisms, normal subgroups, quotients — tools for comparing groups, the centre
4. The rotation group SO(3) — axes, the mismatch from the order of rotations (moving figure)
5. SU(2) — the three-dimensional sphere, half angles
6. The double cover SU(2) → SO(3) — returning at 720° (moving figure)
7. Detour: the shape of SO(3) and the 720° loop — loops that shrink and loops that do not (moving figure)
8. A first look at representations — integer labels, irreducibility, spin
9. The group inside Yang–Mills — gauge group, compact, simple

01

From symmetry to groups

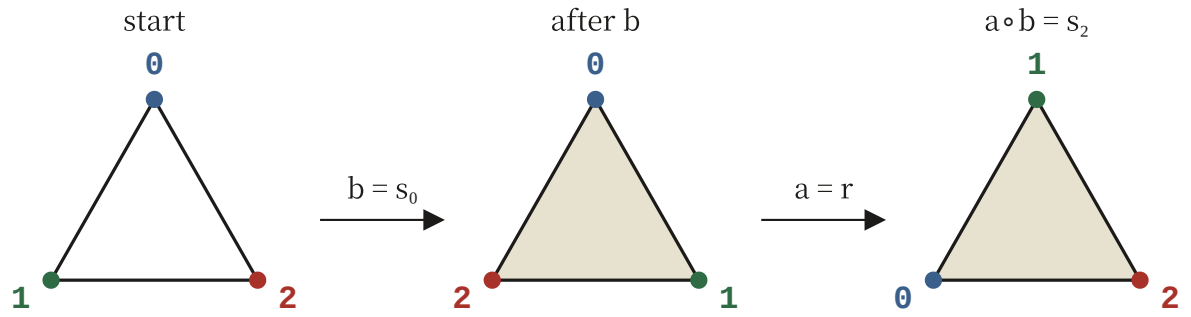
There are six ways to put an equilateral triangular plate back **so that it fits exactly over its original place**: three rotations by 0°, 120° and 240°, and three flips about an axis through a vertex. **Doing two in succession gives one of the six again**, and there is an operation that does nothing and operations that undo. A group is the word for exactly these properties and nothing more.

Label the vertices 0, 1, 2 anticlockwise, and write e (do nothing), r (rotate by 120°), r^2 (240°), and s_k (flip about the axis through vertex k). The product ab is read **"first b, then a"** (the direction of composition of maps $a \circ b$).

$$r \circ s0 = s2, \qquad s0 \circ r = s1$$

Swapping the order changes the result. That **the order of multiplication matters** is the protagonist all the way to the end of this book.

Figure 1 Choosing a cell of the table (row a, column b) lines up the triangle as "start → after b → after a∘b". A coloured triangle is in the flipped state.



`r∘s₀ = s₂ (first s₀, then r). In the opposite order s₀∘r = s₁ (different)`

a \ b	e	r	r²	s₀	s₁	s₂
e	e	r	r^2	s_0	s_1	s_2
r	r	r^2	e	s_2	s_0	s_1
r²	r^2	e	r	s_1	s_2	s_0
s₀	s_0	s_1	s_2	e	r	r^2
s₁	s_1	s_2	s_0	r^2	e	r
s₂	s_2	s_0	s_1	r	r^2	e

Computed The values in the table were computed on this machine as compositions of permutations of the vertices. The rules of a group were also checked for every combination.

That each of the six elements appears exactly once in each row and each column reflects the fact that $ax = b$ has exactly one solution $x = a^{-1}b$, that is, that **division is possible**. Of the 36 ordered pairs, $ab \neq ba$ for 18 Computed .

02

Definition of a group, and examples

Definition If there is a set G and a rule (product) $(a, b) \mapsto ab$ that makes one element of G from two elements, satisfying the following, G is called a **group**.

1. Associative law: $(ab)c = a(bc)$.
2. Identity: there is e with $ea = ae = a$ for all a.
3. Inverses: for each a there is a^{-1} with $aa^{-1} = a^{-1}a = e$.

If $ab = ba$ for every pair, the group is **commutative** (an abelian group); otherwise it is **non-commutative**. The "non-abelian" of "non-abelian gauge theory" has this meaning. The second half of the table below is **groups of matrices**, where $U^\dagger$ is the transpose with complex conjugation (1-01).

Group	Elements (product)	Commutative?	Compact?	Real parameters
$\mathbb{Z}_N$	Remainders on division by N (addition)	Commutative	Finite	0
D_3	The six operations of an equilateral triangle	Non-commutative	Finite	0
$\mathbb{R}$	Real numbers (addition)	Commutative	No	1
U(1)	$e^{i\theta}$ (multiplication)	Commutative	Yes	1
GL(n, $\mathbb{R}$)	Invertible real matrices	Non-commutative	No	n^2
O(n)	Real matrices with $R^\mathsf{T}R = I$	Non-commutative	Yes	$n(n-1)/2$
SO(n)	Those of O(n) with $\det R = 1$	Non-commutative for $n \geq 3$	Yes	$n(n-1)/2$
U(n)	Complex matrices with $U^\dagger U = I$	Non-commutative	Yes	n^2
SU(n)	Those of U(n) with $\det U = 1$	Non-commutative	Yes	$n^2 - 1$

Groups of matrices are $n \geq 2$. SO(2) is a commutative group of the same form as U(1).

For groups of matrices, **compact** means **bounded** and **closed** (the limit of a sequence of elements is again an element) in the space of entries Known . U(n) is bounded because each column has length 1, and $U^\dagger U = I$ is preserved in the limit. GL(n, $\mathbb{R}$) is not bounded. "Real parameters" is the number of real numbers that specify an element

(the dimension of the group, 1-03); for SO(3) it is 2 for the direction of the axis and 1 for the angle.

03

Homomorphisms, normal subgroups, quotients

A **subgroup** is a subset closed under products and inverses ($\{e, r, r^2\} \subset D_3$, SU(n) $\subset$ U(n), and so on). A **homomorphism** is a map $\varphi : G \to H$ satisfying $\varphi(ab) = \varphi(a)\varphi(b)$, and the set of elements sent to the identity is called the **kernel** $\ker\varphi$.

$$\det : \mathrm{U}(n) \to \mathrm{U}(1), \qquad \det(UV) = \det U \, \det V$$

The kernel of this homomorphism is $\mathrm{SU}(n)$. The kernel of $\theta \mapsto e^{i\theta}$ is $2\pi\mathbb{Z}$. The map sending the rotations of D_3 to +1 and the flips to −1 is also a homomorphism. A subgroup N is **normal** if $gNg^{-1} = N$ for all g, and a kernel is always normal. Then the **quotient group** G/N, which identifies g and gn ($n \in N$), can be formed, and for a surjective homomorphism φ the following holds (the homomorphism theorem) Known .

$$G/\ker\varphi \;\cong\; H$$

Example: $\mathbb{R}/2\pi\mathbb{Z} \cong \mathrm{U}(1)$ (an angle is a real number with differences of integer multiples of 2π forgotten). The **centre** $Z(G)$ (the elements that commute with every element) is also a normal subgroup, and the centre of SU(n) is the n elements ωI ($\omega^n = 1$), of the same form as $\mathbb{Z}_n$ Known . It appears again in 17-02, which treats confinement.

04

The rotation group SO(3)

A rotation in three dimensions is a linear map that preserves the inner product ($R^{\mathsf{T}}R = I$) and makes no mirror image ($\det R = 1$).

$$\mathrm{SO}(3) = \{\, R \in M_3(\mathbb{R}) : R^{\mathsf{T}}R = I,\ \det R = 1 \,\}$$

Every rotation has an axis

Every element of SO(3) has a line that does not move (Euler's rotation theorem) Known . A real orthogonal matrix is unitary as a complex matrix (1-01), so its eigenvalues have absolute value 1, and their product is $\det R = 1$. They are roots of a cubic with real coefficients, so complex eigenvalues come in pairs $\lambda, \bar{\lambda}$ (product $|\lambda|^2 = 1$), and the remaining real eigenvalue is 1. If all are real, they are three values ± 1 with

product 1, so again they include 1. The eigenvector of that eigenvalue is the axis. Odd dimension matters here: some rotations in four dimensions have no fixed point other than the origin. With axis n and angle θ it has the following form (Rodrigues' formula) Known .

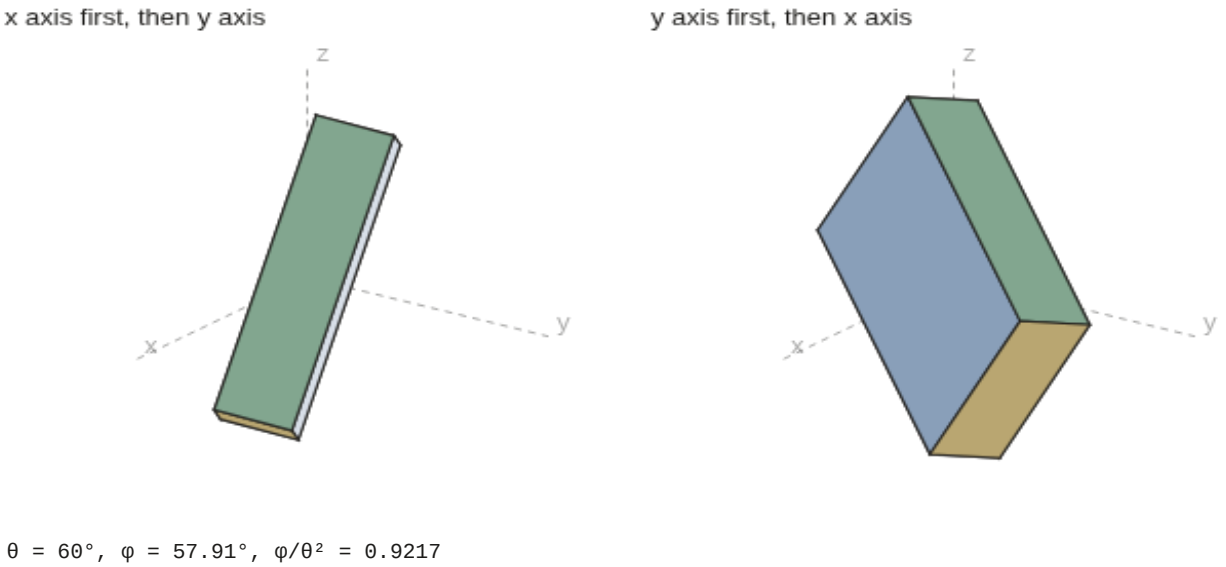
$$R(\theta, n)\, x = \cos\theta\, x + \sin\theta\, (n \times x) + (1 - \cos\theta)(n \cdot x)\, n$$

The trace is $\operatorname{tr} R(\theta, n) = 1 + 2\cos\theta$, from which the angle can be read.

The mismatch from the order of rotations

Rotate by the same angle θ about the x axis and the y axis. "x first" is $A = R_y(\theta) R_x(\theta)$, "y first" is $B = R_x(\theta) R_y(\theta)$. The mismatch $D = AB^{-1}$ is also a rotation, so we measure the difference by its angle ϕ.

Figure 2 Changing the angle θ with the slider changes the difference between the final orientations of the two boxes. "Rotate in turn" replays the two rotations in order.



θ = 60°, φ = 57.91°, φ/θ² = 0.9217

θ	φ	φ / θ² (in radians)
1°	0.0175°	1.0000
10°	1.7409°	0.9975
30°	15.3639°	0.9781
45°	33.6842°	0.9531
60°	57.9100°	0.9217
90°	120.0000°	0.8488
180°	0°	0

Computed The table was computed on this machine from products of rotation matrices. The φ in the figure is computed on the spot from the same formula.

For small θ, $\varphi \approx \theta^2$, and the axis of D is almost in the −z direction ((−0.087, 0.087, −0.992) at θ = 10°) Computed . **The mismatch is second order in θ**, and to first order rotations look commutative. Extracting this gives the commutator $[X, Y] = XY - YX$ of 1-03. At θ = 180° both are diagonal matrices and commute.

05

SU(2)

Next is a group of complex 2×2 matrices.

$$\mathrm{SU}(2) = \{\, U \in M_2(\mathbb{C}) : U^\dagger U = I,\ \det U = 1 \,\}$$

The two columns are orthonormal. If the first column is (a, b), then $|a|^2 + |b|^2 = 1$, and the second column is a vector of length 1 orthogonal to it, $e^{i\chi}(-\bar{b}, \bar{a})$, with $\det U = e^{i\chi} = 1$.

$$U = \begin{pmatrix} a & -\bar{b} \\ b & \bar{a} \end{pmatrix}, \qquad |a|^2 + |b|^2 = 1$$

The real and imaginary parts of a, b form a point on the unit sphere of $\mathbb{R}^4$, and **as a space, SU(2) is the three-dimensional sphere S^3** Known . Using the **Pauli matrices** (1-04), every element can be written with a unit vector n and a real number θ.

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Writing $n \cdot \sigma = n_1\sigma_1 + n_2\sigma_2 + n_3\sigma_3$, it has the following form.

$$U(\theta, n) = \cos\tfrac{\theta}{2}\, I - i \sin\tfrac{\theta}{2}\, n \cdot \sigma = \exp\Big(-\tfrac{i\theta}{2}\, n \cdot \sigma \Big)$$

In components, $a = \cos\frac{\theta}{2} - in_3 \sin\frac{\theta}{2}$, $b = (n_2 - in_1)\sin\frac{\theta}{2}$. The equality on the right follows from $(n \cdot \sigma)^2 = I$ (from $\sigma_j\sigma_k + \sigma_k\sigma_j = 2\delta_{jk} I$ and $|n| = 1$; how to check it is in 1-04) and the matrix exponential (1-01). At θ = 2π it is $U = -I$ whatever n is, and only at θ = 4π does it become $U = I$. This **half angle** is the protagonist of the next section.

06

The double cover SU(2) → SO(3)

Identify vectors of $\mathbb{R}^3$ with traceless Hermitian matrices.

$$x \ \longleftrightarrow\ X = x \cdot \sigma = \begin{pmatrix} x_3 & x_1 - ix_2 \\ x_1 + ix_2 & -x_3 \end{pmatrix}, \qquad \det X = -|x|^2$$

$UXU^\dagger$ is also Hermitian and traceless, so it is the matrix of some vector x'. This fixes a linear map $\rho(U) : x \mapsto x'$ Known .

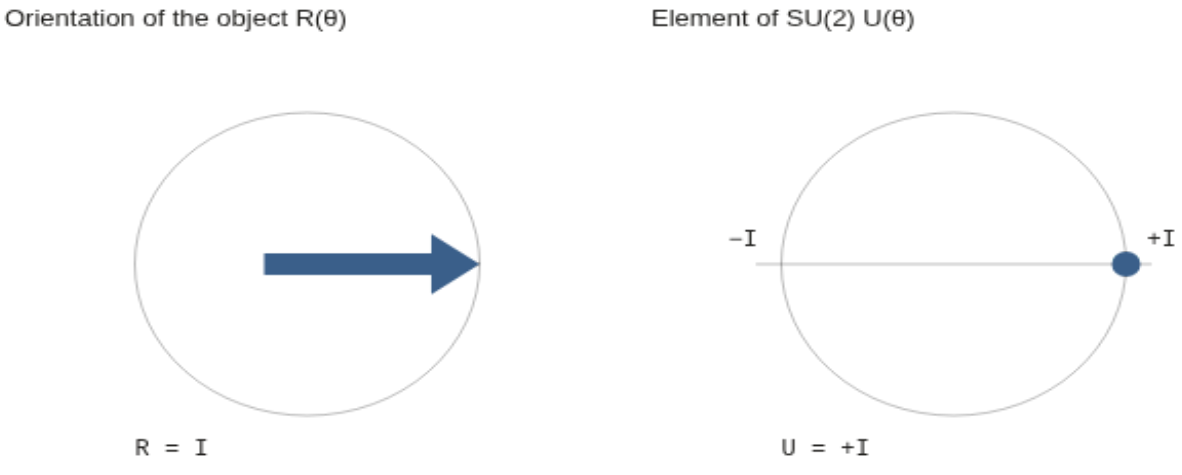
1. **It preserves length**: $|x'| = |x|$ follows from $\det(UXU^\dagger) = \det X$. SU(2) is connected and ρ is continuous, so the determinant cannot move from +1, and $\rho(U) \in \mathrm{SO}(3)$.
2. **It is a homomorphism**: $\rho(UV) = \rho(U)\rho(V)$ follows from $(UV)X(UV)^\dagger = U(VXV^\dagger)U^\dagger$.
3. **Its kernel is $\pm\mathbf{I}$**: if $\rho(U) = I$, then U commutes with σ_3 (it is diagonal) and with σ_1 (its diagonal entries are equal), so $U = \lambda I$, $\det U = \lambda^2 = 1$.
4. **It is surjective, and the angle is halved**: $\rho(U(\theta, n)) = R(\theta, n)$. Every rotation has the form $R(\theta, n)$ (the axis theorem of 04), so $U(\theta, n)$ is its preimage, and the map is surjective.

A simple case of 4: if n is the z axis, $U = \mathrm{diag}(e^{-i\theta/2}, e^{i\theta/2})$ gives $U\sigma 1 U^\dagger = \cos\theta\,\sigma 1 + \sin\theta\,\sigma 2$. **The two half angles multiply together to give the angle θ.** From the homomorphism theorem,

$$\mathrm{SO}(3) \;\cong\; \mathrm{SU}(2)/\{\pm I\}$$

Two elements $\pm U$ correspond to one rotation, so it is called a **double cover**. A 360° rotation goes in SU(2) only from I to $-I$, and returning takes 720°.

Figure 3 Changing the angle θ with the slider, the object on the left (an arrow seen from above) returns to its original state at 360°, while the point on the right goes round its circle at half the speed and returns at 720°. The point on the right is the two coefficients $(\cos\frac{\theta}{2}, \sin\frac{\theta}{2})$ of $U(\theta) = \cos\frac{\theta}{2} I - i\sin\frac{\theta}{2}\sigma 3$, one of the great circles in S^3.



θ	$\mathrm{tr}\,R(\theta) = 1 + 2\cos\theta$	$\mathrm{tr}\,U(\theta) = 2\cos\frac{\theta}{2}$	R	U
0°	3	2	I	$+I$
180°	−1	0	180° about the z axis	$-i\sigma 3$

θ	$\operatorname{tr} R(\theta) = 1 + 2\cos\theta$	$\operatorname{tr} U(\theta) = 2\cos\frac{\theta}{2}$	R	U
360°	3	−2	I	$-I$
540°	−1	0	180° about the z axis	$+i\sigma_3$
720°	3	2	I	$+I$

Computed Computed on this machine by building ρ(U). Orthogonality (maximum deviation 1.2×10⁻¹⁵), the homomorphism property and ρ(−U) = ρ(U) were checked for 20,000 random U, and agreement with Rodrigues' formula (maximum deviation 1.0×10⁻¹⁵) for 2,000 pairs (θ, n).

07

Detour: the shape of SO(3) and the 720° loop

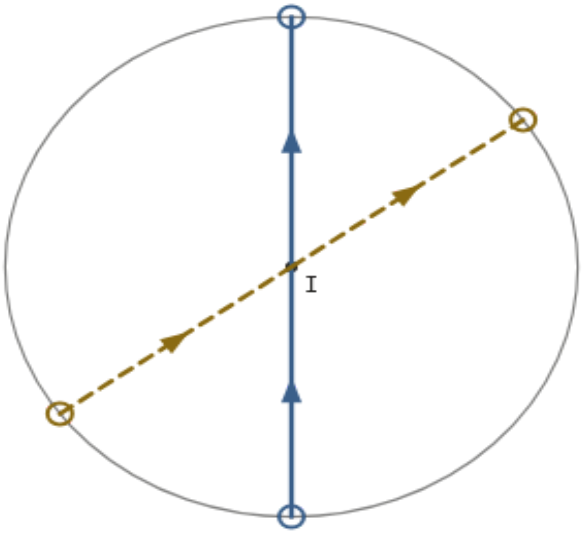
A section that may be skipped. It is not a prerequisite for later chapters. It looks at the double cover as a "shape".

Represent the rotation by angle θ ($0 \le \theta \le \pi$) about the axis n by the point θn in a ball of radius π. Rotations by π about n and about −n are the same, so SO(3) is **a ball of radius π with antipodal points of its surface glued together** Known .

The path 0°→360° about the z axis is a loop that goes from the centre to the north pole, comes out of the south pole, which is the same point, and returns to the centre; it **cannot be shrunk to a point** Known . Lifted to SU(2), it is a path from I to $-I$. Deforming the loop continuously moves the end point of the lift continuously, but the end point cannot jump between the two points $\pm I$. The lift of a one-point loop ends at I, so the loop cannot be deformed into one. The 720° loop, lifted, also ends at I, and does in fact shrink.

Figure 4 The first half of the slider tilts the axis of the second turn, and the second half shrinks the "there and back" path. Switching to "360° loop", however the axis is tilted, the path still crosses the rim once (even under general deformations, the parity of the number of crossings does not change).

First turn (solid) and second turn (dashed). The axis of the second turn is being tilted



Number of crossings of the rim 2, end point in SU(2) +I

Loop	Number of crossings of the rim	End point in SU(2)	Shrinks to a point?
360°	1 (odd)	$-I$	Does not shrink
720°	2, becoming 0 in the second half of the deformation (even)	$+I$	Shrinks

Known The disc is a cross-section of the ball, and two opposite points of the rim are the same rotation. It is a schematic that draws only the coordinates of the path; the parity of the number of crossings (when the path does not touch the rim) corresponds to the classification of loops.

Loops in SO(3) come in two kinds, "shrinks" and "does not shrink" (the fundamental group is $\mathbb{Z}_2$), and in S^3 every loop shrinks Known. The "plate trick", in which turning a plate held level on your hand through one revolution twists your arm and a second revolution untwists it, is often explained as the visible form of this fact. To be exact, the orientations of the points from shoulder to hand form a path in SO(3), and the motion of that path during the 720° turn is a deformation that shrinks the loop.

08

A first look at representations

Definition A (finite-dimensional) **representation** of a group G is a pair of a vector space V and a homomorphism $\pi : G \to GL(V)$ (continuous, for groups of matrices); that is, an assignment of an invertible matrix to each element such that $\pi(gh) = \pi(g)\pi(h)$ ($\dim V$ is the dimension). Examples: the trivial representation (everything to 1), writing D_3 by rotations and reflections of the plane, the action of SO(3) on $\mathbb{R}^3$ and of SU(2) on $\mathbb{C}^2$ (the **defining representation**), and the 3-dimensional representation ρ.

Representations of U(1) are labelled by integers

For $e^{i\theta} \mapsto e^{iq\theta}$ to be a representation of U(1), since θ and θ + 2π are the same element, $e^{2\pi i q} = 1$, that is, **q is an integer**. These exhaust the continuous 1-dimensional unitary representations Known. For $\mathbb{R}$, $x \mapsto e^{iqx}$ is a representation for any real q, so **what makes the labels integers is compactness**. The physical argument that if the gauge group of electromagnetism (2-03) is U(1) then charges line up as integer multiples of a unit rests on this fact (the title of C. N. Yang 1970 states this correspondence) Physics.

Irreducible and completely reducible

A subspace W with $\pi(g)W \subset W$ for all g is called **invariant**, and a representation with no invariant subspaces other than 0 and V is called **irreducible**. The action of SO(3) on $\mathbb{R}^3$ is irreducible; representing D_3 by permutations of the coordinates of $\mathbb{R}^3$ leaves the line (1, 1, 1) invariant, so it is not irreducible. **A finite-dimensional representation of a compact group splits into a direct sum of irreducible representations** (complete reducibility) Known. If the inner product is averaged over the whole group, every $\pi(g)$

preserves that inner product, and the orthogonal complement of an invariant subspace is also invariant. For a finite group the average is the sum over all elements divided by the number of elements. For a compact group, the same can be done with an invariant integral on the group (its existence is known).

Representations of SU(2) and the sign at 360°

SU(2) has exactly one finite-dimensional complex irreducible representation in each dimension, up to equivalence (change of basis) Known . The one of dimension m + 1 is Π_m, acting by $(\Pi_m(U)f)(z) = f(U^{-1}z)$ on the space of homogeneous polynomials of degree m in 2 variables, and in physics $j = m/2$ is called the **spin**. $-I$ acts as $f(-z) = (-1)^m f(z)$, so the ones that can be representations of SO(3) are **only those with m even (odd dimension)**.

m	Dimension	Spin j	Character $\chi(360^\circ)$	A representation of SO(3)?
0	1	0	+1	Yes
1	2	1/2	−2	No
2	3	1	+3	Yes
3	4	3/2	−4	No
4	5	2	+5	Yes

The character χ is the trace of the representation matrix, measured at $U(\theta, e3)$. At 720° every one returns to the value of the dimension Computed . For half-integer spin the sign flips at 360° — the $-I$ of Figure 3 appears, on the side of representations, as this sign.

09

The group inside Yang–Mills

The problem statement (0-03) begins with conditions on the group.

> Prove that for any compact simple gauge group G, a non-trivial quantum Yang–Mills theory exists on R^4 and has a mass gap $\Delta > 0$.

Jaffe and Witten's description of the problem says "at the level of classical theory, replace the gauge group U(1) of electromagnetism by a compact gauge group G". In a word, the **gauge group** is the group of symmetries that can be applied independently at each point of spacetime. To be exact, a choice of one element g(x) of G at each point x is a **gauge transformation**, and G is the group in which it takes values (3-01).

Simple. As an abstract group, "simple" means that the only normal subgroups are {e} and the group itself, and in this sense SU(2) is not simple (its centre {±I} is a normal subgroup). In the terminology of Lie groups the meaning is different: a connected non-commutative group is called "simple" if its Lie algebra is simple (1-03). The description of the problem does not define simple, but in the sense of abstract groups every SU(N) ($N \geq 2$) would be excluded (centre $\mathbb{Z}_N$), so this book reads it in the sense of Lie groups. SU(N) ($N \geq 2$) is simple in this sense (the list is in 1-04) Known , and the commutative U(1) is not included.

Compact. The same description states that with a compact gauge group and a compactified spacetime, Wilson's lattice approximation reduces the functional integral to a finite-dimensional integral. An element of G is put on each edge of the lattice and integrated over the whole of G (5-01), so it matters that the size of G is finite.

Representations. The same description says that in the theory of the strong force (QCD) the gauge group is SU(3), and quarks transform in the fundamental representation (this chapter's defining representation). The gauge field takes values in the Lie algebra of G. Under transformations that apply the same element at each point it transforms in the adjoint representation (1-04), and under general gauge transformations an extra term is added (3-02). The field strength transforms in the adjoint representation. The pure Yang–Mills theory the problem treats has no quarks, only the gauge field (3-07). Lattice models with G a finite group are treated in 5-10.

Next is 1-03 Lie groups and Lie algebras: the "second-order mismatch" is extracted as the commutator of the infinitesimal part of the group.

What this chapter can say

Item	What is said	What is not said
Definition of a group, and examples	Definition and examples. The table of D_3 Computed	The general theory of groups
SO(3)	The outline of the proof that an axis exists. The numbers of the mismatch $\varphi \approx \theta^2$ Computed	The reason it becomes θ^2 (the commutator) is in 1-03
Double cover	That ρ is a homomorphism to SO(3) with kernel $\pm I$. $\rho(U(\theta, n)) = R(\theta, n)$ by formula for the z axis, numerically for general n Computed	A proof by formula for general n Known

Item	What is said	What is not said
Loops in SO(3)	An explanation, by figure and by lifting, that the 360° loop does not shrink and the 720° loop does	A complete proof that the fundamental group is $\mathbb{Z}_2$ (the path lifting theorem) Known
Representations	Why representations of U(1) are labelled by integers. The sign of Π_m at 360° Computed	The classification of irreducible representations of SU(2) and a proof of complete reducibility Known . Quantisation of charge is physics Physics
Yang–Mills	How to read the conditions on the group in the problem statement (compact, simple)	Shows nothing about the mass gap or the construction of the theory

Sources and reproduction

Item	Kind	Source / tool
The problem statement, replacing U(1) by G, QCD and the fundamental representation, finite-dimensional integrals on the lattice. That simple is not defined	Primary source (text checked)	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory”, the Clay Mathematics Institute's description of the problem (PDF on the Institute's website; no date given; bibliographic details in the source table of 0-03)
Definition of compactness, the two-to-one SU(2) → SO(3), that SU(2) is simply connected and the universal cover of SO(3), complete reducibility (Prop. 5.17, using an invariant measure on the group), simplicity of abstract groups (note to Exercise 13 of Chapter 1), simplicity of Lie algebras (beginning of Chapter 7), Π_m	Known (text checked)	B. C. Hall, “An Elementary Introduction to Groups and Representations”, arXiv:math-ph/0005032 (2000)
The model of SO(3) as a ball of radius π with antipodal points of its surface glued, the fundamental group $\mathbb{Z}_2$, that the S^3 loop shrinks	Known (text checked)	A. Hatcher, *Algebraic Topology*, Cambridge University Press (2002). In the version made public by the author, § 3.D ($\mathbf{SO(3)} \approx \mathbb{RP}^3$), Example 1.43 ($\pi_1(\mathbb{RP}^n) \cong \mathbb{Z}_2$), Prop. 1.14 ($\pi_1(S^n) = 0$, $n \geq 2$)
"Simple" for Lie groups (connected, non-commutative, Lie algebra simple)	A terminological convention of this book	The definitions for abstract groups and Lie algebras were checked in Hall's notes above. A textbook stating this convention explicitly for Lie groups has not been checked
Quantisation of charge and compactness of the gauge group	Bibliographic data only (title and abstract)	C. N. Yang, “Charge Quantization, Compactness of the Gauge Group, and Flux Quantization”, Phys. Rev. D 1 (1970) 2360

Item	Kind	Source / tool
The table of $D3$ ($ab \neq ba$ for 18/36), $\varphi(\theta)$, the checks of the double cover, the signs and characters of Π_m (m = 0–6)	**Computed on this machine**	Python 3.12, NumPy 1.26. Compositions of permutations, products of rotation matrices, 20,000 uniform U on S^3 and 2,000 pairs (θ, n), representation matrices on homogeneous polynomials (maximum deviation from the homomorphism property 5.8×10^{-15})
Figures 1–4	Figures	Plain JavaScript. Figure 1 uses a precomputed table; Figures 2–4 compute the formulas for rotations and cos and sin on the spot

Next chapter to read: 1-03 Lie groups and Lie algebras.

Lie groups and Lie algebras — reading a curved group through the space of tangents at the identity

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That the set of "infinitesimal elements" of a group of matrices becomes a vector space (the Lie algebra), and that the mismatch in the group's multiplication survives as the commutator. How far the exponential map gets back into the group. That the structure constants and the Killing form reflect whether a group is simple and whether it is compact.

Prerequisite chapters — 1-02 What a group is (SO(3), SU(2), the double cover, homomorphisms), and its prerequisite 1-01 (the matrix exponential, Hermitian and unitary). We do not use manifolds and restrict to groups of matrices (general Lie groups can be read after 1-05).

Said first — This chapter is preparation, and shows nothing about the mass gap. Many theorems are cited without proof and carry the "Known" label. The numbers are computations with finite-dimensional matrices, and are not a substitute for proofs.

Order of this chapter

1. Looking at infinitesimal elements — a curved group and a flat tangent space
2. Matrix Lie groups and their Lie algebras — the exponential map, dimension
3. How far the exponential map reaches — one-to-one near the identity / Detour: elements of SL(2, ℝ) it does not reach (moving figure)
4. The commutator is the second-order term of the mismatch — parallel parking (moving figure), the θ^2 of 1-02
5. The structure called a Lie algebra — the Jacobi identity, SU(2) and SO(3), the adjoint
6. Structure constants — the two conventions of mathematics and physics
7. The Baker–Campbell–Hausdorff formula — commutators alone determine the product (moving figure)

8. The Killing form — seeing simple and compact through signs (a figure where you choose table rows)
9. The Lie algebra inside Yang–Mills — invariant inner products, the number of coupling constants

---

01

Looking at infinitesimal elements

The SO(3) and SU(2) of 1-02 are groups whose elements move continuously with real parameters. As spaces they are curved (SU(2) is the three-dimensional sphere), so we **take a first-order approximation only very near the identity**. As a metaphor, even on the curved surface of the earth, the map at your feet can be drawn on a plane. To be exact, the **Lie algebra** is the flat vector space that collects all velocities $g'(0)$ of curves $g(t)$ in the group with $g(0) = I$.

$$R(t) = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} = e^{tJ}, \qquad R'(0) = J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

The Lie algebra of the plane rotations SO(2) is the real multiples of J (1-01 § 08). The order of multiplication survives in the second-order term (§ 04), and from that information alone the multiplication near the identity can be recovered (§ 07). In Yang–Mills theory, **gauge transformations take values in the group, and the gauge field in the Lie algebra** (3-02).

02

Matrix Lie groups and their Lie algebras

Definition A subgroup G of $\mathrm{GL}(n, \mathbb{C})$ (the group of invertible complex $n \times n$ matrices) that is closed (if a sequence of elements converges to an invertible matrix, the limit is also an element of G) is called a **matrix Lie group** Known . Its **Lie algebra** is

$$\mathfrak{g} = \{\, X \in M_n(\mathbb{C}) : \text{for every real } t,\ e^{tX} \in G \,\}$$

and $X \mapsto e^X$ is called the **exponential map**. X is the velocity of the curve e^{tX} starting at the identity ((a) of 1-01 § 08). The condition is "for all t", and $e^X \in G$ alone is not enough. For example, for the group $G = \{I\}$ consisting only of the identity and $X = 2\pi J$, $e^X = I \in G$, but e^{tX} for $t = \frac{1}{4}$ is a 90° rotation and is not in G.

Example: U(n). Differentiating the condition $e^{tX\,\dagger}e^{tX} = I$ in $t = 0$ gives $X^\dagger = -X$. Conversely, in that case e^{tX} is unitary (1-01 § 09). From $\det e^{tX} = e^{t\,\mathrm{tr}\,X}$, the condition of determinant 1 becomes $\mathrm{tr}\,X = 0$.

$$\mathfrak{su}(n) = \{X : X^\dagger = -X,\ \mathrm{tr}\,X = 0\}$$

The Lie algebra is a **real** vector space (multiplying by i makes it Hermitian, which falls outside), and its dimension is called the **dimension of the group** (the number of "real parameters" in the table of 1-02). The same computation gives the following table Known .

Group	Condition on group elements	Condition on Lie algebra elements	Dimension
SO(n)	Real, $R^\top R = I$, $\det R = 1$	Real, $X^\top = -X$	$n(n-1)/2$
U(n)	$U^\dagger U = I$	$X^\dagger = -X$	n^2
SU(n)	$U^\dagger U = I$, $\det U = 1$	$X^\dagger = -X$, $\mathrm{tr}\,X = 0$	$n^2 - 1$
SL(n, ℝ)	Real, $\det g = 1$	Real, $\mathrm{tr}\,X = 0$	$n^2 - 1$

03

How far the exponential map reaches

Near the identity the exponential map is one-to-one with a continuous inverse Known . The inverse map is the logarithm series $\log(I + A) = A - \frac{A^2}{2} + \frac{A^3}{3} - \cdots$ ($\|A\| < 1$), and a nearby element can be written uniquely as e^X with a small X (**exponential coordinates**).

Far away, SU(2) reaches every element by the $U(\theta, n) = \exp(-\frac{i\theta}{2}\, n \cdot \sigma)$ of 1-02 (surjective), and shifting θ by 4π gives the same element, so it is not one-to-one. **For a compact connected group the exponential map is surjective** Known .

Detour: SL(2, ℝ) has elements the exponential does not reach

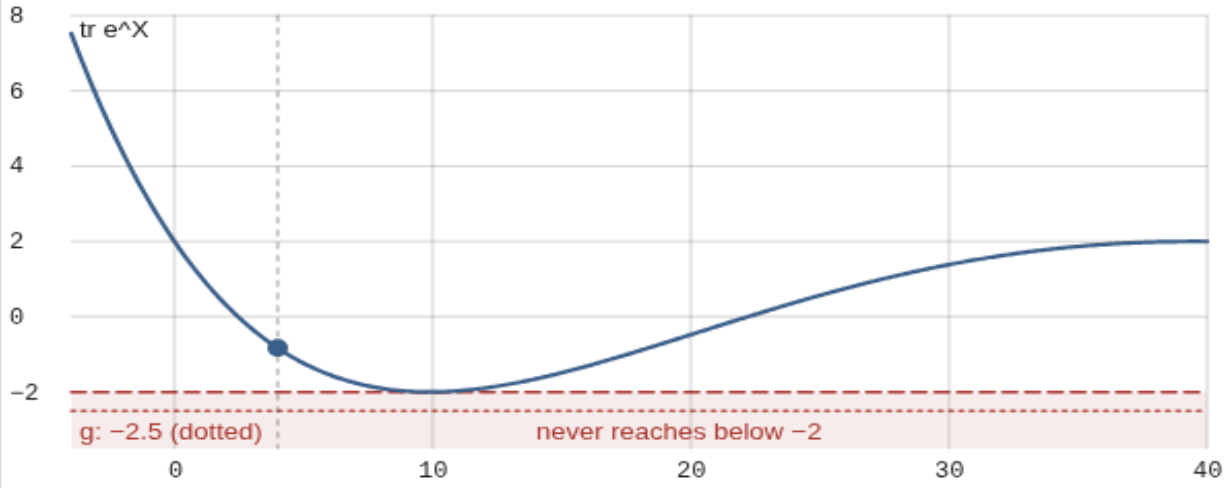
A subsection that may be skipped. It is not a prerequisite for later chapters. It is an example of a non-compact group.

$X \in \mathfrak{sl}(2,\mathbb{R})$ has the form $X = \begin{pmatrix} a & b \\ c & -a \end{pmatrix}$. As can be checked by multiplying out the entries, $X^2 = (a^2 + bc)I = -\delta I$ ($\delta = \det X = -a^2 - bc$), and this always holds for traceless 2×2 matrices. So, as in the example of 1-01 § 08, splitting the series of the exponential into even and odd orders gives

$$\mathrm{tr}\,e^X = 2\cos\sqrt{\delta}\ (\delta > 0),\quad 2\ (\delta = 0),\quad 2\cosh\sqrt{-\delta}\ (\delta < 0) \qquad \text{all } \geq -2$$

Hence no e^X equals $g = \operatorname{diag}(-2, -\frac{1}{2})$ (determinant 1, trace −2.5) Known . Even so, with the J of §01 it can be written as a product of two, $g = e^{\pi J} e^{\operatorname{diag}(\log 2, -\log 2)}$. Computed For 200,000 random X, the minimum of $\operatorname{tr} e^X$ was −2.000000.

Figure 1 Moving $\delta = \det X$ (horizontal axis) with the slider moves the point $\operatorname{tr} e^X$ (vertical axis) along the curve. The curve never goes below the dashed line at −2, so it does not reach the trace −2.5 of g (dotted line). "Play" sweeps δ from −4 to 40.



δ = 4.00, tr e^X = −0.8323, = 2cos√δ (√δ = 2.000; example X = √δ·J: a kind of rotation)

Examples of X	$\delta = \det X$	$\operatorname{tr} e^X$
$\operatorname{diag}(1, -1)$ (stretching)	−1	3.08616
E_{12} (only the (1, 2) entry is 1; nilpotent)	0	2.00000
$3J$ (akin to rotations)	9	−1.97998
πJ	9.8696	−2.00000

Computed The numbers in this figure were computed on this machine. The table by numerical computation of the matrix exponential, the curve by computing the closed formula above on the spot (relative difference between the two at most 4.2×10^{-16}).

04

The commutator is the second-order term of the mismatch

Multiplying four $e^{tX} = I + tX + \frac{t^2}{2}X^2 + O(t^3)$ together, the first-order terms cancel and the second order remains.

$$e^{tX} e^{tY} e^{-tX} e^{-tY} = I + t^2 [X, Y] + O(t^3), \qquad [X, Y] = XY - YX \quad (1)$$

The left-hand side is called the **group commutator**, and $[X, Y]$ the **commutator**. If $[X, Y] \neq 0$, a path that ought to close after four moves fails to close by an amount of order t^2.

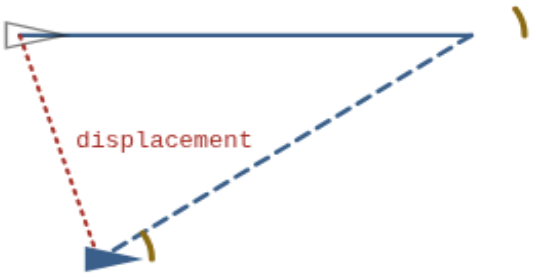
Example: parallel parking. Write the group SE(2) of "points with an orientation" in the plane (position (x, y), orientation α) as 3×3 matrices.

$$g = \begin{pmatrix} \cos\alpha & -\sin\alpha & x \\ \sin\alpha & \cos\alpha & y \\ 0 & 0 & 1 \end{pmatrix}, \qquad [J, P] = P_\perp$$

J is the rotation with the J of § 01 in the upper left, and P, $P_\perp$ are the translation matrices with only the (1, 3) and (2, 3) entries equal to 1. Multiplying e^{tP} on the right means "move forward t in the direction you face", and e^{tJ} means "turn left by t on the spot". As a metaphor, it is a trolley that can turn on the spot but cannot move sideways. To be exact, by (1), "forward → turn left → back → turn right" equals $e^{t^2[P,J]} = e^{-t^2 P_\perp}$ (a translation of t^2 to the right) up to third order, and the exact displacement is $(t(1-\cos t),\ -t\sin t)$.

Figure 2 Changing the step t with the slider changes the displacement after the four moves (dotted line and the label "displacement"). The view is scaled so that a step looks the same length, and the gap shrinking in proportion to t is the sign that "the displacement is of size t^2". "Swap the order" reverses the displacement.

forward → turn → back → turn back



solid: forward, dashed arm back turning on the spot

t = 0.60, displacement = (0.1048, -0.3388), sideways / t² = -0.9411

t	Forward–back displacement Δx	Sideways displacement Δy	$\Delta y/t^2$
0.6	0.1048	−0.3388	−0.94107
0.1	4.996×10^{-4}	-9.983×10^{-3}	−0.99833
0.01	5.000×10^{-7}	-1.000×10^{-4}	−0.99998

Computed The numbers in this figure were computed on this machine. The table by products of matrix exponentials, differing from the exact formula by at most 2×10^{-16}. The figure computes that formula on the spot.

The reason for the θ^2 of 1-02. In 1-02 § 04, the mismatch between rotations about the x and y axes applied in opposite orders had an angle close to θ^2 and an axis almost along −z. With the generators $(L_a)_{bc} = -\varepsilon_{abc}$ ($\varepsilon_{123} = 1$ the totally antisymmetric

symbol), $R_x = e^{\theta L_1}$, $R_y = e^{\theta L_2}$, $[L_1, L_2] = L_3$. The mismatch $D = R_y R_x R_y^{-1} R_x^{-1}$ has the form of (1), so $D = I - \theta^2 L_3 + O(\theta^3)$, a rotation by the angle θ^2 about the −z axis. Computed The difference between $\log D/\theta^2$ and $-L_3$ was 0.12 at $\theta = 10°$ and 0.012 at 1°.

05

The structure called a Lie algebra

The outline of why $\mathfrak{g}$ is a vector space closed under commutators Known . **Sums**: the right-hand side of **the Lie product formula** $e^{t(X+Y)} = \lim_{m\to\infty}(e^{tX/m}e^{tY/m})^m$ (§ 07) is a sequence of elements of G, and G is closed. **Commutators**: from $ge^{sY}g^{-1} = e^{s\,gYg^{-1}}$, $e^{tX}Ye^{-tX} \in \mathfrak{g}$; differentiating this in $t = 0$ gives $XY - YX$.

In the definition that forgets matrices, a **Lie algebra** is a pair of a real vector space and a bilinear operation $[\cdot,\cdot]$ satisfying the following (the second is the **Jacobi identity**).

$$[X,Y] = -[Y,X], \qquad [X,[Y,Z]] + [Y,[Z,X]] + [Z,[X,Y]] = 0$$

The cross product in three dimensions is also a Lie algebra; $L(v) = \sum_a v_a L_a$ satisfies $L(v)w = v \times w$, $[L(a), L(b)] = L(a \times b)$.

The Lie algebra sees only the local. Differentiating the double cover $\rho : \mathrm{SU}(2) \to \mathrm{SO}(3)$ of 1-02 at the identity gives a one-to-one linear map $\mathfrak{su}(2) \to \mathfrak{so}(3)$ that preserves commutators (in § 06 we see that the structure constants agree). The kernel $\{\pm I\}$ is two separated points and does not appear near the identity; the difference between 720° and 360° is global and is not reflected in the Lie algebra.

The adjoint. The group acts on the Lie algebra by $\mathrm{Ad}_g X = gXg^{-1}$, and its derivative is $\mathrm{ad}_X Y = [X,Y]$. $\mathrm{Ad}_{e^X} = e^{\mathrm{ad}X}$ holds Known . The gauge field changes under gauge transformations through Ad (3-02). As a representation it is treated in 1-04.

06

Structure constants

Choosing a basis $X_1, \dots, X_d$, one can write $[X_a, X_b] = \sum_c f_{ab}{}^c X_c$, and the coefficients f are called the **structure constants**. Antisymmetry and the Jacobi identity become conditions on f Known .

- $\mathfrak{so}(3)$: $[L_a, L_b] = \sum_c \varepsilon_{abc} L_c$.
- $\mathfrak{su}(2)$: with $X_a = -\frac{i}{2}\sigma_a$ (the Pauli matrices, 1-02 § 05), $[X_a, X_b] = \sum_c \varepsilon_{abc} X_c$ Computed . The constants are the same, so $X_a \mapsto L_a$ is the isomorphism of § 05, and

the $U(\theta, n)$ of 1-02 is $\exp(\theta \sum_a n_a X_a)$.

In **the physicists' convention** (written e^{-itT}), one writes $[T^a, T^b] = i f^{abc} T^c$ with Hermitian T^a, omitting the summation sign over the repeated index c. Setting $X_a = -iT^a$ gives $[X_a, X_b] = f^{abc} X_c$, so **the real f is the same in the two conventions**. Normalising with $\mathrm{tr}(T^a T^b) = \frac{1}{2}\delta^{ab}$ for $\mathfrak{su}(N)$, f^{abc} is **totally antisymmetric** (the sign changes when any two indices are swapped). Multiplying the commutator formula by T^c and taking the trace gives $\mathrm{tr}([T^a, T^b]T^c) = \frac{i}{2} f^{abc}$, so

$$f^{abc} = -2i\ \mathrm{tr}\big([T^a, T^b]T^c\big) = -2i\big(\mathrm{tr}(T^a T^b T^c) - \mathrm{tr}(T^b T^a T^c)\big)$$

The trace is cyclic ($\mathrm{tr}(ABC) = \mathrm{tr}(CAB)$ from $\mathrm{tr}(PQ) = \mathrm{tr}(QP)$), so the right-hand side does not change under the cyclic permutation of indices $a \to b \to c \to a$, and changes sign under swapping a and b. Combining cyclic permutations with one swap produces every permutation, so it is totally antisymmetric. The f^{abc} of $\mathfrak{su}(3)$ is derived in 1-04.

07

The Baker–Campbell–Hausdorff formula

This is the formula that writes the difference of $e^A e^B \neq e^{A+B}$ seen in 1-01 § 08 in terms of commutators. When X, Y is small enough, $e^X e^Y = e^Z$ gives

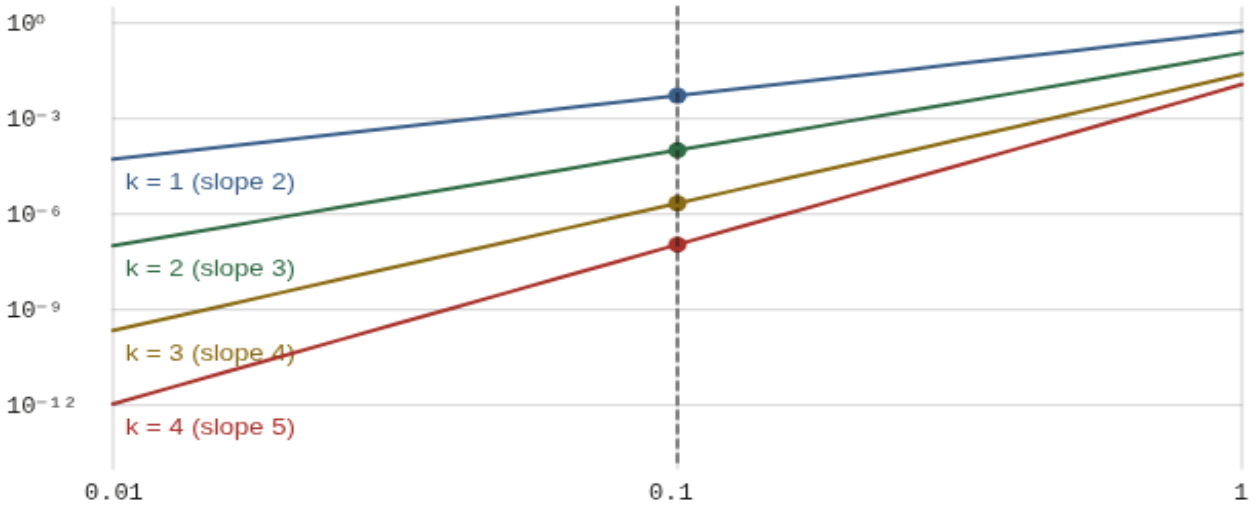
$$Z = X + Y + \tfrac{1}{2}[X,Y] + \tfrac{1}{12}\big([X,[X,Y]] + [Y,[Y,X]]\big) - \tfrac{1}{24}[Y,[X,[X,Y]]] + \cdots \quad (2)$$

The terms up to third order are in Hall and Kirillov Known . The fourth-order term appears in the paper of Casas and Murua in the form $\frac{1}{24}[X,[Y,[Y,X]]]$, and by the Jacobi identity it equals the form of (2) Known . The fourth-order coefficient was also checked numerically on this machine (Figure 3 and the table below) Computed .

Everything beyond $X + Y$ is nested commutators, and the coefficients do not depend on the group Known . So the multiplication near the identity is determined by the Lie algebra. A connected group is generated by products of elements near the identity, so if the group is known as a space, the operation is determined too Known . However, **the Lie algebra alone does not determine the group itself**. SU(2) and SO(3) have the same Lie algebra but are different as groups. The difference lies in the global part seen in § 05 (the kernel $\{\pm I\}$).

For $X = L(1, 0, 0.3)$, $Y = L(0.5, 1, 0)$ of $\mathfrak{so}(3)$ multiplied by t, we measured the error of (2) truncated at order k. If the formula is right, the error is proportional to t^{k+1}.

Figure 3 The difference between $\log(e^{tX}e^{tY})$ and formula (2) up to order k (log–log). The error at order k lies on a line of slope $k+1$. Moving t with the slider shows the error and the error divided by t^{k+1}.



t = 0.100, k=1 5.27e-3 (÷t^2 = 0.527), k=2 1.02e-4 (÷t^3 = 0.102), k=3 2.20e-6 (÷t^4 = 0.022), k=4 1.10e-7 (÷t^5 = 0.011)

t	k = 1	k = 2	k = 3	k = 4
1	0.561	0.116	2.48×10^{-2}	1.19×10^{-2}
0.1	5.28×10^{-3}	1.02×10^{-4}	2.20×10^{-6}	1.10×10^{-7}
0.01	5.27×10^{-5}	1.02×10^{-7}	2.20×10^{-10}	1.10×10^{-12}
Error $/t^{k+1}$ at $t\to0$	0.5274	0.1017	0.0220	0.0110

Computed The numbers in this figure were computed on this machine (the difference in the operator norm; the figure embeds 21 points).

The values 0.5274, 0.1017 and 0.0220 in the last row agree with the size of the next term (the $\frac{1}{2}[X,Y]$, third-order and fourth-order terms), and the coefficients fitted by a polynomial in t also agreed with (2) to within 5×10^{-9} Computed .

The Lie product formula. From (2), $(e^{X/m}e^{Y/m})^m = \exp\big(X+Y+\frac{1}{2m}[X,Y]+O(m^{-2})\big)$, and the error decreases roughly as $1/m$. Computed For the X,Y above, the errors at m = 10, 100, 1000 were 4.57×10^{-2}, 4.57×10^{-3}, 4.57×10^{-4} (multiplied by m, all are about 0.457). This formula has an extension to the case where X,Y are unbounded operators (on an infinite-dimensional Hilbert space, 1-10), called the Trotter product formula Known . That slicing time finely in the path integral takes this form is seen in 2-07.

```
08
```

The Killing form

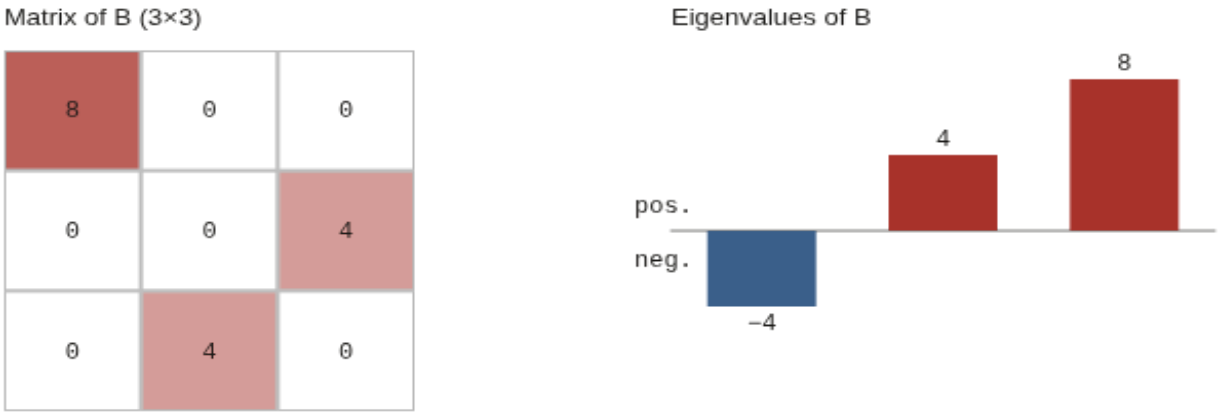
The Yang–Mills action needs a "squared length" on the Lie algebra that does not change under Ad_g (§ 09). The candidate that can be made from the Lie algebra alone is the **Killing form**.

$$B(X,Y) = \mathrm{tr}\big(\mathrm{ad}_X\,\mathrm{ad}_Y\big), \qquad B(X_a,X_b) = \sum_{c,d} f_{ad}{}^{c}\, f_{bc}{}^{d}$$

B is symmetric, and **invariant** in the sense $B(\mathrm{Ad}_g\,X, \mathrm{Ad}_g\,Y) = B(X,Y)$, $B([Z,X],Y) + B(X,[Z,Y]) = 0$ Known . For $\mathfrak{su}(2)$, $\mathrm{ad}_{X_a} = L_a$, so $B(X_a,X_b) = \mathrm{tr}(L_aL_b) = -2\delta_{ab}$: it is **negative definite**.

A subspace $\mathfrak{i}$ is an **ideal** if $[\mathfrak{g},\mathfrak{i}] \subset \mathfrak{i}$. **Simple** means non-commutative with no ideals other than 0 and $\mathfrak{g}$, and **semisimple** means a direct sum of simple ones (the standard definition takes another form, known to be equivalent to this: Kirillov Definition 6.19, Corollary 6.43). The following are known Known . (i) **Cartan's criterion**: semisimple $\Leftrightarrow$ B is non-degenerate. (ii) For the Lie algebra of a compact group, B is negative semidefinite, and the degenerate directions are exactly the centre (the elements that commute with everything). Conversely, if it is semisimple and B is negative definite, it is the Lie algebra of a compact group.

Figure 4 Choosing a row of the table below changes the matrix of B in the basis of that Lie algebra (shade shows size) and the bars of its eigenvalues. A bar below the axis is negative, above it positive, and a circle on the axis is 0.



sl(2, ℝ), signs (negative, 0, positive) = (1, 0, 2), Basis H = diag(1, −1), E = E_{12}, F = E_{21}. B(H, H) = 8 and B(E, F) = 4, so the matrix is not diagonal

Lie algebra	Dimension	Signs of B (negative, 0, positive)	Reading
$\mathfrak{su}(2)$	3	(3, 0, 0)	Compact, simple
$\mathfrak{su}(3)$	8	(8, 0, 0)	Compact, simple
$\mathfrak{so}(4)$	6	(6, 0, 0)	Compact, semisimple $(\mathfrak{su}(2)\oplus\mathfrak{su}(2))$

Lie algebra	Dimension	Signs of B (negative, 0, positive)	Reading
$\mathfrak{sl}(2,\mathbb{R})$	3	(1, 0, 2)	Simple but not compact
$\mathfrak{u}(2)$	4	(3, 1, 0)	Degenerate in the direction of the centre $i\mathbb{R}I$
Lie algebra of SE(2)	3	(1, 2, 0)	Degenerate on the ideal of translations
Heisenberg (upper triangular with zero diagonal)	3	(0, 3, 0)	$B = 0$

Computed The numbers in this figure were computed on this machine (the matrices of ad built from the structure constants, the matrix of B and its eigenvalues computed and embedded). The lengths of the bars change with the choice of basis, but the numbers of negative, 0 and positive do not (Sylvester's law of inertia Known ; the same numbers under 200 random changes of basis). E_{jk} is the matrix with only the (j,k) entry equal to 1.

Computed The matrices of ad were built from the structure constants and the signs of the eigenvalues of B counted, confirming $B = 2n\,\mathrm{tr}(XY)$ ($\mathfrak{su}(2)$–$\mathfrak{su}(4)$) and $B = (n-2)\,\mathrm{tr}(XY)$ ($\mathfrak{so}(3)$–$\mathfrak{so}(5)$). The deviation from invariance was at most 1.3×10^{-12}. That $\mathfrak{so}(4)$ is not simple cannot be read from the signs.

09

The Lie algebra inside Yang–Mills

The symbols d (exterior derivative), $\wedge$ (wedge product), $*$ (Hodge dual) and the integration of forms are defined in 1-06. Here we only look at where Tr appears.

Jaffe and Witten's description of the problem replaces U(1) by a compact group G, takes the curvature to be $F = dA + A \wedge A$, and writes of the Tr in the expression $L = \frac{1}{4g^2}\int \mathrm{Tr}\, F \wedge *F$, which it calls the Lagrangian (integrated over spacetime, so in this book's words it is the action, 2-01), that it is “an invariant quadratic form on the Lie algebra of G”. In this chapter's words:

- **Values of the field**: the gauge field has $A_\mu = \sum_a A_\mu^a X_a \in \mathfrak{g}$ at each point and in each direction ($\mu = 0,1,2,3$ is the number of a direction of spacetime). The components of $A \wedge A$ can be written with the commutator $[A_\mu, A_\nu]$, which vanishes for U(1). The coefficient in front depends on the convention for components ($F = \frac{1}{2} F_{\mu\nu} dx^\mu \wedge dx^\nu$ and so on), so the exact form is left to 3-02.
- **Invariant**: gauge transformations act through Ad, so the action needs a form invariant in the sense of § 08.

- **Compact**: for $\mathfrak{su}(N)$, if $X \neq 0$ then $-\operatorname{tr} X^2 = \sum_{ij} |X_{ij}|^2 > 0$. The energy is bounded below as a sum of squares; if signs are mixed, directions of negative energy appear Physics (3-03).
- **Simple**: on a compact simple Lie algebra ($\mathfrak{su}(N)$ and so on), the invariant symmetric bilinear form is unique up to a constant multiple Known . This is because it stays simple under complexification (extending real coefficients to complex; for $\mathfrak{su}(N)$, $\mathfrak{sl}(N, \mathbb{C})$), and on a complex simple Lie algebra uniqueness follows from Schur's lemma. "Simple" alone is not enough. $\mathfrak{sl}(2, \mathbb{C})$ regarded as a real Lie algebra is simple, but the real and imaginary parts of its Killing form are separate invariant forms. The constant of the unique form is absorbed into the coupling constant g (as is the difference in the normalisation of Tr). For $G_1 \times G_2$ a constant enters for each factor, and the Standard Model SU(3)×SU(2)×U(1) has three coupling constants (3-07).

Next is 1-04 SU(2) and SU(3): the Pauli matrices and Gell-Mann matrices, the adjoint representation, the list of compact simple groups.

What this chapter can say

Item	What is said	What is not said
Lie algebra	That it is closed under commutators (outline) Known . The second-order term of the group commutator, and the numbers for SE(2) and SO(3) Computed	Lie groups as manifolds (after 1-05)
Exponential map	One-to-one locally Known . Surjective for SU(2) (the formula of 1-02), not surjective for SL(2, ℝ) (the trace formula of § 03) Known . A check with 200,000 random elements Computed	The proof that it is surjective for compact connected groups is not given (the theorem itself is known)
BCH, Killing form	The coefficients up to third order and the fourth-order coefficient Known . The numbers for the fourth-order coefficient, the order of the error and the signs in the table Computed . That the Lie algebra does not determine the group itself (SU(2) and SO(3))	The general term and the domain of convergence, and proofs of Cartan's criterion and of compactness, are not given (the theorems themselves are known)
Yang–Mills	That the action needs an invariant quadratic form (primary source). That for compact simple it is unique up to a constant multiple Known . Positive definiteness and energy Physics	Shows nothing about the mass gap or the construction of the theory

Sources and reproduction

Item	Kind	Source / tool
Definitions, the table of § 02, closure under commutators, the Lie product formula and its extension to the Trotter product formula, local one-to-oneness, the image for SL(2, ℝ), $\mathrm{Ad}_{e^X} = e^{\mathrm{ad}X}$, BCH up to third order, simple and semisimple, the physicists' convention	Known (text checked)	B. C. Hall, "An Elementary Introduction to Groups and Representations", arXiv:math-ph/0005032 (2000). Definitions 2.2 and 3.13, the example in Chapter 3 § 5, Theorem 3.9 and the remark just after it (the extension to unbounded operators and the pointer to Reed–Simon), Theorems 3.16 and 3.23, formula (3.17), Chapter 3 Exercise 15, Chapter 4, Chapter 6 § 7
Surjectivity for compact connected groups, the universality of BCH and its terms up to third order, that the operation is determined by the Lie algebra with the proviso "the group itself is not determined", the complexification of $\mathfrak{su}(n)$, the Killing form, Cartan's criterion, the definition of semisimple and the decomposition into direct sums, compactness and negative definiteness, uniqueness of invariant forms (for complex simple Lie algebras)	Known (text checked)	A. Kirillov Jr., *Introduction to Lie Groups and Lie Algebras* (lecture notes; PDF made public on a page of Stony Brook University, the author's institution; the PDF was created in 2005). Numbers are those of the lecture-note version (they may differ from the book version). Remark 3.8 (worded "for compact Lie groups"; the text is read as being about the component containing the identity, that is, connected groups), Theorem 3.34, formula (3.15), Corollary 3.35 and the remark after it, Example 3.54, Exercise 4.4, Definitions 6.19, 6.20 and 6.34, Example 6.35, Theorem 6.37, Corollary 6.43, Theorem 6.49
The book version by the same author	Bibliographic data only	A. Kirillov Jr., *An Introduction to Lie Groups and Lie Algebras*, Cambridge Studies in Advanced Mathematics 113, Cambridge University Press (2008)
The fourth-order term of BCH	Known (text checked)	F. Casas, A. Murua, "An efficient algorithm for computing the Baker–Campbell–Hausdorff series and some of its applications", J. Math. Phys. 50 (2009) 033513, arXiv:0810.2656. The $Z_4 = \frac{1}{24}[X,[Y,[Y,X]]]$ after formula (1.4)
The Trotter product formula	Bibliographic data only	H. F. Trotter, "On the product of semi-groups of operators", Proc. Amer. Math. Soc. 10 (1959) 545–551 / M. Reed, B. Simon, *Methods of Modern Mathematical Physics*, Vol. I (Academic Press, 1972) § VIII.8 (section as pointed to by Hall). Bibliographic data checked in Crossref
A primary source treating the path integral (Feynman integral) as mathematics	Bibliographic data only	E. Nelson, "Feynman Integrals and the Schrödinger Equation", J. Math. Phys. 5 (1964) 332–343 (bibliographic data checked in Crossref; the relation to the form that slices time is in 2-07)

Item	Kind	Source / tool
That the numbers of signs do not depend on the basis (Sylvester's law of inertia)	Bibliographic data only	J. J. Sylvester, “A demonstration of the theorem that every homogeneous quadratic polynomial is reducible by real orthogonal substitutions to the form of a sum of positive and negative squares”, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 4 (1852) 138–142 (bibliographic data checked in Crossref)
That the Tr of the action is "an invariant quadratic form on the Lie algebra of G", $F = dA + A \wedge A$, that formula (1) is called the Lagrangian	Primary source (text checked)	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory”, the official description of the problem by the Clay Mathematics Institute (PDF; no year given in the text; the announcement of the prize in 2000 checked on the Clay page). In: J. Carlson, A. Jaffe, A. Wiles (eds.), *The Millennium Prize Problems*, American Mathematical Society / Clay Mathematics Institute (2006) (bibliographic data checked in Open Library; page numbers not checked)
Tables of Figures 1–4, the numbers of § 03, § 04 and § 06– § 08	Computed on this machine	Python 3.12, NumPy 1.26, SciPy 1.11 (expm, logm). The BCH coefficients by fitting to 20 points of $\|t\| \le 0.05$. The Killing form built from structure constants found by least squares; invariance checked with 50 random pairs of elements, the numbers of signs with 200 random changes of basis. $\operatorname{tr} e^{X}$ by the closed formula (relative difference from expm at most 4.2×10^{-16})
Figures 1–4	Figures	Plain JavaScript (Figures 1 and 2 compute formulas on the spot; Figures 3 and 4 embed precomputed values)

Next chapter to read: 1-04 SU(2) and SU(3).

SU(2) and SU(3) — seeing the matrices in which the gauge field takes values, at a size you can handle

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — su(2) and su(3) through the Pauli and Gell-Mann matrices. Why the irreducible representations of su(2) line up in dimensions 2j+1. The weight diagrams of SU(3). The adjoint representation the gauge field follows, $C_A = N$, and the centre. That the angle between roots can take only four values, and the list of compact simple Lie groups.

Prerequisite chapters — 1-03 Lie groups and Lie algebras (commutators, structure constants, the Killing form). SU(2) and the double cover are in 1-02, and simultaneous diagonalisation in 1-01.

Said first — This is a preparatory chapter, and shows nothing about the mass gap. The theorem of the highest weight, the properties of simple roots and reflections, and the classification theorem are cited without proof (Known).

Order of this chapter

1. Why SU(2) and SU(3) — the number of components and conventions for symbols
2. The Pauli matrices and su(2) — the same Lie algebra as rotations
3. Representations of su(2): the ladder — why 2j is an integer (Figure 1)
4. The Gell-Mann matrices and su(3) — structure constants and rank
5. Weights and roots — drawing the representations of SU(3) on a plane (Figure 2)
   - Detour: the SU(3) of flavour and the SU(3) of colour — putting names of particles on the points of Figure 2 (may be skipped)
6. The adjoint representation — how the gauge field changes, and C_A
7. The centre and N-ality — the part the gauge field cannot see
8. Root systems: four possible angles — growing roots from simple roots and reflections (Figure 3)

9. The list of compact simple Lie groups — the G of the problem statement may be any row

---

01

Why SU(2) and SU(3)

The gauge field has, at each point of spacetime and for each direction, one element of the Lie algebra $\mathfrak{g}$ of the gauge group G (3-02). For G = U(1) of electromagnetism there is one component. For SU(N) the Lie algebra becomes a space of matrices.

$$\mathfrak{su}(N) = \{\, X \in M_N(\mathbb{C}) : X^\dagger = -X,\ \operatorname{tr} X = 0 \,\}, \qquad \dim_{\mathbb{R}} \mathfrak{su}(N) = N^2 - 1 \quad (1)$$

The count: the diagonal entries are purely imaginary, N of them, and the entries above the diagonal are complex, $N(N-1)/2$ of them. From the real degrees of freedom N^2, the trace-zero condition removes one. For SU(3) this is 8, which is why there are 8 kinds of gluon (PDG).

In physics one writes $X = -i\theta^a T_a$ (T_a Hermitian, repeated indices summed). The $U = \exp(-\frac{i\theta}{2}\, n\cdot\sigma)$ of 1-02 is this form with $T_a = \sigma_a/2$. This book normalises $\operatorname{tr}(T_a T_b) = \frac{1}{2}\delta_{ab}$, and writes the structure constants of 1-03 like this.

$$[T_a, T_b] = i f_{abc}\, T_c \quad (2)$$

f_{abc} is real and totally antisymmetric. The PDG review of QCD uses the same convention Known .

02

The Pauli matrices and su(2)

The properties of the Pauli matrices (1-02 § 05) are packed into the following identity. ε_{abc} is the symbol with $\varepsilon_{123} = 1$ whose sign changes when two indices are swapped.

$$\sigma_a \sigma_b = \delta_{ab}\, I + i\, \varepsilon_{abc}\, \sigma_c \quad (3)$$

Taking the trace gives $\operatorname{tr}(\sigma_a\sigma_b) = 2\delta_{ab}$; swapping a and b and subtracting gives $[\sigma_a, \sigma_b] = 2i\varepsilon_{abc}\sigma_c$. With $T_a = \sigma_a/2$ this takes the form of formula (2), and $f_{abc} = \varepsilon_{abc}$. Example: $T_1T_2 = \frac{i}{4}\sigma_3$, $T_2T_1 = -\frac{i}{4}\sigma_3$, so $[T_1, T_2] = iT_3$.

Rotations have the same relation. $(L_a)_{bc} = -i\varepsilon_{abc}$ satisfies $[L_a, L_b] = i\varepsilon_{abc}L_c$, and $e^{-i\theta L_3}$ is the rotation by angle θ about the z axis Computed . **su(2) and so(3) are the**

same Lie algebra. The Lie algebra sees only the neighbourhood of the identity, so the fact that SU(2) covers SO(3) twice (1-02 § 06) does not appear here.

03

Representations of su(2): going up and down the ladder

A representation of su(2) is a triple of Hermitian matrices J_1, J_2, J_3 satisfying formula (2) of $f = \varepsilon$. We build **matrices that go up and down one step at a time**.

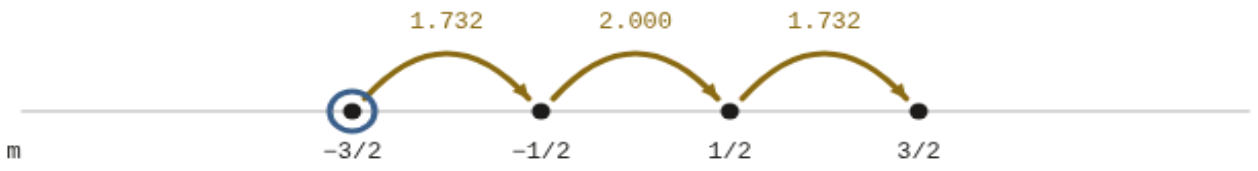
$$J_\pm = J_1 \pm i J_2, \qquad [J_3, J_\pm] = \pm J_\pm, \qquad [J_+, J_-] = 2J_3 \quad (4)$$

If $J_3 v = mv$ then $J_3(J_\pm v) = (m \pm 1) J_\pm v$, so $J_\pm$ moves eigenvalues by 1. $J^2 = J_1^2 + J_2^2 + J_3^2 = J_- J_+ + J_3^2 + J_3$ commutes with every J_a. The dimension is finite, so there is a top v ($J_+ v = 0$); if its eigenvalue is j, then $J^2 v = j(j+1)v$. Going down gives $\|J_- v_m\|^2 = (j+m)(j-m+1)\|v_m\|^2$, and this coefficient cannot be negative, so the sequence must stop exactly at $m = -j$. Going from j to −j in integer steps, **2j is one of 0, 1, 2, ...** Known .

$$J_3 v_m = m\, v_m, \qquad J_+ v_m = \sqrt{(j-m)(j+m+1)}\; v_{m+1}, \qquad m = -j, \ldots, j \quad (5)$$

The dimension is $2j+1$, and there is one irreducible representation in each dimension (the Π_m of 1-02 has m = 2j). Example: j = 1/2 is $J_a = \sigma_a/2$. j = 1, after a change of basis, is the L_a of § 02.

Figure 1 Changing 2j with the slider changes the number of steps and the coefficients. "Climb" keeps multiplying by J_+, which gives 0 at the top.



j = 3/2, dimension 4, J² = 3.75, now m = −3/2, next coefficient √3 = 1.7321

j	Dimension	J²	Coefficients of J₊ (upwards from m = −j)
1/2	2	0.75	$\sqrt{1}$
1	3	2	$\sqrt{2}$, $\sqrt{2}$
3/2	4	3.75	$\sqrt{3}$, $\sqrt{4}$, $\sqrt{3}$
3	7	12	$\sqrt{6}$, $\sqrt{10}$, $\sqrt{12}$, $\sqrt{12}$, $\sqrt{10}$, $\sqrt{6}$

Computed The numbers in this figure and Table 1 were computed on this machine. For the matrices built by formula (5) with j ≤ 3, the deviation in the commutation relations and in $J^2 = j(j+1)I$ was at most 1.8×10⁻¹⁵.

04

The Gell-Mann matrices and su(3)

What plays the role of the Pauli matrices in su(3) is the 8 **Gell-Mann matrices**. All are Hermitian and traceless, with $\mathrm{tr}(\lambda_a \lambda_b) = 2\delta_{ab}$.

$$\lambda_{1,2,3} = \begin{pmatrix} \sigma_{1,2,3} & 0 \\ 0 & 0 \end{pmatrix}, \quad \lambda_4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \lambda_5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}$$

$$\lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \quad \lambda_8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}$$

$\lambda_{4,5}$ does in entries 1 and 3, and $\lambda_{6,7}$ in entries 2 and 3, the same as the Pauli matrices, and λ_8 is a diagonal matrix orthogonal to λ_3. Setting $T_a = \lambda_a/2$ and extracting with $f_{abc} = -2i\,\mathrm{tr}([T_a, T_b]T_c)$ gives the following table.

abc	123	147	156	246	257	345	367	458	678
f_{abc}	1	1/2	−1/2	1/2	1/2	1/2	−1/2	√3/2	√3/2

Table 2: the non-zero ones with a < b < c (the others are fixed by antisymmetry or are 0). The 64 commutators are reproduced, and the deviation from the Jacobi identity is at most 3.3×10^{-16} Computed .

Inside there are three su(2). $\{T_1, T_2, T_3\}$, $\{T_4, T_5, \frac{1}{2}T_3 + \frac{\sqrt{3}}{2}T_8\}$ and $\{T_6, T_7, -\frac{1}{2}T_3 + \frac{\sqrt{3}}{2}T_8\}$ each satisfy formula (2) of $f = \varepsilon$ Computed . They are the SU(2) for each pair of entries (1,2), (1,3), (2,3).

T_3 and T_8 commute. The largest number of linearly independent mutually commuting elements one can take (the dimension of the space they span) is called the **rank**; su(3) has rank 2, and su(N) rank N − 1 Known . Indeed, the elements of su(3) commuting with both T_3 and T_8 are only the linear combinations of these two, and in su(N) too the elements commuting with the N − 1 diagonal ones were only their linear combinations (N = 2–6) Computed .

05

Weights and roots: drawing the representations of SU(3) on a plane

In a representation, T_3 and T_8 are commuting Hermitian matrices, so they can be diagonalised simultaneously, and the pair of eigenvalues (μ_3, μ_8) of a common eigenvector is called a **weight** (the two-dimensional version of the m of su(2)). In the defining representation **3** one just reads off the diagonal entries.

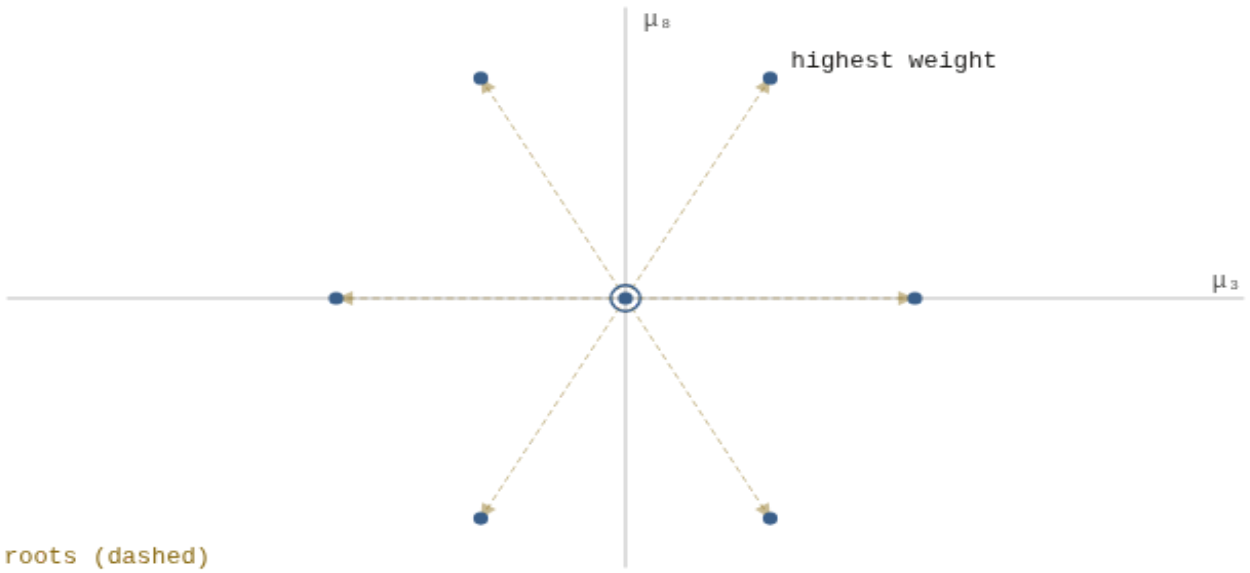
$$\mu(e_1) = \left(\tfrac{1}{2},\ \tfrac{1}{2\sqrt{3}}\right), \qquad \mu(e_2) = \left(-\tfrac{1}{2},\ \tfrac{1}{2\sqrt{3}}\right), \qquad \mu(e_3) = \left(0,\ -\tfrac{1}{\sqrt{3}}\right) \quad (6)$$

The three points form an equilateral triangle, and $\bar{\mathbf{3}}$ (the complex conjugate) is −1 times it. The matrix units E_{ij} (only the (i, j) entry is 1) satisfy $[H, E_{ij}] = (h_i - h_j)E_{ij}$ for diagonal H, and are ladders that move weights by $\mu(e_i) - \mu(e_j)$. These 6 differences are called **roots**, and form a regular hexagon of side 1. $E_{12} = T_1 + iT_2$ is the J_+ of § 03. E_{ij}, like $J_\pm$, is a linear combination of elements of su(3) with complex coefficients, not an element of su(3) itself, but on representations it acts in the same way.

Irreducible representations are labelled by pairs of non-negative integers (p, q) (highest weight $p\,\mu(e_1) - q\,\mu(e_3)$), and the dimension is given by the following formula Known . (1, 0) is **3**, (0, 1) is $\bar{\mathbf{3}}$, and (1, 1) is the adjoint representation.

$$\dim V(p,q) = \tfrac{1}{2}\,(p+1)(q+1)(p+q+2) \quad (7)$$

Figure 2 Choosing (p, q) with the slider shows the weights of V(p, q) on the plane with μ_3 horizontal and μ_8 vertical. The number of rings around a point is its multiplicity. Rows of the table also switch it.



V(1, 1) = 8, dimension 8, Casimir 3.0000, N-ality 0, points 7, largest multiplicity 2

(p, q)	Name	Casimir	N-ality	Weight points	Largest multiplicity
(1, 0)	3	4/3	1	3	1

(p, q)	Name	Casimir	N-ality	Weight points	Largest multiplicity
(0, 1)	$\bar{3}$	4/3	2	3	1
(1, 1)	**8**	**3**	**0**	**7**	**2**
(3, 0)	10	6	0	10	1
(2, 2)	27	8	0	19	3
(3, 3)	64	15	0	37	4

Computed The numbers in this figure and Table 3 were computed on this machine. Weights and multiplicities were found in two ways, by counting monomials and by the Casimir of the representation matrices, and agreed for the 21 cases with p + q ≤ 5. The figure does the counting on the spot, and agreed with the Python computation for all 16. The name is the dimension (overlined if q > p), and the Casimir is the value $(p^2+q^2+pq+3p+3q)/3$ of $\sum_a T_a T_a$.

The rule visible in the figure: the multiplicity increases by 1 from the outer shell inwards, and once the shells have become triangles the multiplicity stops increasing and stays constant to the inside. All 16 computed cases had this form Computed .

Detour: the SU(3) of flavour and the SU(3) of colour

A subsection that may be skipped.

The axes of Figure 2 can be read as physical quantities: $\mu_3 = I_3$ (the third component of isospin) and $\mu_8 = \frac{\sqrt{3}}{2} Y$ (Y is hypercharge). Reading the three basis vectors of the defining representation as the u, d and s quarks gives Y = 1/3, 1/3, −2/3, which are the μ_8 of formula (6) Computed .

According to Rosner's account, in the SU(3) of Gell-Mann and Ne'eman (1961), the simplest families in which charges are integer multiples of the elementary charge are families of 1, 8 and 10, and the 8 baryons including the proton and neutron fitted exactly into the family of 8 Physics . In the (1, 1) of Figure 2, the six outer points are Σ⁺, Σ⁻, p, n, Ξ⁰, Ξ⁻ (for example p is $(I_3, Y) = (\frac{1}{2}, 1)$, the point $(\frac{1}{2}, \frac{\sqrt{3}}{2})$), and the centre with multiplicity 2 is Σ⁰ and Λ Computed . Later Δ, Σ* and Ξ* lined up in the family of 10, and the missing vertex Ω⁻ (the lower vertex of the triangle of (3, 0)) was found by Barnes et al. (1964).

This SU(3) of "flavour" is a broken approximate global symmetry; the gauge group of the strong force is a different SU(3), of three "colours" (PDG). Yang–Mills is of the latter type (3-01).

06

The adjoint representation: the gauge field changes in this representation

A group acts on its own Lie algebra by conjugation. This is the **adjoint representation**, and its derivative is the commutator (1-03).

$$\mathrm{Ad}_g(X) = gXg^{-1}, \qquad \mathrm{ad}_X(Y) = [X, Y], \qquad (T_a^{\mathrm{adj}})_{bc} = -i f_{abc} \quad (8)$$

The matrices on the far right satisfy formula (2) again (equivalent to the Jacobi identity). For SU(2), $-i\varepsilon_{abc}$ is the L_a of § 02, so **the map SU(2) → SO(3) of 1-02 is the adjoint representation itself**, corresponding to j = 1.

Under a gauge transformation g that does not depend on space, the gauge field changes as $A \mapsto gAg^{-1}$ (the general form is in 3-02). For U(1), $gAg^{-1} = A$ and the photon carries no charge, but for SU(N) **the gauge field itself is charged** and interacts with itself. In the PDG too, gluons are in the adjoint representation and quarks in the fundamental representation Known .

$\sum_a T_a T_a$ commutes with every T_b (the same as the J^2 of § 03). By **Schur's lemma** (for an irreducible complex representation, a matrix that commutes with all representation matrices is a constant multiple of the identity), on an irreducible representation it is a constant multiple of the identity, and the constant is called the Casimir Known .

$$\sum_a T_a T_a = C_F\, I, \quad C_F = \frac{N^2-1}{2N}; \qquad f_{acd} f_{bcd} = C_A\, \delta_{ab}, \quad C_A = N \;(9)$$

N	2	3	4	5	6
dim su(N)	3	8	15	24	35
C_F	3/4	4/3	15/8	12/5	35/12
C_A	2	3	4	5	6
Killing form / tr	4	6	8	10	12

Table 4: computed with generalised Gell-Mann matrices Computed . The bottom row is the ratio of the Killing form $B(X,Y) = \mathrm{tr}(\mathrm{ad}_X\, \mathrm{ad}_Y)$ to $\mathrm{tr}(XY)$, which was 2N. For anti-Hermitian $X \neq 0$, $\mathrm{tr}(X^2) < 0$, so B is negative definite, and compactness shows in the sign.

That the Tr of the action is "an invariant quadratic form on the Lie algebra of G" was seen in 1-03 § 09. On a compact simple Lie algebra (whose complexification is also simple), the invariant symmetric bilinear form is unique up to a constant multiple Known , so the difference between B and tr in Table 4 is also absorbed into g. C_A enters the coefficient of asymptotic freedom. The PDG one-loop coefficient $b_0 = (11C_A - 4n_f T_R)/(12\pi)$ is $11N/(12\pi)$ for pure Yang–Mills without quarks Known . The N of "11N" in 17-01 and 4-06 is C_A.

> 07

The centre and N-ality

The set of elements that commute with every element is called the **centre**. For SU(N) it is the scalar matrices of determinant 1 Known .

$$Z(\mathrm{SU}(N)) = \{\, \omega^k I : k = 0, 1, \ldots, N-1 \,\}, \qquad \omega = e^{2\pi i/N} \quad (10)$$

An element of the centre commutes with all representation matrices, so on an irreducible representation it acts as a scalar by Schur's lemma (§ 06). For SU(3), comparing diagonal entries gives $\omega I = \exp(4\pi i T_8/\sqrt{3})$, and a vector of weight (μ_3, μ_8) is multiplied by $e^{4\pi i \mu_8/\sqrt{3}}$. The highest weight of V(p, q) is $\mu_8 = (p+2q)/(2\sqrt{3})$, so the whole representation is multiplied by $\omega^{p+2q} = \omega^{p-q}$. This value held at every weight point (216 points in all) of the 16 representations with p, q ≤ 3 Computed . $k = p - q \bmod 3$ is called the **N-ality**. In the column of Table 3, **3** is 1, $\bar{\mathbf{3}}$ is 2, and the adjoint is 0.

It is 0 for the adjoint because $zXz^{-1} = X$ for a scalar matrix z. **The gauge field does not feel the centre**. So gluons alone cannot screen charges with non-zero N-ality Physics , and the centre plays the leading role in arguments about confinement (17-02, 5-10). Holland et al. (2003) argue that for G_2, whose centre is trivial (§ 09), charges are screened by gluons and the string can break Physics .

08

Root systems: only four possible angles

The roots of § 05 exist for every compact simple Lie algebra. For rank r, the roots line up in an r-dimensional space as the simultaneous eigenvalues of r mutually commuting elements. The inner product (α, β) is, as in § 05, the inner product of the coordinates when the generators are made orthonormal for the invariant quadratic form (for su(3) it is the ordinary inner product on the (μ_3, μ_8) plane, and the roots have length 1).

The heart of the classification is the "2j is an integer" of § 03. For each root α, E_α, $E_{-\alpha}$ and their commutator form an su(2), and seeing another root β from this su(2), the quantity corresponding to m is $(\alpha,\beta)/(\alpha,\alpha)$, which is **an integer multiple of 1/2 (including integers)** Known . In su(3), with α = (1, 0) (E_{12}; the su(2) is $\{T_1, T_2, T_3\}$) and β = $(\frac{1}{2}, \frac{\sqrt{3}}{2})$ (E_{13}), m = 1/2. Multiplying the two directions together gives the following for the angle θ between them.

$$\frac{2(\alpha,\beta)}{(\alpha,\alpha)} \cdot \frac{2(\alpha,\beta)}{(\beta,\beta)} = 4\cos^2\theta \in \{0, 1, 2, 3\} \quad (11)$$

This is because the left-hand side is a product of integers, and is less than 4 if $\beta \neq \pm\alpha$. The ratio of the two factors is the square of the ratio of lengths, so the ratio of the lengths of two non-orthogonal roots is one of $1,\ \sqrt{2},\ \sqrt{3}$ (because the pairs of integers with product 1, 2 or 3 are limited).

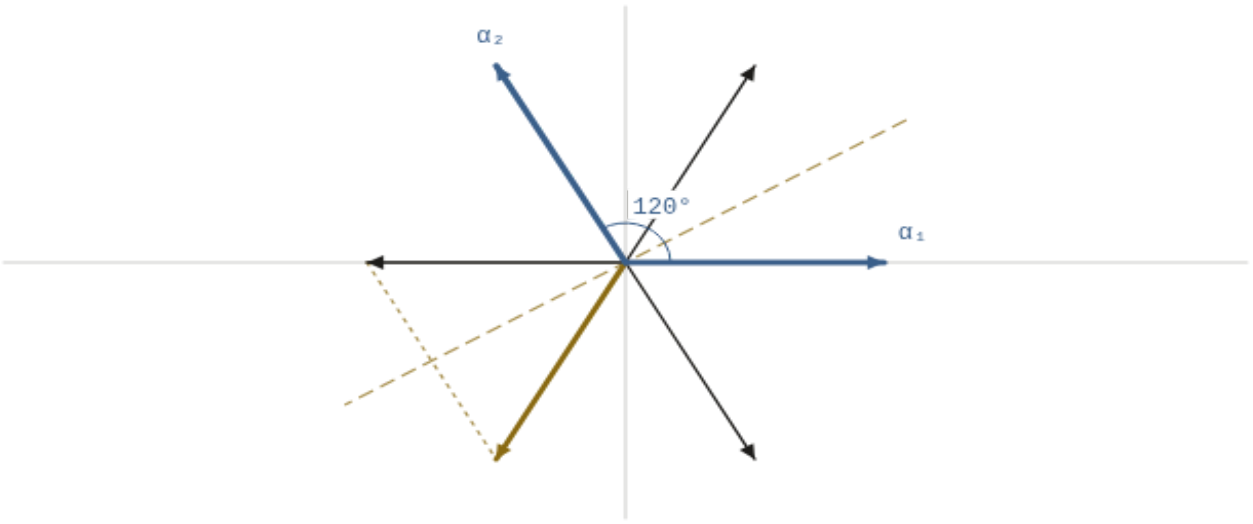
There are many roots, so we choose representatives. Draw a line through the origin containing no root (a hyperplane, for rank 3 or more), and call the roots on one side the **positive roots**. The positive roots that cannot be written as a sum of two positive roots are the **simple roots**; their number equals the rank, and the angles between them are at least 90° Known . So the angles between simple roots take four values: 90°, 120°, 135°, 150°.

What recovers the whole from the simple roots is **reflection**. The reflection in a root α is the folding across the line (hyperplane) perpendicular to α.

$$s_\alpha(\beta) = \beta - \frac{2(\alpha,\beta)}{(\alpha,\alpha)}\,\alpha \quad (12)$$

Reflections take roots to roots, and repeatedly applying the reflections in the simple roots to the simple roots gives all the roots Known . The coefficients in (12) are the integers of (11), so every root is a sum of simple roots with integer coefficients. The coefficients are either all at least 0 or all at most 0 Known .

Figure 3 Choosing a type of rank 2 in a row of the table shows the simple roots α_1, α_2 (thick lines) and their angle. Changing the number with the slider or pressing "Grow" adds roots one by one by the reflections of formula (12) (the dashed line is the mirror, the dotted line the folding).



A₂, roots 6 / 6 −α₁ reflected in the mirror of α₂ gives −α₁ − α₂. That is all. Dimension 8

Type	Angle of simple roots	4cos²θ	Long roots / short roots	Number of roots	Dimens
$A_1 \times A_1$	90°	0	1	4	
A_2	**120°**	**1**	**1**	**6**	
B_2	135°	2	√2	8	

Type	Angle of simple roots	4cos²θ	Long roots / short roots	Number of roots	Dimens
G_2	150°	3	√3	12	

Computed The numbers in this figure and Table 5 were computed on this machine. The order in which roots were grown by repeated reflections from the simple roots, and the coefficients in terms of the simple roots, were found in Python and embedded in the figure. The long roots have length 1.

In rank 2 there are $A_1 \times A_1$ (4 roots; it splits into two su(2) and is not simple), A_2 (6 roots; the regular hexagon of § 05 itself), B_2 (8 roots) and G_2 (12 roots). The dimension is rank + number of roots (for su(3), the 2 diagonal directions and E_{ij} for each of the 6 roots, giving 8). The coefficients in Figure 3 were also all at least 0 or all at most 0 Computed .

09

The list of compact simple Lie groups

The problem statement (0-03) asks for "any compact simple gauge group G". Pushing the restrictions of § 08 through for rank 3 and above, the candidate Lie algebras are exhausted by four series $A_n = \mathfrak{su}(n+1)$, $B_n = \mathfrak{so}(2n+1)$, $C_n = \mathfrak{sp}(n)$, $D_n = \mathfrak{so}(2n)$ and five exceptions G_2, F_4, E_6, E_7, E_8 Known . According to Baez's survey, Killing announced the classification in a letter of 1887 and published it in papers over the following three years (he counted 6 exceptions). Cartan's thesis of 1894 constructed them all and found that the two of dimension 52 are isomorphic, bringing the exceptions to 5.

Type	Simply connected group	Rank	Dimension	$h^\vee$	Centre
A_n ($n \geq 1$)	SU(n+1)	n	n(n+2)	n+1	$\mathbb{Z}_{n+1}$
B_n ($n \geq 2$)	Spin(2n+1)	n	n(2n+1)	2n−1	$\mathbb{Z}_2$
C_n ($n \geq 3$)	Sp(n)	n	n(2n+1)	n+1	$\mathbb{Z}_2$
D_n ($n \geq 4$)	Spin(2n)	n	n(2n−1)	2n−2	$\mathbb{Z}_4$ (n odd), $\mathbb{Z}_2 \times \mathbb{Z}_2$ (n even)
G_2 / F_4	G_2 / F_4	2 / 4	14 / 52	4 / 9	Trivial
E_6 / E_7 / E_8	E_6 / E_7 / E_8	6 / 7 / 8	78 / 133 / 248	12 / 18 / 30	$\mathbb{Z}_3$ / $\mathbb{Z}_2$ / trivial

Table 6: the series agree with the formulas for n ≤ 8 Computed . Spin(m) is the simply connected group that covers SO(m) twice, and Sp(n) is the group of unitary matrices with quaternion entries. The matrix of the integers $2(\alpha_i, \alpha_j)/(\alpha_j, \alpha_j)$ between simple roots (the factors of formula (11)) is called the **Cartan matrix**, and the order of the

centre is its determinant Known . The form of the centre ($\mathbb{Z}_4$ or $\mathbb{Z}_2\times\mathbb{Z}_2$) was also computed from this matrix. $h^\vee$ (the dual Coxeter number) is only cited as a table value in this chapter.

For SU(N), $h^\vee = N$, agreeing with the $C_A = N$ of § 06. If the squared length of the long roots is set to 1 (for SU(N) this corresponds to $\mathrm{tr}(T_a T_b) = \frac{1}{2}\delta_{ab}$, the same as the roots having length 1 in § 05), then $C_A = h^\vee$ for general G as well Known . In the normalisation common in mathematics, "squared length 2", the adjoint Casimir becomes $2h^\vee$. Comparing the two normalisations for SU(3) gave 3 and 6 Computed . **The G of the problem statement may be any row of this table**, and the group enters the coefficients of perturbation theory through numbers like these. Groups divided by the centre, such as SU(2)/$\mathbb{Z}_2$ = SO(3), have the same Lie algebra, and the difference appears in the topology of bundles (3-05 § 03).

Next is 1-05 Manifolds and tangent spaces: the language for putting this chapter's matrices at each point of spacetime.

What this chapter can say

Item	What is said	What is not said
su(2), su(3)	The f_{abc} of Table 2. Why 2j is an integer. That formula (5) is a representation for j ≤ 3. The dimension of the elements commuting with the diagonal ones in su(N) (N = 2–6) Computed	A complete proof that these exhaust the irreducible representations, and that the rank does not depend on the choice of maximal commuting part Known
Weights of SU(3)	Weights and multiplicities for p, q ≤ 3 Computed	Proofs of the theorem of the highest weight and of formula (7) Known . The rule of shells is an observation
Adjoint representation	C_F, C_A and B = 2N·tr for N = 2–6 Computed	Schur's lemma, general N, uniqueness of the invariant quadratic form Known
Centre	The centre of SU(N), that the adjoint representation does not feel the centre. The N-ality of SU(3) (all weights for p, q ≤ 3) Computed	The relation to confinement and G_2 Physics
Root systems and classification	Why there are four angles (11). Growing the four root systems of rank 2 by reflections (Figure 3, Table 5). Table 6 Computed	The properties of simple roots and reflections, that the classification is exhausted by these, and $C_A = h^\vee$ for general G (in the normalisation where the long roots have squared length 1) Known
Yang–Mills	How the gauge field changes, and where the N of 11N comes from	Shows nothing about the mass gap or the construction of the theory

Sources and reproduction

Item	Kind	Source / tool
Irreducible representations of su(2), the highest weights of SU(3) and formula (7)	Known (text checked: Chapter 5 § 4, Chapter 6 § 3, Theorem 6.11)	B. C. Hall, “An Elementary Introduction to Groups and Representations”, arXiv:math-ph/0005032 (2000)
$t = \lambda/2$, C_F, C_A, b_0, the representations of gluons and quarks	Known (text checked)	J. Huston, K. Rabbertz, G. Zanderighi, “Quantum Chromodynamics”, in S. Navas et al. (PDG), Phys. Rev. D 110 (2024) 030001
"Compact simple gauge group", "invariant quadratic form"	Primary source (text checked)	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory”, the Clay Mathematics Institute's description of the problem
The history of the classification (the letter of 1887, the 6 exceptions Killing counted, Cartan's thesis of 1894 and the isomorphism of the two of dimension 52), the series and the exceptions	Survey (text checked)	J. C. Baez, “The Octonions”, Bull. Amer. Math. Soc. 39 (2002) 145–205, arXiv:math/0105155
The history of the SU(3) of flavour (the simplest families with integer charges being 1, 8 and 10, the family of 8 baryons, Δ, Σ*, Ξ* and Ω⁻)	Account (text checked)	J. L. Rosner, “The Eightfold Way”, arXiv:hep-ph/0109241 (2001)
The centre of G_2 and string breaking	Abstract checked	K. Holland, P. Minkowski, M. Pepe, U.-J. Wiese, Nucl. Phys. B 668 (2003) 207, arXiv:hep-lat/0302023
Original papers, the discovery of Ω⁻, literature on the relation between the dual Coxeter number $h^\vee$ and the adjoint Casimir	Bibliographic data only	W. Killing, Math. Ann. 31–36 (1888–1890); É. Cartan, thesis (1894); M. Gell-Mann, CTSL-20 (1961); Y. Ne'eman, Nucl. Phys. 26 (1961) 222; V. E. Barnes et al., “Observation of a Hyperon with Strangeness Minus Three”, Phys. Rev. Lett. 12 (1964) 204; V. G. Kac, *Infinite Dimensional Lie Algebras*, 3rd edition (1990)
Tables 1–6, the checks of § 02, § 04, § 07 and the detour, Figures 1–3	**Computed on this machine**	Python 3.12, NumPy 1.26, SciPy 1.11: `su2_ladder.py`, `su3_algebra.py`, `su3_weights.py`, `classification.py`, `rank_center_roots.py`. The figures use canvas and plain JavaScript; formula (5) and the counting of monomials are computed on the spot, and the order of roots in Figure 3 is embedded from the Python results

Next chapter to read: 1-05 Manifolds and tangent spaces.

Manifolds and tangent spaces — gluing charts together, and differentiating without depending on coordinates

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — How to handle a curved space by gluing together coordinates (charts): the manifold. The space of directions at a point (the tangent space) and the transformation law of components. The bracket, which measures how far flows fail to commute.

Prerequisite chapters — 1-01 and multivariable calculus (the chain rule, the implicit function theorem). If you have read 1-02 and 1-03, the examples connect.

Said first — This chapter is preparation, and shows nothing about the mass gap. Theorems are cited without proof and carry the "Known" label. The numbers are checks at finitely many points.

Order of this chapter

1. Why manifolds are needed
2. Coordinate neighbourhoods and atlases (moving figure)
3. Definition and examples of manifolds
4. Cutting out by equations (computed)
5. Smooth functions and maps
6. Tangent vectors
7. The transformation law and the differential of a map
8. Vector fields and flows / Detour: the hairy ball theorem
9. The bracket (moving figure)
10. Manifolds inside Yang–Mills

01

Why manifolds are needed

Spaces that are not flat appear in Yang–Mills theory. The gauge group SU(2) is, as a space, the sphere S^3 (1-02 § 05). The stage of the problem statement is $\mathbb{R}^4$, but the official description also touches on approximation by the four-dimensional torus T^4 and on general four-dimensional manifolds Known .

Such spaces sometimes cannot be covered by a single set of coordinates. As a metaphor: however a world map is drawn, it tears at a pole or at the date line. To be exact, there is no one-to-one correspondence, continuous in both directions, between the sphere S^2 and an open set of $\mathbb{R}^2$ Known . The sphere is closed and bounded, and so is its continuous image, but a non-empty open set of $\mathbb{R}^2$ is not.

So we glue several charts together smoothly, and **call geometric only those quantities that do not depend on the choice of chart**. It is the same stance as saying that only quantities independent of the gauge are observed (2-03 § 03).

02

Coordinate neighbourhoods and atlases

A pair (U, φ) of a subset U of a set M and a φ that maps U one-to-one onto an open set of $\mathbb{R}^n$ is called a **chart** (coordinate neighbourhood), and $\varphi(q) = (x^1, \dots, x^n)$ are the coordinates of q. A collection of charts covering M that satisfies the following two conditions is called an **atlas** Known . For any two charts $(U_\alpha, \varphi_\alpha)$, (U_β, φ_β), the images of the overlap $\varphi_\alpha(U_\alpha \cap U_\beta)$ and $\varphi_\beta(U_\alpha \cap U_\beta)$ are open sets of $\mathbb{R}^n$ (because derivatives cannot be defined except on open sets). And the change of reading on the overlap (the **transition map**) is C^∞.

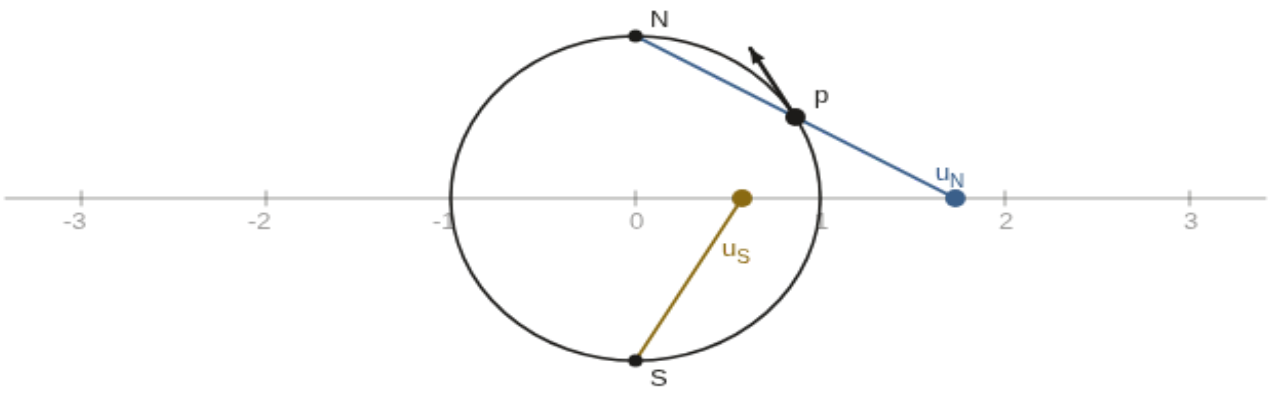
$$\varphi_\alpha \circ \varphi_\beta^{-1} : \ \varphi_\beta(U_\alpha \cap U_\beta) \longrightarrow \varphi_\alpha(U_\alpha \cap U_\beta) \quad \text{is } C^\infty \ (1)$$

Example: the circle S^1. The angle θ does not make a single chart (on $0 \le \theta < 2\pi$ the image is not open, and on $0 < \theta < 2\pi$ one point is missing). So we use **stereographic projection**. Let u_N be the position where the line joining the north pole $N = (0, 1)$ and the point $p = (\cos\theta, \sin\theta)$ meets the horizontal axis, and u_S the position obtained in the same way from the south pole $S = (0, -1)$.

$$u_N = \frac{\cos\theta}{1 - \sin\theta}, \qquad u_S = \frac{\cos\theta}{1 + \sin\theta}, \qquad u_N\, u_S = \frac{\cos^2\theta}{1 - \sin^2\theta} = 1 \ (2)$$

u_N covers everything except N, u_S covers everything except S, and $u_S = 1/u_N$ on the overlap is smooth, so (1) is satisfied.

Figure 1 Moving p with the slider (or a row of the table) changes the coordinates u_N (blue) and u_S (ochre), and their product is always 1. As p approaches the north pole, u_N grows without bound. The arrow in the text colour is the tangent vector of p moving at speed 1 (§ 07).



θ = 30°, u_N = 1.7321, u_S = 0.5774, product = 1.000000, v_N = 2.0000, v_S = -0.6667, v_S/v_N = -0.3333, $-1/u_N^2$ = -0.3333

θ (degrees)	uN	uS	vN	vS	vS/vN	$-1/uN^2$
30	1.7321	0.5774	2.0000	−0.6667	−0.3333	−0.3333
89	114.5887	0.0087	6565.7794	−0.5000	−0.0001	−0.0001
−135 (= 225)	−0.4142	−2.4142	0.5858	−3.4142	−5.8284	−5.8284

Computed The numbers in this figure were computed on this machine (formula (2) and its θ derivative; relative difference from numerical differentiation at most 5×10^{-10}). v_N and v_S are du_N/dt and du_S/dt when going round at speed 1 (§ 05, § 07).

The sphere S^2 too can be covered by the projections $u = (x_1, x_2)/(1 - x_3)$, $u' = (x_1, x_2)/(1 + x_3)$ from the two poles onto the equatorial plane, and the gluing is $u' = u/|u|^2$ Known (difference at most 2×10^{-13} at about 2000 random points Computed).

03

Definition and examples of manifolds

Definition A (smooth) **manifold** of dimension n is a pair of a set M and an atlas covering M. Every chart that can be added without breaking (1) is added (a maximal atlas) Known . Textbooks also impose topological conditions (the Hausdorff property and so on), but in the examples of this book they are satisfied automatically.

Examples: open sets of $\mathbb{R}^n$ (such as $\mathrm{GL}(n,\mathbb{R}) \subset \mathbb{R}^{n^2}$) Known , the sphere S^n, products (the torus $T^2 = S^1 \times S^1$, T^4), groups of matrices (§ 04). **Non-examples**: the crossing point of a figure 8. Removing the crossing point from a small neighbourhood of it leaves four pieces, so it cannot have the same shape as an interval, which splits into

two when one point is removed. The vertex of the double cone $x^2 + y^2 = z^2$ is the same.

This definition does not use an outside space. Things that differ only as atlases (for example, the atlas on $\mathbb{R}$ made from the chart $x \mapsto x^3$) are carried to the usual one by a diffeomorphism (§ 05), so we regard them as the same structure. Even counted this way, genuinely different smooth structures can sit on the same space. A smooth structure that is homeomorphic to $\mathbb{R}^4$ (there is a one-to-one correspondence continuous in both directions) but not diffeomorphic to it is called an **exotic** $\mathbb{R}^4$, and Taubes is credited with first showing that there are uncountably many of them (of the cardinality of the continuum, that is, as many as there are real numbers) Known (1-12 § 07, 3-06, the article on four dimensions).

04

Cutting out by equations — the regular value theorem

Many manifolds appear as solution sets of equations. For a smooth $F : \mathbb{R}^N \to \mathbb{R}^m$ and a value c, the following holds (the **regular value theorem**) Known .

$$M = F^{-1}(c), \qquad \operatorname{rank} dF_x = m \ \ (x \in M) \implies M \text{ is an } N-m\text{-dimensional manifold} \ (3)$$

What lies inside is the implicit function theorem: solve m of the coordinates as functions of the remaining $N - m$, and use those $N - m$ as a chart. For example, $F(A) = A^{\mathsf{T}}A$ maps $\mathbb{R}^9$ to the symmetric matrices (6 entries), and $F^{-1}(I) = \mathrm{O}(3)$ has dimension $9 - 6 = 3$. det is continuous and takes only ± 1, so $\mathrm{O}(3)$ splits into the part with $\det A = 1$ (the rotation group $\mathrm{SO}(3)$) and the part with $\det A = -1$, both of dimension 3. This agrees with a rotation being fixed by 2 numbers for the direction of the axis and 1 for the angle (1-02 § 04).

Table 1 Computed The rank of dF was counted at 200 random points, and for the groups the kernel at the identity (the tangent space of § 06) was also found. The rank is the number of square roots of eigenvalues of $dF\,dF^{\mathsf{T}}$ (singular values) that cannot be regarded as 0 at a relative threshold of 10^{-6}. The points of the orthogonal group were taken from the part with $\det = 1$.

Set	Equation	N	Rank	Dimension	Kernel at the identity
$S^3 = \mathrm{SU}(2)$	$\lvert x\rvert = 1$	4	1	3	—
$\mathrm{O}(3)$ (the part with $\det = 1$ is $\mathrm{SO}(3)$)	$A^{\mathsf{T}}A = I$	9	6	3	antisymmetric (residual 3×10^{-16})
$\mathrm{U}(2)$	$A^{\dagger}A = I$	8	4	4	anti-Hermitian

Set	Equation	N	Rank	Dimension	Kernel at the identity
cone (non-example)	$x^2+y^2=z^2$	3	0 at the vertex	–	–

SL(2, ℝ), SL(3, ℝ) (kernel: trace 0), O(4) (the part with $\det = 1$ is SO(4)) and U(3) also gave dimensions 3, 8, 6, 9. The cone drops rank at the vertex. A closed group of matrices is always a manifold (a Lie group) Known , and this backs up the discussion of 1-03 § 02.

05

Smooth functions and maps

A function $f : M \to \mathbb{R}$ is called **smooth** when $f \circ \varphi^{-1}$ is C^∞ in every chart. On an overlap (4) holds, so if it is smooth in one chart it is smooth in the other (the reason for requiring smooth gluing). Maps $F : M \to N$ are the same, and a bijection whose inverse is also smooth is called a **diffeomorphism**.

$$f \circ \varphi_\beta^{-1} = \left(f \circ \varphi_\alpha^{-1}\right) \circ \left(\varphi\alpha \circ \varphi_\beta^{-1}\right) \quad (4)$$

Example: height on the circle. Assign to the point $p = (\cos\theta, \sin\theta)$ of S^1 the height $f(p) = \sin\theta$. From (2), $u_N^2 = \frac{1+\sin\theta}{1-\sin\theta}$ and $u_S^2 = \frac{1-\sin\theta}{1+\sin\theta}$; solving these for $\sin\theta$, the f read in the two charts of § 02 is

$$f = \frac{u_N^2 - 1}{u_N^2 + 1} \quad \text{(chart N)}, \qquad f = \frac{1 - u_S^2}{1 + u_S^2} \quad \text{(chart S)}$$

The denominators are both at least 1, so both formulas are C^∞, and f is smooth. Substituting $u_S = 1/u_N$ on the overlap carries one formula to the other, a concrete instance of (4) Computed .

The values of partial derivatives differ from chart to chart. At the point $\theta = 0$ ($u_N = u_S = 1$) we have $df/du_N = 1$, $df/du_S = -1$: the scales of the two charts run in opposite directions. Even so, the rate of change $\frac{d}{dt} f(\gamma(t))|_{t=0}$ along a curve γ is written without coordinates. If $\gamma(t) = (\cos t, \sin t)$, the coordinate speeds $v_N = du_N/dt$, $v_S = du_S/dt$ at $t = 0$ are 1 and -1 (the quantities of the table of Figure 1). By the chain rule, in chart N it is 1×1 and in chart S $(-1) \times (-1)$, both $1 = \cos 0$ Computed . This is the seed of differentiation independent of coordinates.

06

Tangent vectors — three views

In a curved space an arrow stretching from point to point sticks out of the space, so we redefine an arrow as **a direction attached to one point**. Three views give the same thing Known . **(a) Velocity of a curve**: regard curves whose velocities read in a chart are equal as the same (independent of the chart by (1)). **(b) Directional derivative**: $v(f) = \frac{d}{dt} f(\gamma(t))|_{t=0}$. Linear, and satisfies the product rule. **(c) Components**: using the chain rule in a chart x,

$$v(f) = \sum_{i=1}^{n} v^i \frac{\partial f}{\partial x^i}(p), \qquad v^i = \frac{d}{dt} x^i(\gamma(t))\Big|_{t=0}, \qquad v(fg) = v(f)\, g(p) + f(p)\, v(g) \quad (5)$$

and one can write $v = \sum_i v^i \partial_i$. The whole set of these directions is the **tangent space** T_pM at p (a vector space of dimension n), and $\partial_1, \dots, \partial_n$ is called the **coordinate basis**.

When cut out by equations. For a curve inside $M = F^{-1}(c)$ we have $F(\gamma(t)) = c$, so $dF_p(\gamma'(0)) = 0$, and counting dimensions gives $T_pM = \ker dF_p$ Known . At the identity of SO(3), differentiating $A^{\mathsf{T}}A = I$ gives $X^{\mathsf{T}} + X = 0$: the tangent space is the set of all antisymmetric matrices, and agrees with the Lie algebra of 1-03 § 02 Known (the right-hand column of Table 1) Computed .

07

The transformation law of components and the differential of a map

The tangent vector v is fixed before any chart, and the components v^i are its shadow cast onto a chart. The components in another chart x' are obtained by the chain rule Known .

$$v'^{\,i} = \sum_{j=1}^{n} \frac{\partial x'^{\,i}}{\partial x^j}(p)\, v^j, \qquad \partial'_i = \sum_{j=1}^{n} \frac{\partial x^j}{\partial x'^{\,i}}(p)\, \partial_j \quad (6)$$

The components change by the Jacobian matrix, the basis by its inverse, and $\sum_i v^i \partial_i$ is invariant. Example: in $(x, y) = (1, 1)$ ($r = \sqrt{2}$, $\theta = \pi/4$), $\partial_x = \frac{1}{\sqrt{2}}\partial_r - \frac{1}{2}\partial_\theta$. For the circle of Figure 1, $v_N = 1/(1 - \sin\theta)$ and $v_S = -1/(1 + \sin\theta)$, and (6) is $v_S = -v_N/u_N^2$. The agreement of the two right-hand columns of the table of Figure 1 is the check of this Computed (v_N is large near the north pole because the scale of chart N stretches).

The differential of a map. A smooth $F : M \to N$ carries tangent vectors by carrying curves.

$$dF_p\big(\gamma'(0)\big) = (F \circ \gamma)'(0), \qquad d(G \circ F)_p = dG_{F(p)} \circ dF_p \quad (7)$$

$dF_p : T_pM \to T_{F(p)}N$ is linear, and read in charts it is the Jacobian matrix. (6) is the case of the identity map read in two charts (a linear example is the Lorentz transformation, 2-04). Treating df as "something that eats a vector and returns a number" is done in 1-06.

08

Vector fields and flows

A smooth assignment $X = \sum_i X^i(x)\,\partial_i$ of a tangent vector to each point is called a **vector field**. A curve following the arrows (an **integral curve**) is a solution of an ordinary differential equation, and is locally unique Known . The point reached by advancing p by time t is written $\varphi_t(p)$ and called the **flow**.

$$\frac{dx^i}{dt} = X^i\big(x(t)\big), \qquad \varphi_s \circ \varphi_t = \varphi_{s+t}, \qquad \varphi_0 = \mathrm{id}$$ (8)

The flow of a linear field $X(x) = Ax$ is $\varphi_t(x) = e^{tA}x$, that is, the matrix exponential of 1-01. Examples in the plane: the rotation $R = -y\,\partial_x + x\,\partial_y$ is rotation by the angle t, the dilation $E = x\,\partial_x + y\,\partial_y$ is scaling by e^t, and $x\,\partial_y$ is $(x, y) \mapsto (x, y + tx)$.

Detour: the hairy ball theorem

A subsection you may skip. It is not a prerequisite for later chapters.

A continuous vector field on S^2 vanishes somewhere Known . As a metaphor: however you comb a hairy ball, a cowlick remains. To be exact, if we call the number of times the arrow turns while going once round a zero on a small circle its **index**, then for a field with finitely many zeros the sum of the indices is said to be 2 regardless of the field (the Poincaré–Hopf theorem Known). This 2 is the **Euler characteristic** of S^2, the same value as vertices − edges + faces counted by dividing the sphere into a polyhedron. Computed Going round the zeros in the charts of § 02, the rotation about the z axis and the gradient of the height x_3 have +1 at each pole, the field $\partial/\partial u^1$ that is constant in chart N has +2 at the north pole only, and the sum was 2 in every case. The Euler characteristic of the torus is 0, and there are fields on it without zeros.

Obstructions of the type "a reference cannot be chosen continuously at each point globally" appear in another form in gauge fixing too (3-04).

09

The bracket — can flows be interchanged?

Applying two fields to a function one after the other produces second derivatives, but swapping the order and subtracting makes the second-order terms vanish (partial

derivatives can be taken in either order), and a new vector field is left. This is the **bracket** Known .

$$[X,Y]f = X(Yf) - Y(Xf), \qquad [X,Y]^i = \sum_{j=1}^{n} \left(X^j \frac{\partial Y^i}{\partial x^j} - Y^j \frac{\partial X^i}{\partial x^j} \right) \quad (9)$$

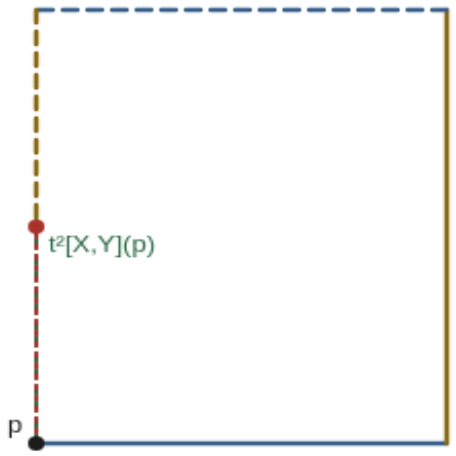
For coordinate basis fields, $[\partial_i, \partial_j] = 0$. Conversely, if any two of n fields forming a basis commute, then locally there is a chart whose coordinate basis they are (a theorem on commuting frames, a form of the Frobenius theorem) Known .

Geometric meaning. Going t by the flow of X, t by Y, $-t$ by X and $-t$ by Y, the path does not close, and the displacement appears with size t^2 in the direction of $[X,Y]$ Known . For example, if $X = \partial_x$ and $Y = x\,\partial_y$, then $(x,y) \to (x+t,y) \to (x+t, y+tx+t^2) \to (x, y+tx+t^2) \to (x, y+t^2)$, and the displacement is exactly $t^2 \partial_y = t^2 [X,Y]$.

$$\varphi^Y_{-t} \circ \varphi^X_{-t} \circ \varphi^Y_{t} \circ \varphi^X_{t}\,(p) = p + t^2\,[X,Y](p) + O(t^3) \quad (10)$$

Figure 2 Choosing a pair with the buttons A–E (or a row of the table) and t with the slider changes the four flows (blue is X, ochre Y; dashed is the direction of $-t$) and the displacement that does not close (red). Green is the leading term $t^2[X,Y](p)$ of (10). The view is scaled so that the sides have the same length, so the displacement shrinking in proportion to t is the sign of "size t^2". $p = (0.6,\ 0.4)$.

Pair B: X = ∂_x (blue), Y = x∂_y (ochre)



t = 0.60, displacement = (0.0000, 0.3600), displacement/t² = (0.0000, 1.0000), [X,Y](p) = (0, 1)

Pair	X	Y	$[X,Y]$	$[X,Y](p)$	displacement / t^2 ($t = 0.5$)	$t = 0.01$
A	∂_x	∂_y	0	(0, 0)	(0, 0)	(0, 0)
B	∂_x	$x\partial_y$	∂_y	(0, 1)	(0, 1.0000)	(0, 1.0000)
C	R	E	0	(0, 0)	(0, 0)	(0, 0)
D	∂_x	R	∂_y	(0, 1)	(0.2448, 0.9589)	(0.0050, 1.0000)

Pair	X	Y	$[X,Y]$	$[X,Y](p)$	displacement / t^2 ($t = 0.5$)	$t = 0.01$
E	$x\partial_y$	$y\partial_x$	$x\partial_x - y\partial_y$	(0.6, −0.4)	(0.9500, −0.7000)	(0.6041, −0.4060)

Computed The numbers in this figure were computed on this machine. The table solves the flows by the Runge–Kutta method (step 10^{-3}) and evaluates the bracket (9) by central differences (the displacement for A and C is at most 10^{-15}). The figure computes with closed formulas for the flows.

A closes, being a coordinate basis. C also commutes: $R = \partial\theta$ and $E = \partial/\partial(\log r)$ are the coordinate basis of the chart $(\log r, \theta)$. D is the same group SE(2) as the parallel parking of 1-03 § 04, seen as motions of points, and the displacement is exactly $(t(1 - \cos t), t \sin t)$. E is an example of a linear field $X_A(x) = Ax$, and from (9) $[X_A, X_B] = X_{BA-AB}$: the **sign is opposite** to that of the matrix commutator.

"The amount by which a small quadrilateral fails to close" appears three times: in the multiplication of a group (1-03 § 04), in flows (this section), and in curvature (1-07: the mismatch of parallel transport around a small loop) Known .

10

Manifolds inside Yang–Mills

- **Spacetime**: $\mathbb{R}^4$ can be covered by a single chart. The official description writes the gauge field of electromagnetism (a connection for the group U(1)) as "locally a 1-form on spacetime", and treats the non-abelian case by replacing this with the group G (Jaffe–Witten § 1) Known . The meaning of 1-forms and connections is given in 1-06 and 1-07.
- **The four-dimensional sphere**: on Euclidean $\mathbb{R}^4$ (with the time coordinate entering distances with the same sign as space; 2-04, 4-02), solutions of the Yang–Mills equations (3-03) with finite action $\int |F|^2$ (F is the field strength, 3-02) are said to extend smoothly, after a change of gauge, to S^4, obtained by adding the point at infinity (Uhlenbeck 1982, 3-05) Known .
- **Gauge groups and lattices**: the tangent space at the identity is the Lie algebra (1-04). The space of lattice configurations is a finite product of copies of the group, and is a compact manifold (here meaning bounded and closed in the space of lined-up matrix entries; 5-01).
- **The space of gauge orbits**: it is infinite-dimensional, and it has been argued that a continuous gauge fixing cannot be chosen globally (Singer 1978, 3-04) Known .

Choosing a chart and choosing a gauge, transition maps and transition functions, charts of the sphere and the obstruction to gauge fixing are much alike. But this is a metaphor. To be exact, the gauge side is about a structure that places a group at each

point (a **principal bundle**, defined in 1-07) and connections on it, and this chapter is its groundwork.

What this chapter can say

Item	What is said	What is not said
Manifolds	Definition, examples, non-examples, the regular value theorem Known . Gluing, Table 1, the example of § 05 Computed	The role of the topological conditions, proofs of the theorems. The numbers are at finitely many points
Tangent vectors	The three views, (6), (7) Known . The transformation law on the circle Computed	Proof that the three views agree. Differential forms (1-06)
Flows and the bracket	(9), (10) Known . Displacement / $t^2 \to [X,Y]$ (p) for five pairs, the indices of the detour Computed	The remainder of (10); proofs of the commuting-frame theorem, the hairy ball theorem and the Poincaré–Hopf theorem
Yang–Mills	Guide only	Shows nothing about the mass gap or the construction of the theory

Sources and reproduction

Item	Kind	Source / tools
Many of the definitions and theorems of § 02, § 03, § 05–§ 09	Known (text checked)	S. M. Carroll, "Lecture Notes on General Relativity", arXiv:gr-qc/9712019 (1997). Around formulas (2.4)–(2.13), (3.69), (3.74), (5.16)
Matrix Lie groups are Lie groups, GL(n, ℝ) is an open set, the Lie algebra is the tangent space at the identity	Known (text checked)	B. C. Hall, "An Elementary Introduction to Groups and Representations", arXiv:math-ph/0005032 (2000). Definition 2.14, Theorem 2.15, Chapter 3 Exercise 19
The $\mathbb{R}^4$ of the problem statement, T^4 and general four-dimensional manifolds, a connection being locally a 1-form (the case of U(1))	Primary source (text checked)	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory", the official description of the problem by the Clay Mathematics Institute. § 1, § 4– § 6
The regular value theorem, formula (10), the topological conditions	Bibliographic data only	J. M. Lee, *Introduction to Smooth Manifolds*, Springer GTM (2013)
The hairy ball theorem	Bibliographic data only	J. Milnor, "Analytic Proofs of the 'Hairy Ball Theorem' and the Brouwer Fixed Point Theorem", Amer. Math. Monthly 85 (1978) 521–524

Item	Kind	Source / tools
The Poincaré–Hopf theorem (the sum of the indices of a field with finitely many zeros)	Bibliographic data only (Poincaré's original paper not checked)	H. Hopf, “Vektorfelder in n-dimensionalen Mannigfaltigkeiten”, Math. Ann. 96 (1927) 225–249 / J. Milnor, *Topology from the Differentiable Viewpoint*, University Press of Virginia (1965)
That there are continuum-many exotic $\mathbb{R}^4$	Bibliographic data only	C. H. Taubes, “Gauge theory on asymptotically periodic 4-manifolds”, J. Differential Geom. 25 (1987) 363–430
The obstruction to gauge fixing	Bibliographic data only	I. M. Singer, “Some remarks on the Gribov ambiguity”, Commun. Math. Phys. 60 (1978) 7–12
Extension to S^4 (Euclidean $\mathbb{R}^4$, finite action, change of gauge)	Bibliographic data only	K. Uhlenbeck, “Removable singularities in Yang–Mills fields”, Commun. Math. Phys. 83 (1982) 11–29
The numbers of Figure 1, Figure 2, Table 1, the example of § 05 and the detour	Computed on this machine	Python 3.12, NumPy 1.26. The figures are plain JavaScript

Next chapter: 1-06 Differential forms and the exterior derivative — on to things that eat tangent vectors and return numbers (1-forms), and Stokes' theorem.

Differential forms and the exterior derivative — gathering gradient, curl and divergence into one d, and counting holes

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — A differential form is an "oriented measuring device" that eats tangent vectors and returns numbers. The exterior derivative d gathers gradient, curl and divergence, and $d^2 = 0$ and Stokes' theorem hold the whole together. On to forms that are closed but not exact counting holes, the electromagnetic field being written $F = dA$, and F and $*F$ being comparable only in four dimensions.

Prerequisite chapters — 1-05 Manifolds and tangent spaces. For determinants, 1-01. The metric used in § 09 is defined on the spot to the minimum needed, and the sign conventions follow 2-04.

Said first — This is a preparatory chapter, and shows nothing about the mass gap. The theorems of Stokes, Poincaré and de Rham are either only sketched or cited without proof (Known). Gauge fields go only as far as U(1); the matrix-valued case is in 1-07.

Order of this chapter

1. Why differential forms — giving a name to what is integrated
2. 1-forms — a line integral is a pullback
3. k-forms and the wedge product — oriented area (Figure 1)
4. The exterior derivative — $d^2 = 0$, gradient, curl, divergence
5. Integration and orientation — the Jacobian comes from the wedge product, the example of the hemisphere
6. Stokes' theorem — the inner edges cancel (Figure 2)
7. Closed and exact — dθ and the winding number (Figure 3)
8. An entrance to de Rham cohomology — counting holes. Detour: differential forms on a lattice (Figure 4)

9. The Hodge dual and Maxwell's equations — making $*$ from the metric, $F = dA$, what is special about four dimensions

---

01

Why differential forms

The $P\,dx + Q\,dy$ of a line integral looks like decoration inside the integral sign. Differential forms are the language that makes this "thing that is integrated" an object in its own right. How it changes under a change of coordinates is fixed by the chain rule, and gradient, curl and divergence gather into a single operation d.

$$\int_\gamma (P\,dx + Q\,dy) = \int_a^b \big(P(\gamma(t))\,x'(t) + Q(\gamma(t))\,y'(t)\big)\,dt, \qquad \gamma(t) = (x(t), y(t)) \tag{1}$$

The rule "replace dx by $x'(t)\,dt$" is the first example of what § 02 calls the **pullback**.

Example. Traversing in the opposite direction changes the sign of (1), while the length $\int |\gamma'(t)|\,dt$ does not change. What differential forms measure are **oriented quantities**. The integrand of (1) is a rule that returns a number for the velocity $(x'(t), y'(t))$ at each point of the curve. This view, "returning a number for a tangent vector", is the starting point of § 02.

In this book the electromagnetic field appears as $F = dA$ (§ 09), the gauge field as a matrix-valued 1-form (1-07), and the instanton number as the integral of a 4-form (3-05).

02

1-forms

A linear function that returns a number for a tangent vector at a point p (1-05) is called a **cotangent vector**, and a smooth assignment of one to each point is called a **1-form**. The basic example is the differential of a function df, which returns the directional derivative for v (repeated indices are summed).

$$df_p(v) = v^i\,\frac{\partial f}{\partial x^i}(p), \qquad dx^i\Big(\frac{\partial}{\partial x^j}\Big) = \delta^i_j, \qquad \alpha = \alpha_i\,dx^i \tag{2}$$

dx^i is the dual basis of $\partial/\partial x^j$, and under a change of coordinates the chain rule gives $dx'^j = \frac{\partial x'^j}{\partial x^i}\,dx^i$. The components α_i change by the transpose of the inverse of the matrix $\partial x'^j/\partial x^i$ that changes the components of tangent vectors, so $\alpha(v) = \alpha_i v^i$ does

not depend on coordinates. The line integral along a curve γ is the one-variable integral $\int_a^b \gamma^*\alpha$ of the **pullback** $\gamma^*\alpha = \alpha_i(\gamma(t))\,\dot\gamma^i(t)\,dt$.

Example 1. Pulling back $\alpha = -y\,dx + x\,dy$ along the unit circle $(\cos t, \sin t)$ gives $\gamma^*\alpha = dt$, and the integral is 2π. **Example 2.** Since $\gamma^* df = \frac{d}{dt} f(\gamma(t))\,dt$,

$$\int_\gamma df = f(\gamma(b)) - f(\gamma(a)) \quad (3)$$

and for a closed curve it is 0. Hence the α of Example 1 is not the df of any f. But α is not "closed" in the sense of § 07 either ($d\alpha = 2\,dx \wedge dy \neq 0$ by (7) of § 04). An example that is closed yet not of the form df is the $d\theta$ of § 07.

The rule of pullback is the same even when it is not a curve. For a map $\varphi : u \mapsto x = \varphi(u)$, $\varphi^*\omega$ is ω with x replaced by $\varphi(u)$ and dx^i by $\frac{\partial \varphi^i}{\partial u^j}\,du^j$. A curve is the case where $u = t$ has one variable, and every L^*, φ^* from § 03 on has this meaning.

03

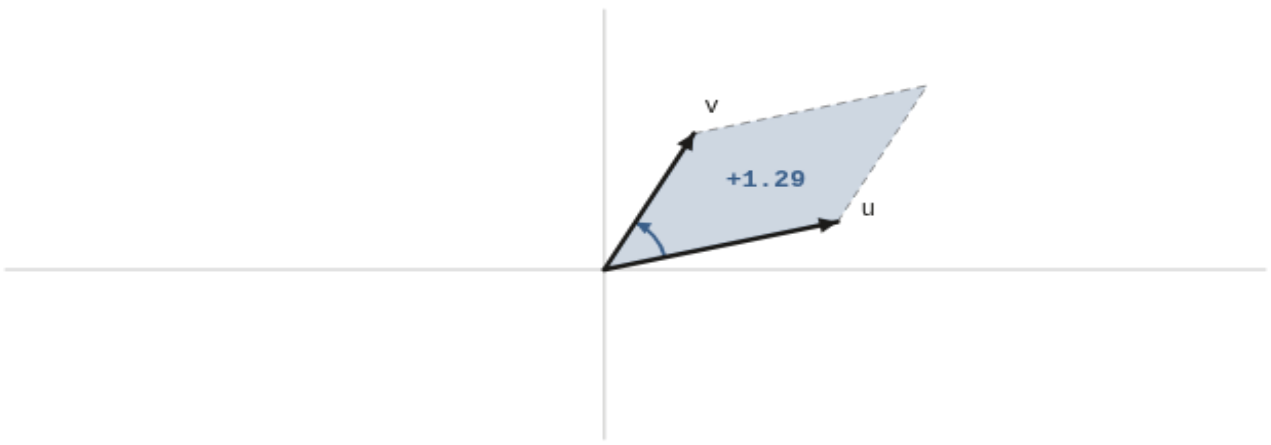
k-forms and the wedge product

A 2-form returns a number from two tangent vectors, is linear in each, and changes sign when they are swapped. As a metaphor, "an area meter that remembers orientation"; to be exact, a smooth arrangement of antisymmetric bilinear forms on the tangent space at each point.

$$(\alpha \wedge \beta)(u,v) = \alpha(u)\,\beta(v) - \alpha(v)\,\beta(u), \qquad (dx \wedge dy)(u,v) = u^1 v^2 - u^2 v^1 = \det[\,u\ v\,] \quad (4)$$

On the left is the **wedge product** of 1-forms, with the rule $\alpha \wedge \beta = -\beta \wedge \alpha$, in particular $\alpha \wedge \alpha = 0$. On the right is the oriented area of the parallelogram spanned by u, v, negative when v lies on the clockwise side of u (Figure 1). A k-form in n dimensions is a sum of $dx^{i1} \wedge \cdots \wedge dx^{ik}$ $(i_1 < \cdots < i_k)$ with function coefficients; it has $\binom{n}{k}$ components, and is 0 if k > n. The wedge product of a k-form and an l-form is a (k+l)-form, and $\alpha \wedge \beta = (-1)^{kl}\,\beta \wedge \alpha$.

Figure 1 Rotating the direction of v with the slider changes the sign of the oriented area $(dx \wedge dy)(u,v) = \det[\,u\ v\,]$ of the parallelogram. It is positive (blue) if the second vector is on the anticlockwise side of the first, negative (ochre) if on the clockwise side, and 0 when they lie on a line. Making it (v, u) with "Swap the order" reverses only the sign for the same parallelogram.




```
Direction of v 60°, (dx∧dy)(u, v) = +1.293, the second is on the anticlockwise side of the first
→ positive
```

Computed $u = (1.5, 0.35)$, $|v| = 1.15$. The area is computed on the spot from the formula $u^1v^2 - u^2v^1$ (the correspondence between the sign and "the anticlockwise side" was checked for 360 directions).

Example. In $\mathbb{R}^3$ both 1-forms and 2-forms have 3 components. Reading the wedge product of $a = a_i\,dx^i$ and $b = b_i\,dx^i$ through the coefficients of $dy \wedge dz,\ dz \wedge dx,\ dx \wedge dy$ gives the cross product $a \times b$ (agreement for 10^4 random pairs Computed). A 2-form on $\mathbb{R}^4$ has 6 components, exactly enough for the 3 of the electric field and the 3 of the magnetic field (§ 09). Pulling back by a linear map L gives $L^*(dx \wedge dy) = (\det L)\,dx \wedge dy$.

04

The exterior derivative

We extend $f \mapsto df$ to an operation from k-forms to (k+1)-forms (the exterior derivative in its present form is said to begin with É. Cartan (1899)). The rule is just "differentiate the coefficients, and multiply the resulting dx^j from the left by the wedge product".

$$d\big(f\,dx^{i1} \wedge \cdots \wedge dx^{ik}\big) = \frac{\partial f}{\partial x^j}\,dx^j \wedge dx^{i1} \wedge \cdots \wedge dx^{ik} \quad (5)$$

d is linear, and for a k-form α, $d(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^k \alpha \wedge d\beta$. And

$$d \circ d = 0 \quad (6)$$

The reason: $d(df) = \partial_i \partial_j f\,dx^i \wedge dx^j$, where the coefficients are symmetric in i, j (partial derivatives can be swapped) and $dx^i \wedge dx^j$ is antisymmetric, so the sum vanishes. For k-forms too it vanishes in the same way, by the symmetry of second derivatives of the coefficients and the antisymmetry of the wedge product. **Example.** In the plane, using only $dx \wedge dx = 0$ and $dy \wedge dx = -dx \wedge dy$,

$$d(P\,dx + Q\,dy) = \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dx \wedge dy \quad (7)$$

Known In $\mathbb{R}^3$, the d of a 0-form is the gradient, the d of a 1-form $a_i\,dx^i$ is the curl ($(\partial_y a_z - \partial_z a_y)\,dy \wedge dz + \cdots$), and the d of a 2-form $b_x\,dy \wedge dz + \cdots$ is the divergence $(\nabla \cdot \mathbf{b})\,dx \wedge dy \wedge dz$. (6) states $\nabla \times \nabla f = 0$ and $\nabla \cdot (\nabla \times \mathbf{a}) = 0$ at once, and is the identity used when 2-03 § 02 introduced potentials.

Since the pullback of § 02 and d can be interchanged ($\varphi^* d\omega = d\varphi^* \omega$) Known , d does not depend on coordinates, and is defined on a manifold without a metric. The $df = \partial_r f\,dr + \partial_\theta f\,d\theta$ in polar coordinates has no $1/r$ of the gradient formula; $1/r$ comes in from the metric.

05

Integration and orientation

A k-form is integrated over an oriented k-dimensional piece. Describe the piece by $\varphi : D \subset \mathbb{R}^k \to M$, and pull back to an ordinary multiple integral: $\int_\varphi \omega = \int_D \varphi^* \omega$. From (4) follows:

$$\varphi^*(dx^1 \wedge \cdots \wedge dx^k) = \det\left(\frac{\partial \varphi^i}{\partial u^j}\right) du^1 \wedge \cdots \wedge du^k \quad (8)$$

In place of the $|\det|$ of the change-of-variables formula, the signed $\det$ appears, and reversing the orientation changes the sign. On a whole manifold we compute chart by chart and glue, so the orientation must be consistent over the whole (**orientable**) (the Möbius strip is not) Known .

Example. In polar coordinates $dx \wedge dy = (\cos\theta\,dr - r\sin\theta\,d\theta) \wedge (\sin\theta\,dr + r\cos\theta\,d\theta) = r\,dr \wedge d\theta$. The r of the area element comes from the wedge product.

Example (a surface). Orient the upper hemisphere outwards and describe it in spherical coordinates $\psi(\vartheta, \varphi) = (\sin\vartheta\cos\varphi, \sin\vartheta\sin\varphi, \cos\vartheta)$ $(0 \le \vartheta \le \pi/2)$. The order (ϑ, φ) matches the outward orientation. Computing the $\det$ of (8) (the 2×2 for the two components x, y) gives $\psi^*(dx \wedge dy) = \cos\vartheta\sin\vartheta\,d\vartheta \wedge d\varphi$, so

$$\int_{\text{upper hemisphere}} z\,dx \wedge dy = \int_0^{2\pi}\int_0^{\pi/2} \cos^2\vartheta\,\sin\vartheta\,d\vartheta\,d\varphi = \frac{2\pi}{3}$$

(numerical integration and a check of the orientation Computed). The value is the same as the volume of the solid hemisphere. This is not a coincidence but an instance of

Stokes' theorem in the next section, § 06: $d(z\,dx \wedge dy) = dx \wedge dy \wedge dz$, and on the base disc, part of the boundary of the solid hemisphere, $z = 0$.

06

Stokes' theorem

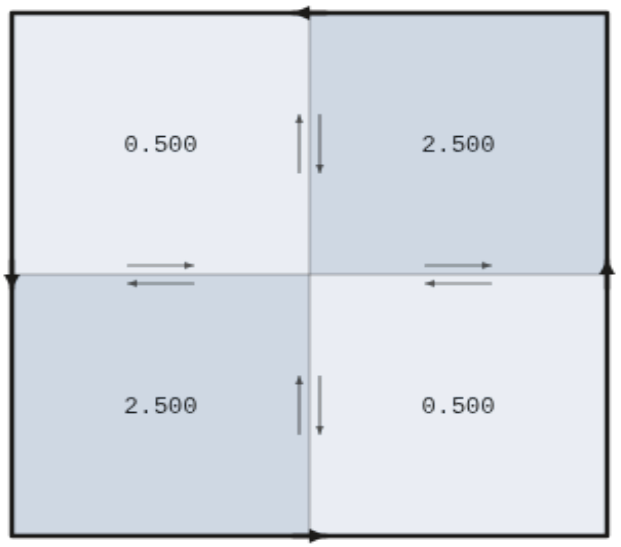
"The total of the derivative inside is the total of the values on the boundary." It is the form of (3) gathered for all dimensions. If M is a compact oriented n-dimensional manifold (it may have a boundary ∂M) and ω an (n−1)-form, then

$$\int_M d\omega = \int_{\partial M} \omega \quad (9)$$

The orientation of ∂M is taken so that putting the outward normal first gives the orientation of M. For an interval it is (3); for a planar region it is Green's theorem via (7); for a surface in $\mathbb{R}^3$ it is Stokes' theorem for the curl; for a solid it is Gauss's divergence theorem Known .

Sketch of the proof. On a rectangle $[a, b] \times [c, d]$, if $\omega = Q\,dy$ then $\iint \partial_x Q\,dx\,dy = \int_c^d (Q(b, y) - Q(a, y))\,dy$, which is the sum of the line integrals along the right edge (upwards) and the left edge (downwards). A general region is divided into small squares. Adjacent squares traverse their common edge in opposite directions, so the inner edges cancel and only the outer boundary remains. Meanwhile, the circulation around a small square approaches "$d\omega$ at the centre × area".

Figure 2 However the number of divisions N is changed, the total of the circulations of the small squares (numbers and shading) stays at the value 6 of the outer boundary. "Circulation ÷ area" approaches $d\omega$ at the centre. For N ≤ 4, the opposite arrows traversed by the two neighbours are drawn on the inner edges.



N = 2, Σ circulation = 6.000000, ∮ on the outer boundary = 6.000000, max|circulation/area − dω| = 0.250000 = 1/4

$\omega = (-\frac{y}{4} - xy^2 - y^3)\,dx + (\frac{x}{4} + x^2 y)\,dy$, $d\omega = (\frac{1}{2} + 4xy + 3y^2)\,dx \wedge dy$, region $[-1,1]^2$. Computed The circulations are computed on the spot from closed formulas for the line integrals along the four edges (agreeing with Table 1).

Computed **Table 1** For the ω of Figure 2, the circulation of each small square found as the sum of the line integrals along its four edges (no surface integrals used).

N	Total of circulations	max |circulation/area − dω(centre)|
1	6.000000000000	1.000000
2	6.000000000000	0.250000
8	6.000000000000	0.015625
64	6.000000000000	0.000244

For every N the total agrees with the $\oint \omega = 6$ of the outer boundary (difference at most 3×10^{-15}), and the deviation decreases exactly as $1/N^2$ (on a square of side $h = 2/N$ the average of $3y^2$ is larger than the value at the centre by $h^2/4$). On the unit disc too both sides agree at $5\pi/4$ Computed .

07

Closed and exact

If $d\omega = 0$ the form is called **closed**, and if there is an η with $\omega = d\eta$ it is called **exact**. By (6), exact implies closed. The converse depends on the shape of the space. On the plane with the origin removed:

$$d\theta := \frac{x\,dy - y\,dx}{x^2 + y^2} \quad (10)$$

By (7), $\partial_x \frac{x}{r^2} + \partial_y \frac{y}{r^2} = \frac{2r^2 - 2x^2 - 2y^2}{r^4} = 0$, so it is closed. On the unit circle, as in Example 1, $\oint d\theta = 2\pi$, so by (3) it is not exact. The argument θ increases by 2π in one turn, so it is a function only on a region with a cut (for example the negative x axis). **dθ is defined without a cut.**

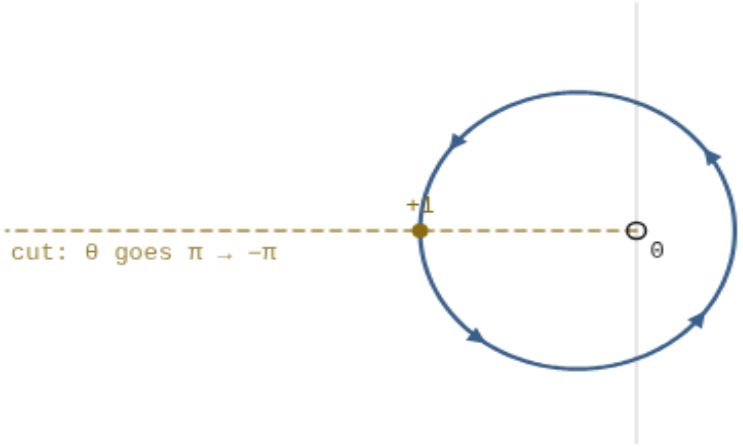
The Poincaré lemma. On a star-shaped region (one containing the segment joining some point to each point) every closed form is exact Known . If the region is star-shaped with respect to the origin, then for a closed 1-form α

$$\eta(x) = \int_0^1 x^i\, \alpha_i(tx)\, dt \quad (11)$$

satisfies $d\eta = \alpha$ (checked with $d(\sin x\, e^y + x^3 y)$; moving the base point to (1, 0), θ is recovered from dθ on the right half-plane Computed).

The winding number. $\frac{1}{2\pi}\oint_\gamma d\theta$ is the number of times the closed curve γ goes round the origin, and does not change under deformations that do not cross the origin (dθ is closed on the band swept by the deformation, so by (9) the difference is 0). θ jumps from π to −π each time it crosses the negative x axis from above to below, so the winding number is also **the oriented number of crossings of the cut**.

Figure 3 Moving the centre c_x of the closed curve makes $\frac{1}{2\pi}\oint d\theta$ jump through the integers 0, 1, 2. The value changes only when the curve crosses the origin, and equals the oriented number of crossings of the cut (dashed line).




```
Circle, centre (-0.375, 0), ∮dθ/2π = 1.000000, crossings of the cut = 1, distance from the origin 0.625
```

Computed The sum of differences of the argument at 720 points is computed on the spot (agreeing with the computation on this machine at 16 positions). The limaçon is the curve $r = \frac{1}{2} + \cos t$ shifted by $(c_x, 0)$.

Computed **Table 2** $\frac{1}{2\pi}\oint d\theta$. Computed in two ways at 2×10^5 points; the difference from an integer is below 10^{-9}.

Closed curve	Winding number
Circle of radius 1, centre (0, 0) / reversed / twice round	1 / −1 / 2
Circle of radius 1, centre (2, 0) (does not enclose the origin)	0
Limaçon $c_x = -0.25 \,/\, -1 \,/\, 0.5$	2 / 1 / 0

The index of the zeros in the detour of 1-05 (the hairy ball theorem) is also this winding number. Outside the solenoid of 2-03 § 06, $A = \frac{\Phi}{2\pi}\,d\theta$, and although $dA = 0$ the circulation is Φ. The Aharonov–Bohm effect is the physics of a form that is closed but not exact.

08

An entrance to de Rham cohomology

Among closed forms, regard the exact ones as 0, and count what remains.

$$H^k(M) = \frac{\{\omega : d\omega = 0\}}{\{d\eta\}}, \qquad b_k(M) = \dim H^k(M) \tag{12}$$

b_k is called a **Betti number**. **Example.** H^0 consists of functions constant on each connected component, so b_0 is the number of connected components. $H^1(\mathbb{R}^2 \setminus \{0\})$ is the one dimension spanned by [dθ] (subtracting $c\,d\theta$ from a closed form ω so that the circulation on the unit circle is 0 makes it exact. Every closed curve can be deformed, without crossing the origin, into "a curve that goes round the unit circle some number of times", and deformation does not change circulation (the argument of § 07), so every circulation is 0. Then the line integral from a base point does not depend on the path and defines a function [Known]). The area form σ of the sphere is closed and $\int_{S2}\sigma = 4\pi$, but if $\sigma = d\eta$, then since S^2 has no boundary, (9) would make the integral 0. Hence it is not exact.

De Rham's theorem (1931), in today's language: H^k agrees with cohomology (with real coefficients) defined from topology alone, without forms [Known] . Examples: $\mathbb{R}^n$ is (1, 0, 0, ...), $\mathbb{R}^3 \setminus \{0\}$ is (1, 0, 1, 0), S^4 is (1, 0, 0, 0, 1) (the stage of 3-05).

The point magnetic charge. The 2-form $\beta = (x\,dy \wedge dz + y\,dz \wedge dx + z\,dx \wedge dy)/r^3$ on $\mathbb{R}^3 \setminus \{0\}$ is the magnetic field $\hat{\mathbf{r}}/r^2$, and $d\beta = 0$, but its integral over a closed surface enclosing the origin is 4π (difference below 10^{-9} for a sphere and a shifted cube, and 0 for a sphere not enclosing it [Computed]). Hence there is no A with $dA = \beta$ on the whole space. Off the negative z axis $A_N = (1 - \cos\vartheta)\,d\varphi$ can be used, and off the positive z axis $A_S = -(1 + \cos\vartheta)\,d\varphi$ (spherical coordinates); the difference $2\,d\varphi$ is a gauge transformation by the multivalued function 2φ [Computed] . Taking potentials separately on the two regions, north and south, and gluing them on the overlap by a gauge transformation is in § III of Wu–Yang (1975). The language of this gluing is bundles and connections (1-07).

If the total magnetic flux is Φ, the difference is $\frac{\Phi}{2\pi}\,d\varphi$ (here Φ = 4π). The phase factor $e^{iq\chi/\hbar}$ of a particle of charge q (2-03) is single-valued under $\chi = \frac{\Phi}{2\pi}\varphi$ when $q\Phi/h$ is an integer. The quantisation of the point magnetic charge is due to Dirac (1931), and the reading in the language of bundles is in 1-07 § 08.

Detour: differential forms on a lattice

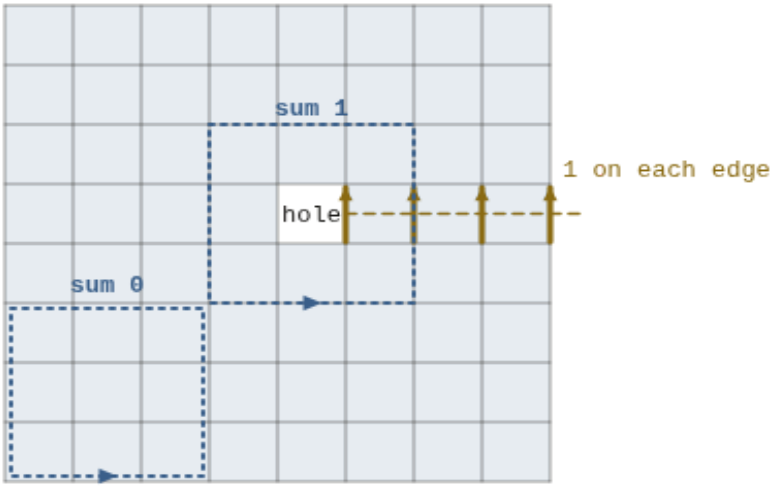
A subsection you may skip.

Values placed on the vertices, the oriented edges and the faces of a square lattice are taken as 0-, 1- and 2-forms; d_0 is the difference between the two ends of an edge, and d_1 the sum around the boundary of a face traversed anticlockwise. Each vertex appears once with + and once with − on the boundary of a face, so $d_1 d_0 = 0$ holds exactly as a matrix of integers (the reason the

gauge function disappeared from the face values on the lattice of 2-03). Table 3 shows the computation of $b_0 = \dim\ker d_0$, $b_1 = \dim\ker d_1 - \operatorname{rank} d_0$, $b_2 = F - \operatorname{rank} d_1$ (F is the number of faces).

Computed **Table 3** Betti numbers on lattices (side L = 8; the same for L = 4) and the known values for the corresponding smooth surfaces. The table sits inside Figure 4; choosing a row changes the figure.

Figure 4 Choosing a row of the table below shows that lattice (side 8), the numbers of vertices, edges and faces V, E, F, and the Betti numbers. For "Remove one face", the 1-form with 1 on the edges crossing the cut (dashed line) is drawn (thick arrows), together with its sums along two closed paths (dotted). It is 1 on a closed path around the hole and 0 on one that does not go around it.



`Remove one face  V = 81  E = 144  F = 63  (b₀, b₁, b₂) = (1, 1, 0)  V − E + F = 0 = b₀ − b₁ + b₂`

Lattice	(b₀, b₁, b₂)	Smooth case
Square	(1, 0, 0)	Disc (1, 0, 0)
Remove one face	(1, 1, 0)	Disc with one hole (1, 1, 0)
Remove two faces	(1, 2, 0)	Disc with two holes (1, 2, 0)
Periodic lattice	(1, 2, 1)	Torus T^2 (1, 2, 1)
Surface of a cube	(1, 0, 1)	Sphere S^2 (1, 0, 1)

Computed The numbers in this figure were computed on this machine (from the ranks of the integer matrices d_0, d_1). The figure only embeds and draws those numbers.

Regardless of fineness, the values of the smooth case come out (a small check of the theorem that the two cohomologies agree Known). In the numbers of Figure 4, $V - E + F = b_0 - b_1 + b_2$ also holds. It holds for every lattice, because the dimension of the kernel plus the rank is the number of columns Known . On the lattice with one hole, putting 1 only on the edges crossing a cut from the hole to the right edge gives a closed form; its sum around the hole is 1, and on a closed path not around the hole 0 Computed . It is the lattice version of dθ/2π. Putting group elements rather than numbers on the edges is Wilson's lattice gauge theory (5-01).

09

The Hodge dual and Maxwell's equations

Up to here we have not used a metric. The Hodge dual needs one, so in this section we define on the spot only as much as we use. A metric is given by the table of inner products η_{ij} between coordinate directions (here diagonal, with entries ± 1). In Minkowski spacetime (2-04), for coordinates (t, x, y, z), $\eta = \operatorname{diag}(-1, 1, 1, 1)$ (signature $(-,+,+,+)$), $c = \varepsilon_0 = \mu_0 = 1$); in Euclidean $\mathbb{R}^4$, $\eta = \operatorname{diag}(1, 1, 1, 1)$.

- **Volume form**: $\mathrm{vol} = dt \wedge dx \wedge dy \wedge dz$ (in the Euclidean case $dx^1 \wedge dx^2 \wedge dx^3 \wedge dx^4$).
- **Inner product of forms**: for basis elements $\langle dx^{i1} \wedge \cdots \wedge dx^{ik},\ dx^{i1} \wedge \cdots \wedge dx^{ik} \rangle = \eta_{i1 i1} \cdots \eta_{ik ik}$, and basis elements with different sets of indices give 0. For general forms extend bilinearly in the coefficients.
- **Hodge dual**: the linear map taking a k-form β to an (n−k)-form $*\beta$ such that $\alpha \wedge *\beta = \langle \alpha, \beta \rangle \,\mathrm{vol}$ for every k-form α.

Example. In Minkowski, from $\langle dt \wedge dx, dt \wedge dx \rangle = \eta_{tt}\eta_{xx} = -1$ and $dt \wedge dx \wedge dy \wedge dz = \mathrm{vol}$, $*(dt \wedge dx) = -dy \wedge dz$. From $\langle dy \wedge dz, dy \wedge dz \rangle = +1$ and $dy \wedge dz \wedge dt \wedge dx = \mathrm{vol}$, $*(dy \wedge dz) = dt \wedge dx$ (checked for all 6 basis 2-forms Computed). A −1 enters only for basis elements containing time, and this affects the sign in (14) and the sign of $**$ below.

We use it for the electromagnetic field. Gathering the potentials into a 1-form gives

$$A = -\phi\, dt + A_x\, dx + A_y\, dy + A_z\, dz, \qquad F = dA = -E_i\, dt \wedge dx^i + B_x\, dy \wedge dz + B_y\, dz \wedge dx$$

($\mathbf{E} = -\nabla\phi - \partial_t \mathbf{A}$, $\mathbf{B} = \nabla \times \mathbf{A}$). The four Maxwell equations become two.

$$dF = 0, \qquad d{*}F = *j, \qquad j = -\rho\, dt + J_x\, dx + J_y\, dy + J_z\, dz \quad (14)$$

$dF = 0$ is automatic from (6) and is the two source-free Maxwell equations of 2-03; $d{*}F = *j$ is the two Maxwell equations with sources (checked with random potentials Computed). That F does not change under the gauge transformation $A \mapsto A + d\chi$ is also (6) (the formula for gauge transformations in 2-03). What enters the action is $F \wedge *F = (|\mathbf{B}|^2 - |\mathbf{E}|^2)\,\mathrm{vol}$, and also $F \wedge F = -2(\mathbf{E} \cdot \mathbf{B})\,\mathrm{vol}$ Computed .

What happens only in four dimensions

$*F$ is an (n−2)-form, and has the same degree as F only for n = 4. The $**$ on 2-forms is +1 in Euclidean $\mathbb{R}^4$ (the eigenvalues ± 1 of $*$ three each) and −1 in Minkowski (eigenvalues $\pm i$) (agreeing with the formula $(-1)^{k(n-k)+s}$, where s is the number of negative signs, for every k Computed). In the Euclidean case it splits as $F = F^+ + F^-$ (self-dual and anti-self-dual), and $|F|^2 = |F^+|^2 + |F^-|^2 \geq \big||F^+|^2 - |F^-|^2\big| = |\langle F, *F \rangle|$.

Integrating gives $\int F \wedge *F \geq |\int F \wedge F|$, with equality when $F = \pm * F$; this is the skeleton of the instanton inequality (3-05).

In Minkowski, $** = -1$ on 2-forms, so if $*F = \lambda F$ then $-F = \lambda^2 F$, and the only real F satisfying $*F = \pm F$ is 0. In physics, instantons are considered in imaginary time (2-04, 4-02). For U(1), $F \wedge F = d(A \wedge F)$ is exact, so its integral is fixed by the boundary term (at infinity) alone. In the non-abelian case this becomes an integer-valued topological quantity (3-05).

See also the article on four dimensions and 1-12. In the non-abelian case $F = dA + A \wedge A$ (A is the geometric way of writing with values in anti-Hermitian matrices; the translation to the physicists' way is in 3-02), and in place of $dF = 0$ we have $dF + A \wedge F - F \wedge A = 0$ (Bianchi) (1-07, 3-03).

What this chapter can say

Item	What is said	What is not said
Exterior derivative, Stokes	(4)–(8), the reason for $d^2 = 0$, the example of the hemisphere, Table 1 Computed	That pullback commutes with d, the proof of the general Stokes theorem Known
Closed and exact, de Rham	Why dθ, the area form and the point magnetic charge are not exact, Tables 2 and 3, the 1-form of the cut in Figure 4 Computed	The general form of the Poincaré lemma, de Rham's theorem Known
Maxwell and the Hodge dual	The values of $*$ on the basis, the sign in (14), $F \wedge *F$, $F \wedge F$, the sign and eigenvalues of $**$ Computed	The non-abelian case (1-07, 3-03), the existence of instantons (3-05)
Yang–Mills	Why gauge fields can be written as forms, one of the special features of four dimensions	Shows nothing about the mass gap or the construction of the theory

Sources and reproduction

Item	Kind	Source / tools
Tables 1–3, the numbers in the text, Figures 1–4	**Computed on this machine**	Python 3.12, NumPy 1.26: `forms_algebra.py`, `stokes_cells.py`, `winding.py`, `cochains.py`, `monopole.py`, `maxwell_forms.py`, `hodge_basis.py`, `hemisphere.py`, `lattice_fig.py`. The figures are plain JavaScript (Figure 1 computes the determinant on the spot; the functions of Figures 2 and 3 were checked with node against Table 1 and `winding.py`; Figure 4 embeds the numbers from `lattice_fig.py`)

Item	Kind	Source / tools
Exterior derivative, Stokes, the Poincaré lemma, de Rham's theorem	Known (bibliographic data only)	M. Spivak, *Calculus on Manifolds* (1965); R. Bott, L. W. Tu, *Differential Forms in Algebraic Topology* (1982)
The origin of the exterior derivative	Primary source (bibliographic data only)	É. Cartan, Ann. Sci. ÉNS (3) 16 (1899) 239–332. Its placing as the origin of the exterior derivative relies on secondary sources
De Rham's theorem	Primary source (bibliographic data only; the attribution of the theorem checked in secondary sources)	G. de Rham, J. Math. Pures Appl. (9) 10 (1931) 115–200
Treating the point magnetic charge by gluing two regions (§ III)	Primary source (text checked)	T. T. Wu, C. N. Yang, Phys. Rev. D 12 (1975) 3845–3857
Quantisation of the point magnetic charge	Primary source (bibliographic data only; attribution following the text of Wu–Yang)	P. A. M. Dirac, Proc. R. Soc. Lond. A 133 (1931) 60

Next chapter: 1-07 Vector bundles and connections (the gluing of § 08).

Vector bundles and connections — a rule for comparing separate spaces at each point along a path

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — We write the rule for comparing the vector spaces at each point along a path (a connection) as a matrix-valued 1-form A, and derive in one sequence parallel transport, curvature $F = dA + A \wedge A$, holonomy and the gauge transformation $A \mapsto gAg^{-1} - (dg)g^{-1}$. From the gluing of the monopole an integer (the first Chern number) comes out.

Prerequisite chapters — 1-04 (Lie algebras, the adjoint representation) and 1-06 (1-forms, the exterior derivative, the point magnetic charge of § 08). Knowing the phase factor of 2-03 makes for quicker reading (what § 08 needs is defined on the spot).

Said first — This is a preparatory chapter, and shows nothing about the mass gap. Gauss–Bonnet, Ambrose–Singer and the integrality of Chern numbers are cited without proof (Known). The space of gauge orbits is in 3-04.

Order of this chapter

1. Why bundles and connections — you cannot subtract until you have carried
2. Vector bundles — transition functions, trivial bundles and twisted bundles
3. Connections and covariant derivatives — the formula for gauge transformations
4. Parallel transport — once round the sphere (Figure 1). Detour: Foucault's pendulum
5. Curvature — the mismatch around a small loop, a connection that curves even when constant (Figure 2)
6. Holonomy — the Wilson loop
7. The geometry of gauge transformations — the Bianchi identity
8. Gluing and integers — phase factors from north and south (Figure 3), the first Chern number

01

Why bundles and connections — you cannot subtract until you have carried

Suppose a field ψ takes values in $\mathbb{C}^N$ at each point ($N = 3$ for the colour of quarks). In gauge theory we allow the basis of $\mathbb{C}^N$ at each point (which direction to call "the first colour") to be chosen afresh point by point (3-01). Then $\psi(x+h) - \psi(x)$ is a difference of components written in separate bases, and its value changes when the bases are rechosen. Tangent vectors of the sphere are the same: the tangent planes at the north pole and at the equator are different things. So we prepare a rule P for carrying one space to the other along a path, and subtract after carrying (e_i is a coordinate direction).

$$D_i\psi(x) = \lim_{h\to 0} \frac{P_{x+he_i\to x}\,\psi(x+he_i) - \psi(x)}{h} \quad (1)$$

The carrying rule is called a **connection**, and the spaces at each point bundled together a **vector bundle**. **Example.** In electromagnetism, $N = 1$, and the carrying rule was the phase factor $e^{\frac{iq}{\hbar}\int \mathbf{A}\cdot d\mathbf{l}}$ (2-03). This chapter extends that to matrices, and treats the "gluing of two sheets" of 1-06 § 08 in the same language.

02

Vector bundles — locally a product, possibly twisted globally

A **vector bundle** $E \to M$ attaches to each point x of a manifold M (1-05) an N-dimensional space E_x (the **fibre**), so that over a small open set U_α it can be identified with $U_\alpha \times \mathbb{C}^N$. N is called the **rank**, and one with $N = 1$ a **line bundle**. It can be defined in the same way with fibre $\mathbb{R}^N$, and this is called a **real vector bundle**. A basis giving the identification is a **local frame**, and on an overlap the components in the two frames are related by a matrix.

$$\psi_\alpha = g_{\alpha\beta}\,\psi_\beta, \qquad g_{\alpha\beta}\,g_{\beta\gamma} = g_{\alpha\gamma} \quad (2)$$

$g_{\alpha\beta}$ is called the **transition function**, and the formula on the right, where three overlap, the cocycle condition; when the transition functions take values in a group G (U(1), SU(N) and so on), G is called the **structure group**. Given transition functions satisfying (2), a bundle can be made by gluing Known . A smooth choice of an element of E_x at each point is a **section**, and the field ψ is a section.

Trivial bundles and twisted bundles

A bundle that can be identified with $M \times \mathbb{C}^N$ ($M \times \mathbb{R}^N$ if real) over the whole space is called **trivial**. The following three are not trivial Known : the Möbius strip (a real line bundle with fibre $\mathbb{R}$; covering the circle with two arcs, the transition function is +1 on one part of the overlap and −1 on the other), the tangent bundle TS^2 (a real bundle of rank 2), and the monopole bundle of § 08 ($n \neq 0$). A trivial bundle would have a section that is nowhere 0. On the Möbius strip, the value of a section written in the frame changes sign in one turn, so by the intermediate value theorem it is 0 somewhere. On TS^2 the hairy ball theorem (1-05) forbids it.

Making the fibre complex untwists it with the same transition functions. $-1 = e^{i\pi}$ can be moved continuously inside U(1) to 1 through e^{it} (t from π to 0), and moving transition functions continuously keeps the bundle isomorphic Known , so the complex line bundle with transition functions +1 and −1 on the overlap is trivial. On the circle, whether a twist remains is decided by whether the transition function can be moved continuously to 1 inside the group where it takes values. Inside $\{\pm 1\}$ it cannot; inside U(1) it can. In § 08 the winding number in U(1) carries this impossibility of moving. Bundles over $\mathbb{R}^4$ are trivial Known , but on S^4, with infinity regarded as one point, and on periodic lattices, non-trivial bundles appear (3-05).

03

Connections and covariant derivatives — $D = d + A$

Once a frame is fixed, the first-order term of the P of (1) becomes a matrix-valued 1-form $A = A_i\, dx^i$. $A_i(x)$ is an element of the Lie algebra $\mathfrak{g}$ of the structure group (1-03), and for U(N), SU(N) it is an anti-Hermitian matrix.

$$D\psi = d\psi + A\,\psi, \qquad D_i\psi = \partial_i\psi + A_i\,\psi \quad (3)$$

This is what we get by putting $P_{x+he_i \to x} = 1 + hA_i + O(h^2)$ in (1). For $D'\psi' = g\,D\psi$ to hold even when the frame is changed to $\psi' = g\psi$ ($g(x) \in G$), since $(d + A')(g\psi) = dg\,\psi + g\,d\psi + A'g\psi$ equals $g(d\psi + A\psi)$,

$$A' = g\,A\,g^{-1} - (dg)\,g^{-1} \quad (4)$$

is needed. This is the **gauge transformation**. The A_α and A_β of an overlap are also related by (4) with $g = g_{\alpha\beta}$: **gluing and gauge transformation are the same formula**.

Example (U(1)). Putting $A = -\frac{iq}{\hbar}\,\mathbf{A}\cdot d\mathbf{x}$, $g = e^{iq\chi/\hbar}$, (3) is the covariant derivative of 2-03 and (4) is $\mathbf{A} \mapsto \mathbf{A} + \nabla\chi$. Physics books write $A = -i\,A^a_\mu T_a\,dx^\mu$ with Hermitian T_a (1-04), and call the real A^a_μ the gauge field (for where the coupling constant is put, see 3-

02). The difference of two connections changes as $g(A_1 - A_2)g^{-1}$ (the dg term cancels), and is a quantity that can be added and subtracted. The set of all connections is then an **affine space** — a set with no fixed origin, in which the difference of two points is an element of a vector space. Choosing one reference connection A_0, every connection can be written as A_0 + (a $\mathfrak{g}$-valued 1-form) (3-04).

04

Parallel transport — an ordinary differential equation along a path

Along a path $\gamma : [0,1] \to M$, carry v so that the covariant derivative is 0.

$$\frac{dv}{dt} + A\big(\dot\gamma(t)\big)\, v = 0, \qquad v(1) = P_\gamma\, v(0) \tag{5}$$

The solution is unique, and P_γ is called **parallel transport**. If A is anti-Hermitian, $\frac{d}{dt}|v|^2 = -v^*(A + A^*)v = 0$, so P_γ is unitary. The $A(\dot\gamma)$ at different times do not commute, so it cannot be written $\exp(-\int A)$; instead the solutions over short intervals are multiplied in order of time.

$$\begin{aligned} P_\gamma &= \lim_{n\to\infty} e^{-A_n \Delta t} \cdots e^{-A_1 \Delta t} \\ &=: \mathcal{P}\exp\Big(-\int_\gamma A\Big) \end{aligned} \tag{6}$$

$A_k = A(\dot\gamma(t_k))$ is the **path-ordered product**, with later times placed on the left. Joining paths gives $P_{\gamma_2\gamma_1} = P_{\gamma_2} P_{\gamma_1}$, and under gauge transformations it only undergoes the change of frame at the two ends.

$$P'_\gamma = g\big(\gamma(1)\big)\, P_\gamma\, g\big(\gamma(0)\big)^{-1} \tag{7}$$

Example: once round the sphere

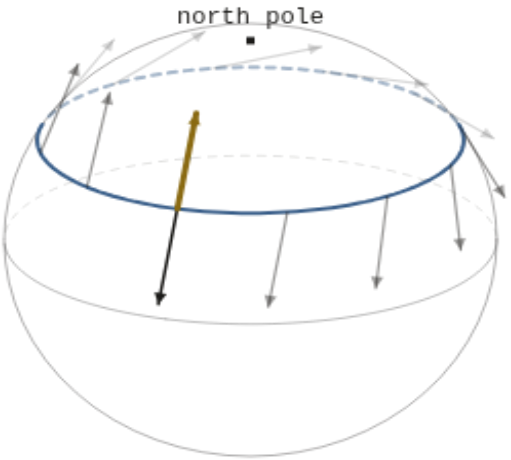
Carrying a tangent vector so that it does not bend as seen from inside the sphere is the Levi-Civita parallel transport; on the unit sphere it is $\dot v = -(v\cdot\dot\gamma)\,\gamma$ (allowing only change in the normal direction). A curve that keeps its own velocity vector $\dot\gamma$ under this parallel transport is called a **geodesic**; on the sphere these are great circles. Going east round the circle of latitude at colatitude θ_0, the angle relative to the south-pointing and east-pointing frame (e_θ, e_φ) changes by $\alpha(\varphi) = \alpha_0 - \varphi\cos\theta_0$, and in one turn

$$\Delta\alpha \equiv 2\pi(1 - \cos\theta_0) \pmod{2\pi} \tag{8}$$

The right-hand side is the area of the spherical cap enclosed by the circle of latitude. Circles of latitude other than the equator are not great circles (not geodesics), so the

vector turns relative to the frame. (8) is a special case of the Gauss–Bonnet theorem, "on a surface of curvature 1 the angle turned in one circuit equals the area enclosed" Known . For a **geodesic triangle**, whose three sides are geodesics, the angle turned also equals the area; the bottom row of Table 1 is an example.

Figure 1 Change the colatitude θ_0 with the slider (or a row of Table 1), and go east once round the circle of latitude with "Carry". After one turn the thick arrow is displaced from the thin arrow at the start (the faint arrows are in between), and the angle equals the area of the cap (8).



θ₀ = 60°, φ = 360°, α = -180.0°, (8) = 180.0°

θ_0	Turn in one circuit (RK4)	Difference from (8)
10°	5.469209°	2×10^{-12}
60°	**180.000000°**	**4×10^{-11}**
90°	0° (= 360°)	6×10^{-14}
Geodesic triangle (area $\pi/2$)	90.000000°	–

Computed **Table 1** The turn in one circuit obtained by integrating (5) with RK4 (angle relative to the frame, mod 360°). The bottom row is the geodesic triangle joining the north pole and the points at longitude 0° and 90° on the equator. The figure computes $\alpha = -\varphi\cos\theta_0$ in the browser. If JS is disabled, this table is the content of the figure.

Detour: Foucault's pendulum

A subsection you may skip.

The plane of oscillation appears to turn relative to the ground, but to be exact what turns is the frame of the ground (north, east), and the plane of oscillation is parallel-transported along the circle of latitude (in the approximation of slow rotation of the Earth) Known . At latitude λ north, put $\theta_0 = 90^\circ - \lambda$ in (8): in one sidereal day $-2\pi\sin\lambda$ (clockwise in the northern hemisphere). At latitude 48.85° north (near Paris) it is 271.08° Computed .

05

Curvature — the mismatch around a small loop

Whether parallel transport depends on the path can be measured by the mismatch after going round a small loop. In the difference of covariant derivatives taken in swapped order, the derivatives of ψ cancel, and only matrix multiplication remains.

$$[D_i, D_j]\,\psi = F_{ij}\,\psi, \qquad F_{ij} = \partial_i A_j - \partial_j A_i + [A_i, A_j] \quad (9)$$

Gathered into forms, it is $F = dA + A \wedge A$. dA is the exterior derivative component by component, and the wedge product of matrix-valued forms multiplies the component matrices and takes the wedge product of dx.

$$A \wedge A = \sum_{i,j} A_i A_j\, dx^i \wedge dx^j = \tfrac{1}{2} \sum_{i,j} [A_i, A_j]\, dx^i \wedge dx^j$$

The second equality is the average with the formula with the names i and j swapped ($dx^j \wedge dx^i = -dx^i \wedge dx^j$). For a scalar 1-form α, $\alpha \wedge \alpha = 0$ (1-06), but matrices do not commute, so it is not 0. The coefficient of $F = \sum_{i<j} F_{ij}\, dx^i \wedge dx^j$ is the F_{ij} of (9), exactly the formula of the Jaffe–Witten problem description (1-03). F is called the **curvature**, and in physics the **field strength**.

Checking with a small loop

Going once anticlockwise round a square of side ε in the x and y directions from the point p,

$$P_{\square} = 1 - \varepsilon^2 F_{xy}(p) + O(\varepsilon^3) \quad (10)$$

Putting (4) into (9) gives $F' = gFg^{-1}$. For U(1), F does not change under gauge transformations and is (apart from the factor $-iq/\hbar$) the electric and magnetic fields themselves (1-06 § 09), but in the non-abelian case F turns with the frame, and what does not change are combinations such as $\operatorname{tr}(F_{ij}F_{kl})$ (the action of 3-03).

Computed **Table 2** Integrating (5) with RK4 for a smooth $\mathfrak{su}(2)$-valued connection and comparing with (10) (operator norm, $|F_{xy}| = 0.369$).

ε	$\lvert \frac{P_{\square} - 1}{\varepsilon^2} + F_{xy} \rvert$	With dA in place of F
0.2	0.076538	0.348489
0.1	0.038844	0.324000
0.05	0.019532	0.311139

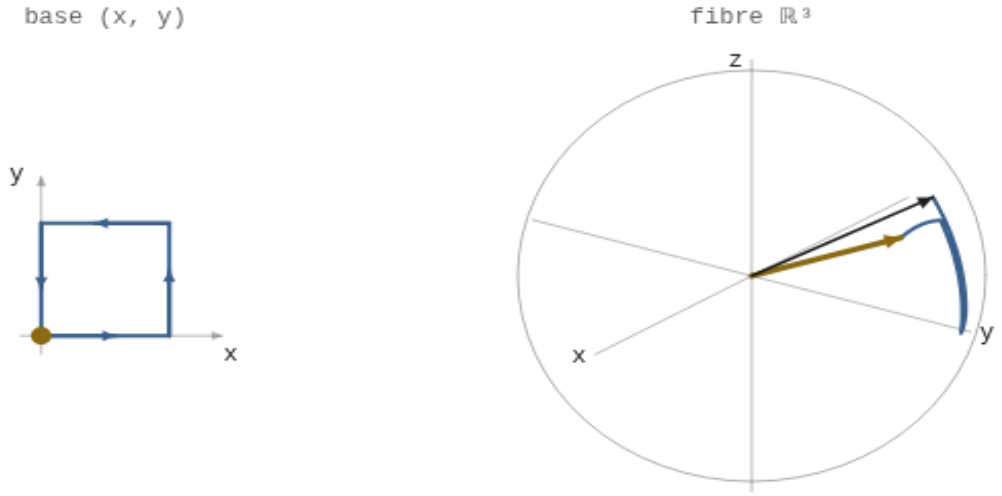
ε	$\lvert\frac{P_\square - 1}{\varepsilon^2} + F_{xy}\rvert$	With dA in place of F
0.025	0.009790	0.304592

Halving ε halves the error too ($O(\varepsilon)$). With dA, where the commutator is dropped, the error approaches $\lvert[A_x, A_y]\rvert = 0.298$ and does not vanish. For a random g, (7) differs by 1.5×10^{-12}, and $F' = gFg^{-1}$ by 3×10^{-7} (the error of differencing) Computed .

A connection that curves even when constant

In the non-abelian case F is not 0 even when A is constant. Take the group to be SO(3) (the adjoint representation of SU(2), 1-04) and the fibre $\mathbb{R}^3$ (the real vector bundle of § 02). Here L_x, L_y, L_z are the real antisymmetric matrices $(L_a)_{bc} = -\varepsilon_{abc}$, the Hermitian L_a of 1-04 § 02 multiplied by $-i$. They satisfy $L_a v = e_a \times v$, $[L_x, L_y] = L_z$, and e^{tL_z} is the rotation by angle t about the z axis (the same as the e^{-itL3} of 1-04). Putting $A_x = aL_x$, $A_y = aL_y$, we have $dA = 0$ and yet $F_{xy} = a^2 L_z$. A constant A for U(1) can be removed by $g = e^{A_x x + A_y y}$ (pure gauge), but here $F' = gFg^{-1}\ \not= 0$, so no g can remove it.

Figure 2 On the left is the square in the base, on the right the fibre $\mathbb{R}^3$. With "Carry", while the point on the left goes once round, the vector on the right moves by (5), and after one turn the thick arrow is displaced from the thin arrow at the start. The slider (or a row of Table 3) is $s = \varepsilon a$, and "Make commutative" is $A_y = aL_x$.



s = 0.60, angle between v_0 and v 19.52°, about the axis (-0.28, 0.28, -0.92) by 20.04° (s^2 = 20.63°)

s	Rotation angle θ (rad)	θ / s²	Rotation axis
0.1	0.009992	0.999168	(−0.050, 0.050, −0.998)
0.4	0.157919	0.986994	(−0.195, 0.195, −0.961)
1.2	1.297942	0.901348	(−0.492, 0.492, −0.719)

Computed **Table 3** The rotation angle and axis of the composition of the rotations along the four sides $e^{sL_y} e^{sL_x} e^{-sL_y} e^{-sL_x}$ ($\lvert P_\square - 1\rvert < 10^{-15}$ if commutative). The figure computes the same composition in the browser. If JS is disabled,

this table is the content of the figure.

θ/s² approaches 1 as s → 0, and the axis is almost −z, agreeing with the $1 - s^2 L_z$ of (10) (a positive rotation about −z). This is the geometric face of "the gauge field interacting with itself" (3-02).

06

Holonomy — what a closed path leaves behind

The parallel transport along a path γ that leaves the point x_0 and returns is called the **holonomy** Hol_γ. By (7), since both ends are the same point, it changes by conjugation under gauge transformations, and its trace does not change.

$$\begin{aligned} \mathrm{Hol}'_\gamma &= g(x_0)\,\mathrm{Hol}_\gamma\, g(x_0)^{-1}, \\ W(\gamma) &= \tfrac{1}{N}\,\mathrm{tr}\,\mathrm{Hol}_\gamma \end{aligned} \qquad (11)$$

The gauge-invariant $W(\gamma)$ is called the **Wilson loop**.

- **It remains even when flat.** If $F = 0$, the holonomy does not change under continuous deformation of the loop Known . In a space with a hole, the holonomy of a loop that cannot be shrunk can be other than 1; this is the Aharonov–Bohm effect (2-03).
- **Curvature is the infinitesimal version of holonomy.** If M is connected, the Lie algebra of the set (subgroup) of holonomies of loops at x_0 is spanned by the curvatures $F_{ij}(y)$ at all points y carried to x_0 along every path joining y and x_0 (Ambrose–Singer 1953) Known .
- **The lattice.** Placing parallel transports $U_\ell \in G$ on the edges and using the plaquette product U_p in place of curvature is Wilson's lattice (5-01). From (10), with spacing a, $U_p \approx \exp(-a^2 F_{\mu\nu})$, hence $\mathrm{Re}\,\mathrm{tr}(1 - U_p) \approx -\frac{a^4}{2}\,\mathrm{tr}\,F_{\mu\nu}^2 \geq 0$ (μ, ν fixed and not summed; $F_{\mu\nu}^2 = -F_{\mu\nu}^* F_{\mu\nu} \leq 0$ because $F_{\mu\nu}$ is anti-Hermitian) Known .

07

The geometry of gauge transformations — changing frames, and what does not change

A gauge transformation is a smooth map $g : M \to G$, which moves a section to $g\psi$, the connection by (4), and the curvature to gFg^{-1}. The whole $\mathcal{G}$ is an infinite-dimensional group. As a metaphor, it is a rechoosing of the orientation of the ruler at each point; to be exact, it can be read either as a change of frame (passive) or as an **automorphism** of the bundle — a map of the bundle that maps each fibre linearly and isomorphically

and does not move the points of the base — (active), and in both readings the formula is (4). What has meaning in physics are quantities that do not change under $\mathcal{G}$: Wilson loops, the $\operatorname{tr}(F \wedge *F)$ of the action, the topological quantity $\operatorname{tr}(F \wedge F)$ (3-05).

Curvature satisfies an identity independent of the gauge. $A \wedge F$ and $F \wedge A$ are also fixed by the same rule as in § 05 (multiply the component matrices and take the wedge product of dx), and the order of the matrices matters.

$$DF := dF + A \wedge F - F \wedge A = 0 \quad (12)$$

This is the **Bianchi identity**. Putting in $F = dA + A \wedge A$, it vanishes by $d(A \wedge A) = dA \wedge A - A \wedge dA$, $d^2 = 0$ and the associativity of the wedge product alone. For U(1) it is $dF = 0$ (the two Maxwell equations without sources). For a $\mathfrak{su}(2)$-valued connection in three dimensions, the left-hand side is within the differencing error 3.5×10^{-7}; dropping the commutator gives 0.553 Computed .

The $\mathcal{A}/\mathcal{G}$ obtained by dividing the set of all connections $\mathcal{A}$ by $\mathcal{G}$ is the space of configurations that physics can distinguish, and choosing a representative continuously from each orbit (a global gauge fixing) is in general impossible (Singer, 3-04, 4-08). In the translation table of Wu–Yang (1975), gauge potential = connection, phase factor = parallel transport and field strength = curvature are listed, and only the entry for sources is "?".

08

Gluing and integers — the monopole and the first Chern number

In 1-06 § 08, the potential of the point magnetic charge $\hat{\mathbf{r}}/r^2$ could be written only with two sheets, north and south, and the difference on the overlap was $2\,d\varphi$. Multiplying the magnetic charge by g_m to get the magnetic field $g_m\hat{\mathbf{r}}/r^2$ (total flux $\Phi = 4\pi g_m$), the difference is $2g_m\,d\varphi$, and multiplying by $-iq/\hbar$ to make a connection as in the example of § 03 gives

$$A_N - A_S = -\frac{iq}{\hbar}\,2g_m\,d\varphi$$

On the other hand, (4) in the commutative case is $A_N - A_S = -g_{NS}^{-1}dg_{NS}$, so the transition function on the equator is $g_{NS} = e^{in\varphi}$, $n = 2qg_m/\hbar$. In the language of bundles, it is the U(1) bundle obtained by gluing the north and south sheets along the equator by this g_{NS}. g_{NS} must return to itself after going once round the equator, so **n is an integer**.

Seen through phase factors

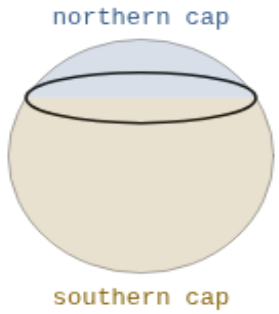
Seen through phase factors, the meaning of n takes a measurable form. The **phase factor** of a closed path C is the parallel transport (the holonomy of C, § 06) obtained by putting $A = -\frac{ig}{\hbar}\,\mathbf{A}\cdot d\mathbf{x}$ in (6). For U(1) no path ordering is needed, and by Stokes' theorem (1-06)

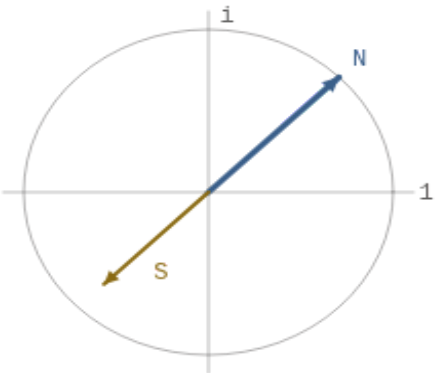
$$\mathrm{Hol}_C = \exp\!\left(\frac{iq}{\hbar}\oint_C \mathbf{A}\cdot d\mathbf{l}\right) = e^{iq\Phi C/\hbar} = e^{2\pi i\,\Phi C/\Phi 0}, \qquad \Phi_0 = \frac{h}{q}$$

(ΦC is the magnetic flux enclosed by C, and we take q > 0; for details see 2-03). $\Phi 0$ is called the **flux quantum**, and writing $f = \Phi/\Phi_0$, we have $n = 2qg_m/\hbar = q\Phi/(2\pi\hbar) = f$. **From here on n and f are the same quantity**; when we write f we also consider values that are not necessarily integers.

Computing the phase factor of the circle of latitude C at colatitude ϑ with A_N, usable on the northern cap, the enclosed flux is $\Phi(1-\cos\vartheta)/2$, giving $e^{i\pi f(1-\cos\vartheta)}$. Computing with A_S, usable on the southern cap, C goes round the boundary of the southern cap in the opposite direction, giving $e^{-i\pi f(1+\cos\vartheta)}$. The ratio is $e^{2\pi i f}$, and they agree only when f is an integer. This is the condition for the two descriptions to be a connection on one bundle, and is the reading, in the language of bundles, of Dirac's quantisation condition (1931).

Figure 3 Changing f and the colatitude ϑ with the sliders (or a row of Table 4) moves two arrows on the unit circle — the phase factor computed from the northern cap and the one computed from the southern cap. The two coincide only when f is an integer, and then they move together even when ϑ is changed. "Move f" increases f from 0 to 3.





f = 0.50, ϑ = 60°, north 45.0°, south -135.0°, mismatch 180.0°

f	From the north	From the south	Winding number of the transition function
0.5	0.707+0.707i	−0.707−0.707i	0.5 (does not return after one turn)
1	i	i	1
2	−1	−1	2

Computed **Table 4** On the circle of latitude at colatitude 60°. The winding number is the increase of the argument of $e^{if\varphi}$ once round the equator ÷ 2π. $\frac{i}{2\pi}\int F$ is f in every row. The figure computes the closed formulas of the two phase factors in the browser, and the angle of mismatch is $360^\circ \times f$ (integer multiples of 360° removed). If JS is disabled, this table is the content of the figure.

Read from curvature

The integer can also be read from curvature. Using Stokes' theorem on the northern and southern caps, and (4) on the equator (commutative, so $A_N - A_S = -g^{-1}dg$),

$$c_1 = \frac{i}{2\pi}\int_{S^2} F = \frac{1}{2\pi i}\oint_{\text{equator}} g_{NS}^{-1}\, dg_{NS} = n \quad (13)$$

The left-hand side is $\frac{i}{2\pi}\cdot\left(-\frac{ig}{\hbar}\right)\Phi = f$, and the right-hand side is the winding number of the transition function (1-06 § 07); the n = f seen above appears here too. c_1 is called the **first Chern number**. Changing the connection continuously cannot make an integer jump, so it is fixed by the bundle alone Known .

In four dimensions

SU(2) bundles over S^4 are classified by an integer $c_2 = \frac{1}{8\pi^2}\int \mathrm{tr}(F\wedge F)$ (the sign depends on the book) Known . On $\mathbb{R}^4$ too, solutions of the Yang–Mills equations with finite action approach a pure gauge $-(dg)g^{-1}$ ($g : S^3 \to SU(2)$) far away, and the winding number of g is the instanton number Known . This is not an unconditional fact. To be exact, it rests on the theorem that such solutions extend to connections on a bundle over S^4, obtained by adding one point at infinity (Uhlenbeck's removable singularity theorem); for general fields that are not solutions of the equations, the same is said by assuming suitable decay far away. For details see 3-05, 1-12 and the article on four dimensions.

What this chapter can say

Item	What is said	What is not said
Bundles, connections, parallel transport	Derivation of (2)–(8). Table 1, Figure 1 Computed	The non-triviality of the examples of § 02, that moving transition functions continuously keeps the bundle isomorphic, the general form of Gauss–Bonnet Known
Curvature and holonomy	(9)–(12). Tables 2 and 3, Figure 2, checks of (7) and $F' = gFg^{-1}$ Computed	Proofs of Ambrose–Singer and of the properties of flat connections Known

Item	What is said	What is not said
The monopole and the Chern number	That an integer comes out of the gluing, $n = \Phi/\Phi_0$, (13), Table 4, Figure 3 Computed	The general theory of Chern numbers, the classification by c_2, the extension of fields on $\mathbb{R}^4$ to S^4 Known
Yang–Mills	The geometric meaning of gauge fields, field strengths and gauge transformations	Shows nothing about the mass gap or the construction of the theory

Sources and reproduction

Item	Kind	Source / tools
Tables 1–4, the numbers of § 04–§ 07, Figures 1–3	**Computed on this machine**	Python 3.12, NumPy, SciPy (`sphere_holonomy.py` and 3 others). The figures are plain JavaScript, computing closed formulas on the spot
The ordering of the path-ordered product, the translation table (Table I), the monopole	Primary source (text checked)	T. T. Wu, C. N. Yang, Phys. Rev. D 12 (1975) 3845–3857
$F = dA + A \wedge A$	Primary source (text checked)	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory” (Clay)
Parallel transport, holonomy	Known (text checked)	S. M. Carroll, arXiv:gr-qc/9712019 (1997), Chapter 3
Levi-Civita, Ambrose–Singer, Dirac, Chern classes	Primary source (bibliographic data only)	T. Levi-Civita, Rend. Circ. Mat. Palermo 42 (1916) 173–204; W. Ambrose, I. M. Singer, Trans. AMS 75 (1953) 428–443; P. A. M. Dirac, Proc. R. Soc. A 133 (1931) 60–72; S. S. Chern, Ann. of Math. 47 (1946) 85–121, doi:10.2307/1969037 (page range checked in the zbMATH bibliographic data)
Four dimensions in § 08: removable singularities	Primary source (bibliographic data only)	K. Uhlenbeck, “Removable singularities in Yang–Mills fields”, Comm. Math. Phys. 83 (1982) 11–29, doi:10.1007/BF01947068
Seeing Foucault's pendulum as parallel transport	Paper (bibliographic data only), checked in secondary sources	J. B. Hart, R. E. Miller, R. L. Mills, “A simple geometric model for visualizing the motion of a Foucault pendulum”, Am. J. Phys. 55 (1987) 67–70, doi:10.1119/1.14972 (bibliographic data only); Wikipedia “Foucault pendulum” (text checked)
The examples of § 02, Gauss–Bonnet, c_2	Known (bibliographic data only)	Kobayashi–Nomizu, *Foundations of Differential Geometry* I (1963); Nakahara, *Geometry, Topology and Physics* (1990)

Next chapter: 1-08 Probability and measure. This chapter continues in 3-01–3-04.

Probability and measure — the mean settles down, and the fluctuations become Gaussian

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — The language that writes probability as "a rule that distributes weight over sets (a measure)". That the Gaussian distribution is fixed by its covariance alone, and higher moments are a count of pairings. That the law of large numbers and the central limit theorem say "the mean settles down, and the fluctuations become Gaussian", and that the rate is $1/\sqrt{n}$.

Prerequisite chapters — None (the example of § 03 is easier to read with 1-02).

Said first — We do not prove measure theory. Probability in infinite dimensions is the subject of 1-09. Constructing four-dimensional Yang–Mills theory as a probability measure is unsolved, and this site does not show it either. What the construction published in 2026-11 (6-00) made is correlation functions and a field theory reconstructed from them, not a probability measure (6-04).

Order of this chapter

1. Why probability
2. Probability spaces
3. Random variables and distributions — the uniform distribution on a group and the semicircle (Figure 1)
4. Expectation, variance, characteristic function
5. Independence
6. The Gaussian distribution — covariance and counting pairings (Figure 2, Detour: counting the Catalan numbers in two ways)
7. Gaussian fields on a lattice — correlation length and mass (Figure 3)
8. The law of large numbers — the mean settles down (Figure 4)
9. The central limit theorem — the fluctuations become Gaussian (Figure 4)

10. Detour: what does not become Gaussian

01

Why probability — field theory is written with probability measures

Moving quantum field theory to imaginary time (Euclideanisation, 4-02) attaches a weight e^{-S} (S is the action) to each configuration of the field, and after normalisation it can be read as a **probability**. The problem statement too lists 'non-Gaussian, Euclidean-invariant measures' among the by-products of constructing field theories Known . In lattice Yang–Mills a group element U_ℓ sits on each edge ℓ (5-01), and the probability has the following form.

$$d\mu(U) = \frac{1}{Z}\, e^{-S(U)} \prod_{\ell} dU_\ell, \qquad Z = \int e^{-S(U)} \prod_{\ell} dU_\ell \tag{1}$$

dU_ℓ is the uniform weight on the group (§ 03), and Z the constant that makes the total 1 (2-08). **Example:** if there is one edge, the group is the circle $U(1)$ and $S = -\beta \cos\theta$ (β is the inverse of the lattice coupling; the same letter as the inverse temperature of statistical mechanics), then (1) has density $e^{\beta\cos\theta}/Z$, and the larger β is, the more it gathers near $\theta = 0$. This chapter assembles the language for this.

02

Probability spaces — distributing weight over sets

A probability is a rule that distributes weights from 0 to 1 over subsets (**events**) of the whole Ω of what can happen. The family $\mathcal{F}$ of events that receive weights is assumed to contain Ω and to be closed under complements and countable unions (a **σ-algebra**). The triple $(\Omega, \mathcal{F}, P)$ is a **probability space**, and on P we impose the following Known .

$$P(\Omega) = 1, \qquad P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i) \quad (A_i \text{ pairwise disjoint}) \tag{2}$$

A non-negative set function with $\mu(\emptyset) = 0$ satisfying the second half of (2) (**countable additivity**) is a **measure**, and a measure with total 1 is a probability (Durrett § 1.1; the axiomatisation is said to begin with Kolmogorov's book of 1933). $\mathcal{F}$ is not taken to be all subsets because a construction using the axiom of choice produces subsets to which no length invariant under rotation and satisfying (2) can be assigned (Vitali 1905) Known . **Example:** a die is $P(A) = |A|/6$. The uniform distribution on $[0,1]$ is $P([a,b]) = b - a$, that is, length (**Lebesgue measure**). The probability of each point is 0, yet some

point always comes out; but (2) cannot be used for uncountable unions, so there is no contradiction. Something that holds with probability 1 is said to hold "**almost surely**".

03

Random variables and distributions — the uniform distribution on a group and the semicircle

A **random variable** is a real-valued function X on Ω such that every $\{X \le a\}$ is an event (**measurable**). Its **distribution** is the measure obtained by pushing the weight of Ω onto the number line by X.

$$\mu_X(B) = P\big(X^{-1}(B)\big), \qquad F_X(a) = P(X \le a), \qquad \mu_X(B) = \int_B \rho(x)\,dx \quad (3)$$

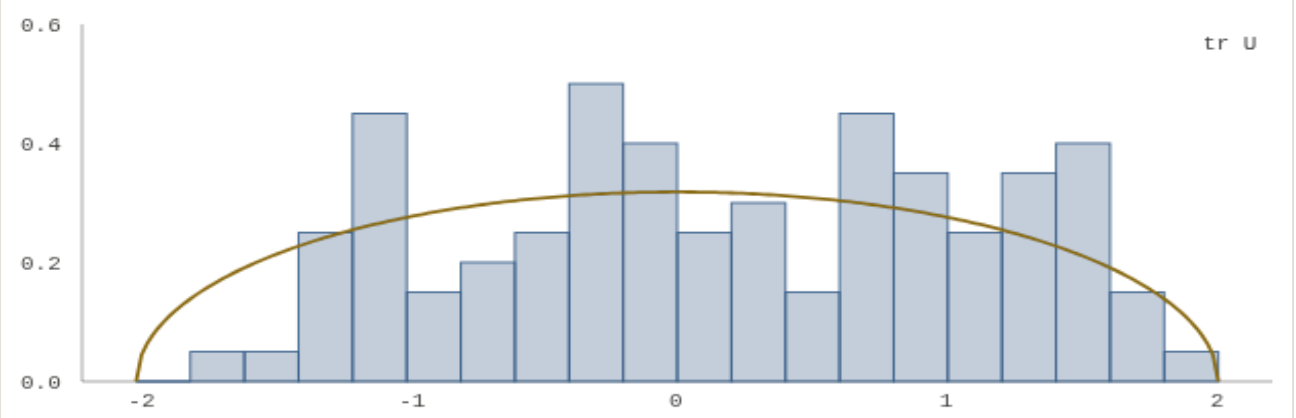
F_X is the **distribution function**, and ρ the **density**. **Example:** the number of heads in n tosses of a coin with probability of heads p follows the **binomial distribution** $\binom{n}{k}p^k(1-p)^{n-k}$, and keeping $np = \lambda$ while $n \to \infty$ gives the **Poisson distribution** $e^{-\lambda}\lambda^k/k!$.

A group of matrices that is closed and bounded, such as $SU(2)$, is called a **compact** group. A compact group G has exactly one probability measure unchanged by multiplication from either left or right, that is, satisfying $\mu(gA) = \mu(Ag) = \mu(A)$ ($gA = \{ga : a \in A\}$) for every $g \in G$ and subset A; it is called the **Haar measure** (for the circle, $d\theta/2\pi$) Known . The dU_ℓ of (1) is this. **Example:** elements of $SU(2)$ are $U = a_0 + i(a_1\sigma_1 + a_2\sigma_2 + a_3\sigma_3)$ (σ_i are the Pauli matrices, $|a| = 1$), and $SU(2)$ is the sphere S^3 (1-02). The Haar measure is the uniform distribution on S^3, and $\operatorname{tr} U = 2a_0$. The cross-section at height a_0 is a sphere of radius $r = \sqrt{1-a_0^2}$ (area $\propto r^2$), and the width of the band is da_0/r, so the density of a_0 is proportional to r, and $\operatorname{tr} U$ has the **semicircle** density Known .

$$\rho(x) = \frac{1}{2\pi}\sqrt{4 - x^2} \qquad (-2 \le x \le 2) \quad (4)$$

Making a histogram (20 bins) of $\operatorname{tr} U$ from 10^6 uniform points on S^3, the difference between the fraction falling in each bin and the probability of each bin given by (4) (0.019–0.064) was within 6.4×10^{-4} Computed . Figure 1 shows how it goes as the points increase. The same semicircle also appears as the distribution of eigenvalues of a large symmetric matrix with independent Gaussian entries, in the limit where it is suitably normalised and the size of the matrix goes to infinity (17-04).

Figure 1 The histogram of $\operatorname{tr} U$ (bars, 20 bins, area 1) made from the first n of the uniform points on S^3 (U following the Haar measure of $SU(2)$), and the semicircle (4) (curve). Increasing n from 10 to 10^6 with the slider or "Play" brings the bars towards the curve



n = 100, largest difference between bin fraction and bin probability 0.038 (1/√n = 0.10)

Computed The histogram of Figure 1 embeds the counts per bin of the first n points of a sample made on this machine (the same 10^6 points as in the text above). The curve and the bin probabilities are computed from formula (4). If JavaScript is disabled, it can be read from the numbers in the text above.

04

Expectation, variance, characteristic function — writing the mean as an integral

The **expectation** is the total of "value × weight", and for continuous cases an integral. The Lebesgue integral slices the values instead, and adds up c times "the weight of the set where the value is near c". As a metaphor, it is counting a pile of coins by denomination; to be exact, it is the limit of the sums $\sum_i c_i P(A_i)$ for functions taking only finitely many values.

$$E[X] = \int_\Omega X \, dP = \int_{\mathbb{R}} x \, d\mu_X(x), \qquad \operatorname{Var}(X) = E\big[(X - E[X])^2\big] \tag{5}$$

Taking the expectation of $\mathbf{1}_{\{|Y| \ge t\}} \le Y^2/t^2$ ($\mathbf{1}_A$ is 1 on A and 0 outside) as $Y = X - E[X]$ gives **Chebyshev's inequality** Known .

$$P\big(|X - E[X]| \ge t\big) \le \frac{\operatorname{Var}(X)}{t^2} \tag{6}$$

The **characteristic function** $\varphi_X(t) = E[e^{itX}]$ determines the distribution uniquely (Durrett § 3.3). **Example:** for a die, $E = 7/2$, $\operatorname{Var} = 35/12$. The $E[(-1)^N] = e^{-2\lambda}$ of the Poisson distribution is used in the counting of centre vortices (17-02). For (4), $E[\operatorname{tr} U] = 0$, $E[(\operatorname{tr} U)^2] = 1$.

05

Independence — how it differs from being uncorrelated

X and Y being **independent** means that the joint probability splits into a product.

$$P(X \in A,\ Y \in B) = P(X \in A)\,P(Y \in B) \qquad (\text{for all } A, B) \tag{7}$$

Then the distribution of (X, Y) is the measure that gives the weight $P(X \in A)P(Y \in B)$ to the rectangle $A \times B$ in the plane (the **product measure**), and $E[XY] = E[X]E[Y]$, $\operatorname{Var}(X+Y) = \operatorname{Var} X + \operatorname{Var} Y$, $\varphi_{X+Y} = \varphi_X \varphi_Y$ Known . If the **covariance** $E[XY] - E[X]E[Y]$ is 0 they are **uncorrelated**; independent implies uncorrelated, but the converse is false. **Example:** if X takes $\{-1, 0, 1\}$ with equal probability and $Y = X^2$, the covariance is 0 but $P(X = 0, Y = 0) = 1/3 \ \neq 1/9$.

In field theory we would like the fields at two distant points to be "almost independent". The problem statement gives **clustering** as a consequence of the mass gap Δ (4-11) Known . It is the property that the size of the correlation of local quantities $O(\underline{x})$, $O(\underline{y})$ with expectation 0 placed at points $\underline{x}, \underline{y}$ of space $\mathbb{R}^3$ (since the expectations are 0, the covariance above is just the expectation of the product) is, if $0 < C < \Delta$, at most $e^{-C|\underline{x}-\underline{y}|}$ where $|\underline{x} - \underline{y}|$ is large enough. To be exact, O are operators of quantum theory, and the correlation is the expectation in the ground state (the vacuum) (an example is in § 07).

06

The Gaussian distribution — covariance and counting pairings

A Gaussian distribution with mean 0 in n dimensions is fixed by the **covariance matrix** $C_{ij} = E[x_i x_j]$ alone (a positive definite symmetric matrix, 1-01).

$$\rho(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det C}} \exp\Big(-\frac{1}{2}\, x \cdot C^{-1} x\Big), \qquad E\big[e^{i\,t\cdot x}\big] = e^{-\frac{1}{2}\,t\cdot C t} \tag{8}$$

There are three properties Known . (i) The image under a linear map $y = Mx$ is also Gaussian, with covariance MCM^{T}. (ii) If jointly Gaussian, uncorrelated implies independent (if C is diagonal the density splits into a product). (iii) Higher moments are the sum over **all ways of pairing up** the variables of products of C (Isserlis 1918; in physics a formula of the same form is used as Wick's theorem (1950)).

$$E[x_1 x_2 x_3 x_4] = C_{12}C_{34} + C_{13}C_{24} + C_{14}C_{23} \tag{9}$$

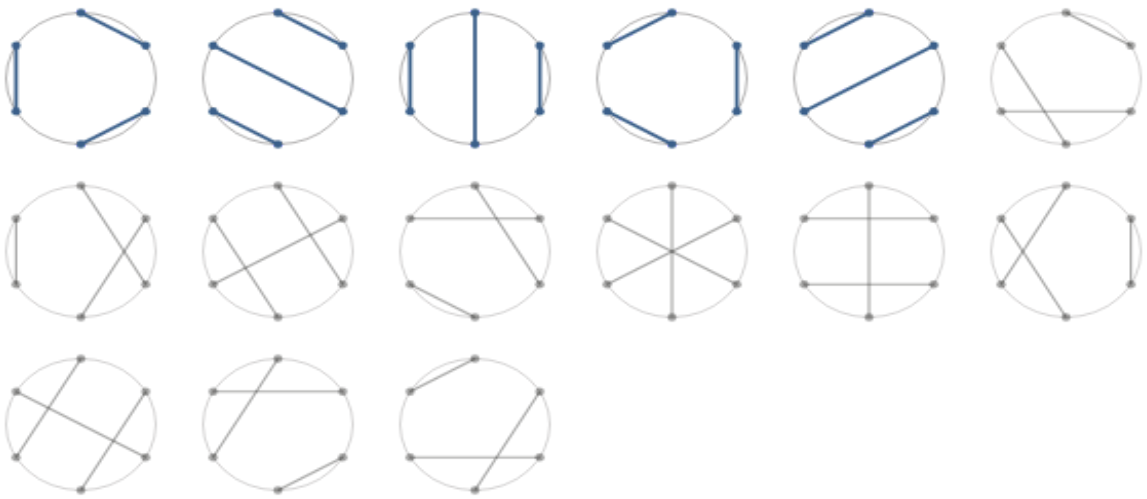
For random C of 4×4, against the right-hand side -0.0769 of (9), 4×10^6 samples gave -0.0774 ± 0.0005 Computed . In one variable it is $E[x^{2k}] = (2k-1)!!\,\sigma^{2k}$. $(2k-1)!! = 1 \cdot 3 \cdot 5 \cdots (2k-1)$ (the **double factorial**, multiplying only the odd numbers) is the number of ways of pairing up $2k$ points (the partner of point 1 can be chosen in $2k-1$ ways, and the same is repeated for the rest). On the other hand, the moments of the semicircle (4) are the number of ways of joining $2k$ points on a circle by chords that **do**

not cross, the **Catalan numbers** $\mathrm{Cat}_k = \binom{2k}{k}/(k+1)$ Paper (the detour computes both). Counting the non-crossing pairings and numerically integrating (4) both agreed with Cat_k for $k \le 6$ (the integral to 6 digits) Computed . Choosing a row of Table 1 lines up those ways of pairing for k in Figure 2.

Table 1 All pairings (Gaussian) and non-crossing pairings (semicircle). Choosing a row changes Figure 2 Computed

k	$(2k-1)!!$	Cat_k	$E[(\mathrm{tr}\,U)^{2k}]$ **(sample 10^6)**
1	1	1	1.001
2	3	2	2.001
3	**15**	**5**	**5.001**
4	105	14	14.00
5	945	42	42.01
6	10395	132	132.05

Figure 2 Ways of pairing $2k$ points on a circle by chords. Choose k by a row of Table 1 (or the slider). The coloured ones are the pairings whose chords do not cross (Cat_k ways), the faint ones the pairings that cross. The three for $k = 2$ correspond to the three terms $C_{12}C_{34}$, $C_{14}C_{23}$, $C_{13}C_{24}$ of (9) (number the points 1–4 clockwise)



k = 3, pairings 15 ((2k−1)!!), non-crossing 5 (Cat_3)

Computed The pairings of Figure 2 are listed in the browser, and their numbers have been checked to agree with the counts of Table 1 (computed on this machine). If JavaScript is disabled, they can be read from Table 1.

(9) is the prototype of Feynman diagrams (4-03), and non-crossing pairings lead to the planar diagrams of large N (17-04).

Detour: counting the Catalan numbers in two ways

A subsection you may skip.

The number of non-crossing pairings. Number the points on the circle 1 to $2k$ clockwise, and let c_k ($c_0 = 1$) be the number of non-crossing pairings. If the partner of point 1 is point p, the points between 1 and p cannot cross the chord $\{1, p\}$, so they can only pair among themselves. There is an even number of points between, so $p = 2j$, and the $j-1$ pairings inside and the $k-j$ pairings outside can be chosen separately. Written with the coefficients lined up in one power series (the **generating function**) $c(x) = \sum_{k\ge 0} c_k x^k$,

$$c_k = \sum_{j=1}^{k} c_{j-1} c_{k-j} \quad (k \ge 1) \qquad \Longleftrightarrow \qquad c(x) = 1 + x\, c(x)^2$$

Solving the quadratic equation, taking the root that is finite at $x \to 0$, and expanding $\sqrt{1-4x}$ by the binomial series gives the coefficients.

$$c(x) = \frac{1 - \sqrt{1-4x}}{2x}, \qquad \sqrt{1-4x} = 1 - \sum_{n\ge 1} \frac{2}{n}\binom{2n-2}{n-1} x^n \qquad \Longrightarrow \qquad c_k = \frac{1}{k+1}\binom{2k}{k}$$

The moments of the semicircle. Putting $x = 2\sin\theta$ in (4) gives $\sqrt{4-x^2} = 2\cos\theta$, $dx = 2\cos\theta\, d\theta$. By integration by parts $I_n = \int_{-\pi/2}^{\pi/2} \sin^n\theta\, d\theta$ satisfies $I_n = \frac{n-1}{n} I_{n-2}$, $I_0 = \pi$, so $I_{2k} = \pi\,(2k-1)!!/(2k)!!$ ($(2k)!! = 2\cdot 4 \cdots 2k = 2^k k!$). Using $\cos^2\theta = 1 - \sin^2\theta$,

$$\int_{-2}^{2} x^{2k} \rho(x)\, dx = \frac{2^{2k+2}}{2\pi}\left(I_{2k} - I_{2k+2}\right) = \frac{2^{2k+1}}{\pi} \cdot \frac{I_{2k}}{2k+2} = \frac{2^{2k}\,(2k-1)!!}{(k+1)\,(2k)!!} = \frac{1}{k+1}\binom{2k}{k}$$

The last equality follows from $(2k-1)!! = (2k)!/(2^k k!)$. The two computations arrived separately at the same Cat_k Paper . Why the semicircle counts non-crossing pairings becomes visible in the counting of planar diagrams of large N (17-04).

07

Gaussian fields on a lattice — correlation length and mass

To N variables $x_0, \dots, x_{N-1}$ ($x_N = x_0$) arranged on a ring, attach a weight under which neighbours prefer close values and the whole is pulled back towards 0.

$$\rho(x) \propto \exp\Big(-\frac{1}{2}\sum_j \left[(x_{j+1} - x_j)^2 + m^2 x_j^2\right]\Big) = \exp\Big(-\frac{1}{2}\, x \cdot (-\Delta_{\mathrm{lat}} + m^2)\, x\Big) \qquad (10)$$

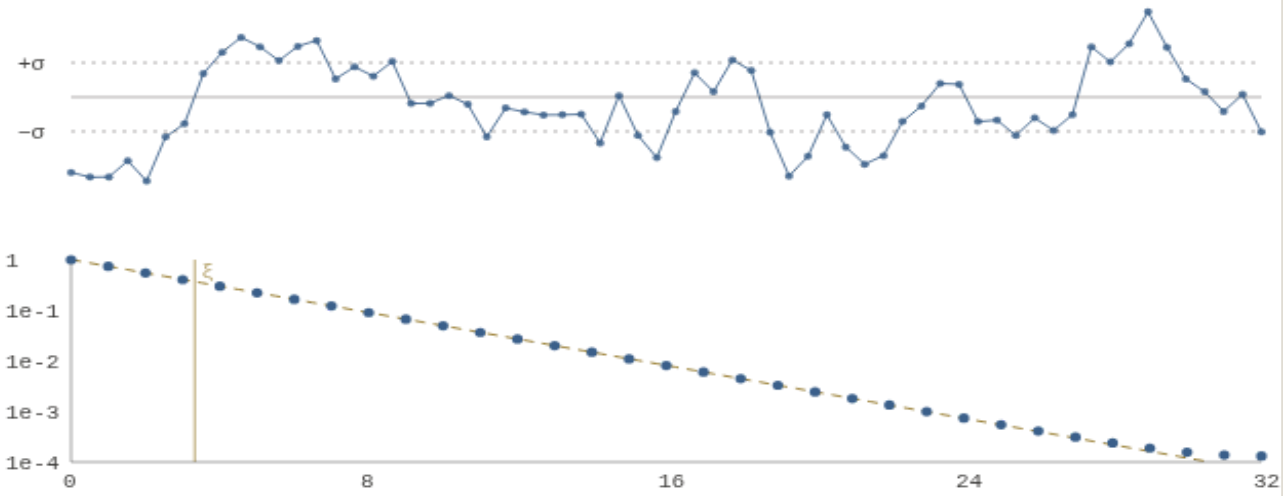
$(\Delta_{\mathrm{lat}} x)_j = x_{j+1} - 2x_j + x_{j-1}$ is the **lattice Laplacian**. It is different from the mass gap Δ of § 05, so we distinguish it with the subscript lat. Comparing with (8), (10) is the Gaussian distribution with covariance $C = (-\Delta_{\mathrm{lat}} + m^2)^{-1}$. The problem statement calls the Gaussian measure on the continuum $\mathbb{R}^d$ with a covariance of the same form (with the lattice Laplacian replaced by the continuum one) 'the free, mass-m0 field'. (10) is its analogue on a one-dimensional lattice (a ring), and on a d-

dimensional lattice too the covariance has the same form $(-\Delta_{\text{lat}} + m^2)^{-1}$ Known . If the ring is infinitely long, $C(0,x) = e^{-|x|/\xi}/(2\sinh(1/\xi))$ exactly Known .

$$\frac{C(0,x)}{C(0,0)} \approx e^{-|x|/\xi}, \qquad \cosh\frac{1}{\xi} = 1 + \frac{m^2}{2}, \qquad \xi \approx \frac{1}{m} \quad (m \text{ small}) \tag{11}$$

ξ is the **correlation length**, and the mass m is roughly its inverse (both measured in units of the lattice spacing). This rate of decrease corresponds to the "gap" of quantum theory (2-06). (10) is the discretised imaginary-time path integral of an oscillator, and m corresponds to the $a\omega$ of 2-07 (there $\omega = 1$, so it is the step a itself). The general theory of correlation length is in 2-08.

Figure 3 A Gaussian field on a ring ($N = 64$). Moving m changes the smoothness of the sample above (random numbers fixed; the dotted line is $\pm\sigma$) and the $C(0,x)/C(0,0)$ below (log scale, dashed line $e^{-x/\xi}$)



m = 0.302, ξ = 3.32, 1/m = 3.31, σ = 1.279, C(0,8)/C(0,0) = 0.0901

Table 2 Correlations on a ring ($N = 64$) Computed

m	ξ ((11))	$C(0,8)/C(0,0)$	ξ read from the slope
0.05	20.00	0.703	24.1
0.1	10.00	0.452	10.2
0.3	3.35	0.0915	3.35
1	1.04	4.5×10^{-4}	1.04
2	0.57	7.5×10^{-7}	0.57

Computed The numbers of Figure 3 and Table 2 were computed on this machine (the figure is drawn from the exact correlations and fixed random numbers; the slope from $x = 4$–12). If JavaScript is disabled, they can be read from Table 2.

For $m = 0.05$ the ξ read from the slope is larger than 20 because the ring is only about 3 times the correlation length, and the contribution $e^{-(N-x)/\xi}$ from going the other way round matters (a finite-box effect, 5-09). (10) is Gaussian, so by (9) all correlations are fixed by C, and there is no interaction (a **free field**).

08

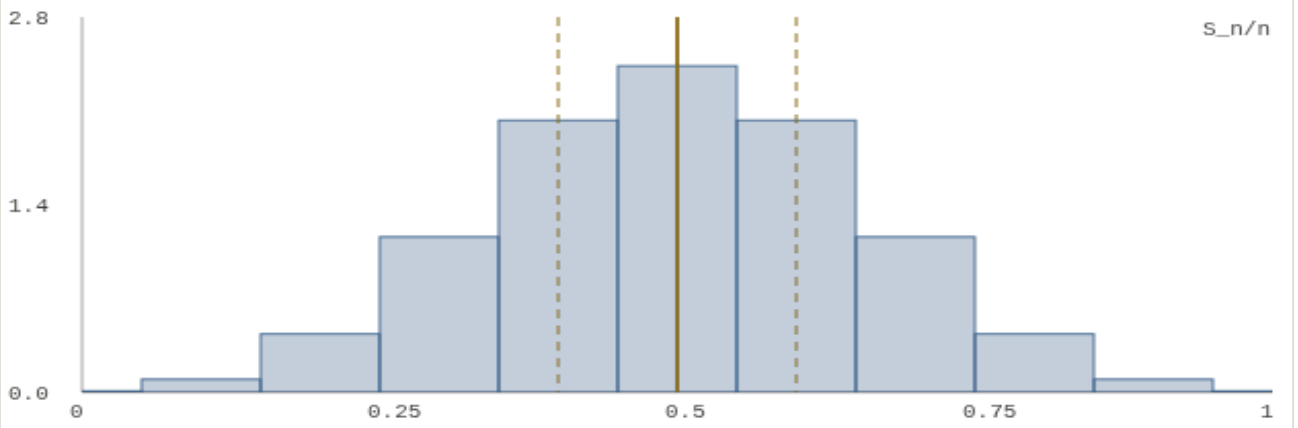
The law of large numbers — the mean settles down

Let S_n be the sum of independent $X_1, X_2, \ldots$ following the same distribution (mean μ, variance σ^2). From § 05, $\mathrm{Var}(S_n/n) = \sigma^2/n$, and putting this into (6) gives the **weak law of large numbers** Known .

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2} \longrightarrow 0 \qquad (n \to \infty) \quad (12)$$

Moreover, with $E|X_1| < \infty$ alone, $S_n/n \to \mu$ holds almost surely (the **strong law**, Durrett Theorem 2.4.1) Known . In "As is" in Figure 4, you can see the distribution of S_n/n shrinking to p and the left-hand side of (12) ($\varepsilon = 0.1$).

Figure 4 The distribution of the number of heads in n tosses of a coin with probability of heads p (bars, area 1). Increasing n, in "As is" it shrinks to p (vertical line) and the probability outside the dashed lines decreases; in "Magnify by $\sqrt{n}$" it approaches the standard Gaussian (curve)



n = 10, p = 0.50, P(|S_n/n − p| ≥ 0.1) = 0.754, right-hand side of (12) 2.50

Computed The numbers of Figure 4 and Table 3 were computed on this machine (the figure computes on the spot from the binomial distribution and an approximation formula for Φ, agreeing with the Python values to 4 digits). If JavaScript is disabled, they can be read from the next paragraph and Table 3.

For $p = 1/2$, $\varepsilon = 0.1$, the left-hand side of (12) is 0.754, 0.0569, 2.7×10^{-10} at $n = 10, 100, 1000$, and the right-hand side $25/n$ is 2.5, 0.25, 0.025 Computed . The right-hand side is a crude upper bound, but enough for the limit. The practical reading is that **the error of a sample mean is about $\sigma/\sqrt{N}$**, so ten times the precision needs a hundred times the samples. Estimating $E[(\mathrm{tr}\, U)^2] = 1$ by sample means gave 0.916 ± 0.092 for $N = 10^2$ and 1.0015 ± 0.0010 for 10^6 (a different sample from Table 1; the error bar is the sample standard deviation $/\sqrt{N}$) Computed . In lattice Monte Carlo, successively generated configurations are not independent, so the effective N decreases (5-07).

09

The central limit theorem — the fluctuations become Gaussian

The deviation $S_n/n - \mu$ is of size $1/\sqrt{n}$. Magnifying it by $\sqrt{n}$ makes the **Gaussian distribution** appear, whatever the original distribution (the **central limit theorem**, Durrett Theorem 3.4.1) Known .

$$P\Big(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le a\Big) \longrightarrow \Phi(a) = \int_{-\infty}^{a} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx \qquad \text{(for all } a\text{)} \quad (13)$$

The reason is visible through characteristic functions. If $\mu = 0$, then $\varphi(s) = 1 - \sigma^2 s^2/2 + o(s^2)$, and by independence $\varphi(t/\sigma\sqrt{n})^n = (1 - t^2/2n + o(1/n))^n \to e^{-t^2/2}$, which is the characteristic function of the one-dimensional standard Gaussian of (8). The continuity theorem, which derives convergence of distributions from convergence of characteristic functions (Durrett Theorem 3.3.17), gives (13) (the classical proof under general conditions is Lindeberg 1922). The rate of approach is given by the **Berry–Esseen inequality** ($\rho = E|X - \mu|^3$).

$$D_n = \sup_a \Big| P\Big(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le a\Big) - \Phi(a) \Big| \le \frac{C_0\, \rho}{\sigma^3 \sqrt{n}} \quad (14)$$

Durrett proves it with $C_0 = 3$, and Shevtsova (2011) has shown $C_0 < 0.4748$ Known . Both sides can be compared in the magnified view of Figure 4 and in Table 3.

Table 3 D_n of the binomial distribution (exact) and the right-hand side of (14) ($C_0 = 0.4748$) Computed

p	n	D_n	Right-hand side	$D_n\sqrt{n}$
0.5	10	0.1230	0.1501	0.389
0.5	1000	0.0126	0.0150	0.399
0.1	100	0.0832	0.1298	0.832
0.02	100	0.1767	0.3258	1.767

For $p = 1/2$, $D_n\sqrt{n}$ settles at $1/\sqrt{2\pi} \approx 0.399$. If n is even, the distribution function has a step of height about $\sqrt{2/\pi n}$ at $a = 0$, and the smooth Φ passes through its middle, so half the step always remains (Durrett § 3.4.4). If n is odd there is no step in the centre, and between the two central steps ($a = \pm 1/\sqrt{n}$) the distribution function stays flat at $1/2$, leaving a deviation of the same size at its ends ($D_n\sqrt{n} = 0.393$ at $n = 11$, 0.399 at $n = 199$) Computed . Either way it is not faster than $1/\sqrt{n}$. For $p = 0.02$, $n = 100$ we have $np = 2$: the distribution is close to an asymmetric Poisson distribution, and still far from Gaussian. Limit theorems do not say "when it is close", so an estimate of the rate is needed.

10

Detour: what does not become Gaussian — heavy tails, triviality

A section you may skip.

(13) assumes that the variance is finite. The **Cauchy distribution** (density $1/\pi(1+x^2)$, characteristic function $e^{-|t|}$) has $E|X| = \infty$, and the mean of n of them follows the same distribution as the original (Durrett Exercise 3.3.6) Known .

In field theory, conversely, "collapsing to a Gaussian" becomes the problem. For the ϕ^4 model on a four-dimensional lattice, it has been shown that the limit of refining the lattice near the critical point is Gaussian and the interaction vanishes (**triviality**) (Aizenman–Duminil-Copin 2021, 7-08) Known . As a metaphor it is the central limit theorem's way of collapsing; to be exact, it is a theorem about strongly correlated fields, and the tools of proof are different.

The "non-trivial" Yang–Mills theory of the problem statement is, roughly, a theory that is not Gaussian in the limit. As the problem statement (§ 6.2) says of scalar fields, the free field is obtained from a Gaussian measure, and 'one requires non-Gaussian dμ to obtain nonlinear fields'. The lattice (1) is not Gaussian to begin with, but there is no result showing that it survives in the limit as a non-Gaussian measure with clustering, and the problem statement itself writes that at present not a single non-trivial relativistic field theory satisfying the axioms in four dimensions is known (as written in 2000) Known . The construction published in 2026-11 made a non-trivial theory in four dimensions, at the level not of a measure but of correlation functions (the committee's assessment is that it is proved in Lean; 6-01). This site does not show the non-triviality of a measure (0-03, 7-10).

What this chapter can say

Said	Not said
Known The definitions, (6), the proof of (12), the derivation of (4). The other theorems are cited from the literature.	Proofs of measure theory (the extension theorem, the convergence theorems, the continuity theorem).
Paper That the moments of the semicircle and the number of ways of joining $2k$ points on a circle by non-crossing chords are both Cat_k (the two computations of the detour of § 06). For $k \le 6$ also checked by counting and by numerical integration Computed .	Why the two agree (the planar diagrams of large N, treated in 17-04).
Computed Tables 1–3, Figures 1–4, the numbers of § 03, § 06, § 08, § 09.	The numbers are checks with finite k, n and samples, and are not a substitute for proofs of the limits.

Said	Not said
Known From the problem statement: that the free field is a Gaussian measure, clustering, that a non-trivial theory needs a non-Gaussian measure. The triviality of four-dimensional ϕ^4.	Measures in infinite dimensions (1-09). The existence of the measure of four-dimensional Yang–Mills (unsolved; the construction published in 2026-11 did not make a measure either; 6-04). This site does not show the non-triviality or clustering of the measure either.

Sources and reproduction

Item	Kind	Source / tools
The theorems and remarks of § 02, § 04, § 08– § 10	Known	R. Durrett, *Probability: Theory and Examples*, 5th ed., Cambridge Univ. Press (2019) (Theorems 2.4.1, 3.4.1, 3.3.17, 3.4.17, the remark of § 3.4.4, Exercise 3.3.6). Text checked in the author's public version.
Original papers (axiomatisation, Vitali, Haar, (9), (13), (14))	Known	A. Kolmogoroff, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer (1933). G. Vitali, Bologna (1905). A. Haar, Ann. of Math. 34 (1933) 147. L. Isserlis, Biometrika 12 (1918) 134–139. G. C. Wick, Phys. Rev. 80 (1950) 268–272. J. W. Lindeberg, Math. Z. 15 (1922) 211–225. A. C. Berry, Trans. AMS 49 (1941) 122–136. C.-G. Esseen, "On the Liapounoff limit of error in the theory of probability", Ark. Mat. Astron. Fys. A 28, No. 9 (1942) 1–19. All bibliographic data only.
$C_0 < 0.4748$, the triviality of § 10	Known	I. Shevtsova, arXiv:1111.6554 (2011). M. Aizenman, H. Duminil-Copin, Ann. of Math. 194 (2021) 163–235, arXiv:1912.07973 (abstract checked). Same authors, Corrigendum, Ann. of Math. 199 (2024) 479 (bibliographic data only).
Quotations from the problem statement (§ 01, § 05, § 07, § 10)	Known	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory" (the problem description of the Clay Mathematics Institute). Included in J. Carlson et al. (eds.), *The Millennium Prize Problems*, Clay Math. Inst. / AMS (2006) 129–152. Text checked in Clay's public version (formula (2) of § 5, § 6.2).
The Catalan numbers of § 06 (detour)	Paper	Computations within this chapter (the recursion by integration by parts and the binomial series). The semicircle of Gaussian matrices is in 17-04 and its table of sources.
Tables, figures, numbers	Computed	Python 3.12 on this machine (NumPy, SciPy): moments.py (Table 1, § 03, § 08; the numerical integration of the semicircle is $k \le 6$), pairings.py (counting for $k \le 6$, (9)), fig_semicircle.py (the counts per bin for Figure 1), lattice_gauss.py (Table 2, Figure 3), clt.py (Table 3, Figure 4), clt_odd.py (odd n in § 09). The functions of the figures were checked against the tables with fig_check.js and fig_check2.js (node).

Next chapter: 1-09 Probability in infinite dimensions — probability on fields themselves, and the Gaussian measure as the starting point.

Probability in infinite dimensions — there is no uniform weight, so we start from Gaussian measures

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That in infinite dimensions there is no weight unchanged by translation (Lebesgue measure), and the starting point becomes the Gaussian measure. That samples fall, with probability 1, outside the set of directions in which one may shift (the Cameron–Martin space). That Brownian motion, white noise and the free field stop even being functions as the dimension grows.

Prerequisite chapters — 1-08. ℓ^2, Fourier coefficients on the torus and distributions are defined on the spot as far as needed (properly in 1-10 and 1-11).

Said first — The big theorems of measure theory are only cited. All the measures made in this chapter are Gaussian, and measures with interaction are touched on only by citation in § 10. Constructing four-dimensional Yang–Mills theory as a measure is unsolved, and this site does not show it either. What the construction published in 2026-11 (6-00) made is correlation functions and a reconstructed field theory, not a measure (6-04).

Order of this chapter

1. Why infinite dimensions — field variables are infinitely many
2. There is no Lebesgue measure — packing balls in orthogonal directions
3. Gaussians in high dimensions live on a shell — reducing the variance
4. Gaussian measures in infinite dimensions — every shadow is Gaussian
5. The Cameron–Martin theorem — directions in which one may shift (Figure 1)
6. Brownian motion — the path settles, the energy diverges (Figure 2)
7. White noise — no value at a point
8. The free field and dimension — the variance at a point diverges (Figure 3)
9. Products of fields — distributions cannot be multiplied
10. Detour: mutually singular measures — scale, mass, interaction

01

Why infinite dimensions — field variables are infinitely many

The lattice probability of 1-08 could be written as "density × volume element" because the variables were finitely many. Refining the lattice and enlarging the box makes the variables infinitely many. Physics books write the probability of a free scalar field of mass m (Euclideanised, 4-02) like this.

$$d\mu(\varphi) \overset{\text{formally}}{=} \frac{1}{Z}\exp\Big(-\frac{1}{2}\int_{\mathbb{R}^d}\big(|\nabla\varphi|^2+m^2\varphi^2\big)\,dx\Big)\prod_{x\in\mathbb{R}^d} d\varphi(x) \quad (1)$$

Neither of the two factors on the right-hand side has a meaning on its own. The measure corresponding to $\prod_x d\varphi(x)$ does not exist (§ 02), and the integral in the exponent is $+\infty$ for almost every sample (§ 05, § 06). Even so, μ as a product can be constructed as a Gaussian measure (§ 04). The problem statement too describes the measures produced by constructive field theory as 'non-Gaussian, Euclidean-invariant measures on spaces of generalized functionals' Known . That means measures on spaces of distributions (§ 07) that are not Gaussian and are unchanged by rotations and translations. This chapter treats the Gaussian case that is their starting point.
Example: dividing the interval into N equal parts with $d=1$, (1) is the Gaussian distribution in N variables of 1-08 § 07, and where $N\to\infty$ leads is the subject of this chapter.

02

There is no Lebesgue measure — packing balls in orthogonal directions

Length and area (Lebesgue measure) are unchanged by translation and give finite weight to bounded sets. In infinite dimensions these two cannot both hold. We state this with the space $\ell^2=\{x=(x_1,x_2,\dots):\sum_k x_k^2<\infty\}$ of sequences with finite sum of squares, the length $\|x\|=(\sum_k x_k^2)^{1/2}$, and the sequence e_n whose n-th component alone is 1.

Theorem Known If a measure on ℓ^2 is unchanged by translation and gives finite weight to every open ball, it is 0.

Proof. If $n\ /=m$ then $\|e_n-e_m\|=\sqrt{2}>1$, so the balls $B(e_n,\frac{1}{2})$ are mutually disjoint, and all lie in $B(0,\frac{3}{2})$.

$$\sum_{n=1}^{\infty}\mu\big(B(e_n,\tfrac{1}{2})\big)\ \le\ \mu\big(B(0,\tfrac{3}{2})\big)\ <\ \infty \quad (2)$$

Each term on the left is the same value c by translation, so $c = 0$. ℓ^2 can be covered by countably many balls of radius $\frac{1}{2}$ centred at sequences with rational components that are 0 except for finitely many, so the total weight is 0 too. ∎ If $\mathbb{R}^n$, only up to 3^n balls fit (the ratio of volumes), and there is no contradiction.

03

Gaussians in high dimensions live on a shell — the variance must decrease

The density of the standard Gaussian $N(0, I_n)$ on $\mathbb{R}^n$ (1-08 (8)) is largest at the origin, but most of the weight lies on a thin shell far away. $|x|^2 = \sum_{k \le n} x_k^2$ is a sum of n independent terms, so by the law of large numbers (1-08 (12)), $|x|^2/n \to 1$.

$$E|x|^2 = n, \qquad \operatorname{Var}|x|^2 = 2n, \qquad |x| = \sqrt{n} + O(1) \quad (3)$$

$E|x|/\sqrt{n} = 0.975,\ 0.9975,\ 0.99975$ at $n = 10,\ 100,\ 1000$; the standard deviation of $|x|$ is $0.698,\ 0.706,\ 0.707$, and the thickness of the shell settles at $1/\sqrt{2}$ regardless of n Computed . Letting $n \to \infty$ with variance 1 gives $|x| \to \infty$, and the limit is not in ℓ^2. So we reduce the variance σ_k^2. For a sequence $\xi_1, \xi_2, \dots$ of independent $N(0,1)$,

$$x = \sum_{k=1}^{\infty} \sigma_k \xi_k e_k, \qquad E\|x\|^2 = \sum_{k=1}^{\infty} \sigma_k^2 \quad (4)$$

the finiteness of the sum on the right is necessary and sufficient for x to be a Gaussian random variable with values in ℓ^2 (Hairer Exercise 4.7) Known . **Example:** if $\sigma_k = 1/k$, then $E\|x\|^2 = \pi^2/6$. A sequence with variance 1 can also be built, on the space $\mathbb{R}^\infty$ of all sequences. What is used is **Kolmogorov's extension theorem** (the theorem that if the distributions of finitely many at a time are given consistently, the distribution of the infinite sequence is fixed uniquely; Durrett Theorem 2.1.21) Known , and this is the seed of the white noise of § 07.

04

Gaussian measures in infinite dimensions — every one-dimensional shadow is Gaussian

Since densities cannot be used, Gaussian measures are defined by their "shadows". A probability measure μ on a space E (ℓ^2 or a space of continuous functions) is a **Gaussian measure** if, for every continuous linear map $\ell : E \to \mathbb{R}$, $\ell(x)$ follows a one-dimensional Gaussian distribution (Hairer Definition 4.4) Known . With mean 0, the covariance $C(\ell, \ell') = E[\ell(x)\ell'(x)]$ determines μ (the form of 1-08 (8)).

$$\int_E e^{\,i\ell(x)}\,d\mu(x) = \exp\Big(-\frac{1}{2}\,C(\ell,\ell)\Big) \quad (5)$$

On ℓ^2 it can be written with $C(\ell,\ell') = \langle \ell, C\ell'\rangle$ and the **covariance operator** C; for (4), $Ce_k = \sigma_k^2 e_k$. If a symmetric positive C has finite trace (sum of eigenvalues), there is a Gaussian measure with it as covariance (Hairer Proposition 4.17) Known . **Example:** $\ell(x) = \sum_k a_k x_k$ is Gaussian with variance $\sum_k a_k^2\sigma_k^2$. The problem statement gives the free field as the Gaussian measure with mean 0 and covariance $(-\Delta + m_0^2)^{-1}$ on the space of distributions $\mathcal{S}'(\mathbb{R}^d)$ Known . The two factors of (1) have been replaced by a single piece of information, the covariance.

05

The Cameron–Martin theorem — the set of directions in which one may shift has probability 0

Two probability measures are **equivalent** if they have the same null sets (sets of weight 0). Conversely, when a set of probability 1 for one has probability 0 for the other, the two are said to be **mutually singular**. A finite-dimensional Gaussian measure acquires a density factor when shifted (in one dimension $N(h,1)$ is $e^{hx-h^2/2}$ times $N(0,1)$). In general a function D with $\mu_h(A) = \int_A D(x)\,d\mu(x)$ for every set A is called the **relative density** and written $d\mu_h/d\mu$. For μ_h, obtained by shifting μ by h in the first n coordinates of (4), we have the following.

$$\frac{d\mu_h}{d\mu}(x) = \exp\Big(\sum_{k\le n}\frac{h_k x_k}{\sigma_k^2} - \frac{1}{2}\|h\|_H^2\Big), \qquad \|h\|_H^2 = \sum_{k\le n}\frac{h_k^2}{\sigma_k^2} \quad (6)$$

$\|h\|_H$ is "the length of the shift measured by the fluctuation in that direction". The degree to which the two can be told apart (the total variation distance) reduces to a one-dimensional problem in the direction of $C^{-1}h$, and is exactly the following formula (Φ is 1-08 (13)).

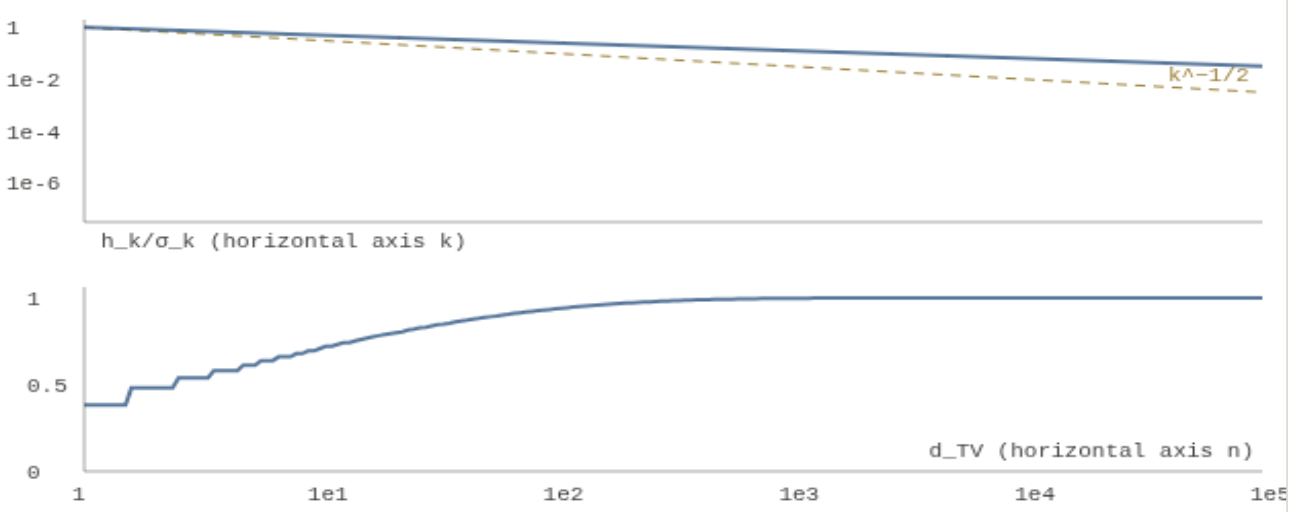
$$d_{\mathrm{TV}}(\mu,\mu_h) = \sup_A\big|\mu(A)-\mu_h(A)\big| = 2\,\Phi\big(\tfrac{1}{2}\|h\|_H\big) - 1 \quad (7)$$

(7) was checked by numerical integration and with samples in 40 dimensions Computed . As $n\to\infty$, if $\|h\|_H$ is bounded then $d_{\mathrm{TV}} < 1$; if it diverges then $d_{\mathrm{TV}}\to 1$, and in the latter case the two measures become mutually singular.

The Cameron–Martin theorem Known μ_h is equivalent to μ only when $\|h\|_H < \infty$, and otherwise they are mutually singular (Hairer Theorem 4.44 and its proof; the original paper is Cameron–Martin 1944). The set H of these h is the **Cameron–Martin space**. Yet for samples of (4), $\|x\|_H^2 = \sum_k \xi_k^2 = +\infty$ (the law of large numbers), and in infinite

dimensions $\mu(H) = 0$ (Hairer Proposition 4.45) Known . Samples fall, with probability 1, outside the set H of directions in which one may shift.

Figure 1 Shifting the measure of $\sigma_k = 1/k$ by $h_k = k^{-a}$. Moving a changes whether the ratio $h_k/\sigma_k = k^{1-a}$ above (log–log) decreases faster than the dashed line $k^{-1/2}$, and whether d_{TV} below ((7) in the first n coordinates) sticks to 1 or stays at the dashed limit



a = 1.30, at n = 10^5 ‖h‖² = 248.0、d_TV = 1.0000, limit ∞、1

Table 1 $\|h\|_H^2 = \sum_{k \le n} k^{2-2a}$ and (7) for $\sigma_k = 1/k$, $h_k = k^{-a}$ Computed

a	$\|h\|_H^2$ $(n = 10^5)$	limit	d_{TV} $(n = 10^5)$	limit
1	10^5	∞	1.0000	1
1.25	631	∞	1.0000	1
1.5	12.09	∞	0.9179	1
1.75	2.606	2.612	0.5804	0.5810
2	1.645	$\pi^2/6$	0.4787	0.4787

Computed The numbers of Figure 1 and Table 1 were computed on this machine (the figure is drawn on the spot from (7) and partial sums, agreeing with the table to 4 digits). If JavaScript is disabled, they can be read from Table 1.

At the borderline $a = 3/2$, $\|h\|_H^2 \approx \ln n + 0.577$ grows only logarithmically, and even at $n = 10^{20}$, $d_{\mathrm{TV}} = 0.9994$ Computed .

06

Brownian motion — the path settles, the energy diverges

The Gaussian measure with values in continuous functions on $[0, 1]$, mean 0 and covariance $E[B(s)B(t)] = \min(s, t)$ is the **Wiener measure**, and its sample B is **Brownian motion** (said to begin with Wiener 1923) Known . The paths are Hölder continuous with any exponent below $1/2$ (if $\gamma < 1/2$, then with a constant C depending on the path, $|B(t) - B(s)| \le C|t - s|^{\gamma}$ for all s, t), and with probability 1 they are

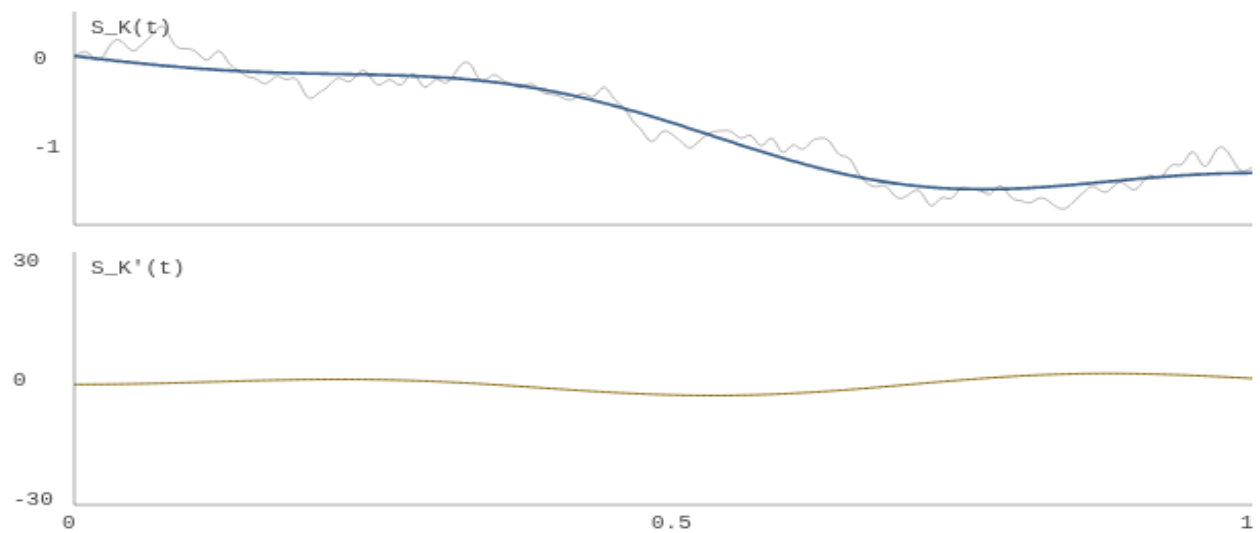
differentiable at no point (Durrett Theorems 7.1.5, 7.1.6) Known . Durrett gives the former as a result of Wiener 1923 and the latter as a discovery of Paley–Wiener–Zygmund 1933.

To put it in the form (4), we look for the eigenfunctions of the covariance operator $(Cf)(s) = \int_0^1 \min(s,t) f(t)\,dt$. $g = Cf$ satisfies $g'' = -f$, $g(0) = 0$, $g'(1) = 0$, so the eigenvalue problem becomes the ordinary differential equation $-g'' = \lambda^{-1} g$, and from $\cos\omega = 0$ the following comes out.

$$B(t) = \sum_{k=1}^{\infty} \xi_k \frac{\sqrt{2}\,\sin(\omega_k t)}{\omega_k}, \qquad \omega_k = \left(k - \frac{1}{2}\right)\pi \tag{8}$$

$\sigma_k = 1/\omega_k \approx 1/(\pi k)$ decreases at the same rate as the example of § 05, and the covariance with 4000 terms agrees with $\min(s,t)$ to within 5.1×10^{-5} Computed . The Cameron–Martin space is the set of functions with $h(0) = 0$, $\|h\|_H^2 = \int_0^1 \dot{h}^2\,dt < \infty$ (Hairer Exercise 4.29) Known , and this $\|h\|_H^2$ is the content of the exponent $m = 0$ in the $d = 1$ of (1). For the partial sum S_K of K terms, $\int_0^1 \dot{S}_K^2\,dt = \sum_{k \le K} \xi_k^2 \approx K$: the path of Figure 2 settles, yet this "energy" keeps growing.

Figure 2 Partial sums of (8). Moving K (the number of terms), S_K above settles (the faint line is $K = 128$), and $\dot{S}_K$ below (an approximation of white noise, § 07) grows in amplitude roughly as $\sqrt{K}$



K = 4 ∫S_K'² = Σξ_k² = 4.62 (÷K 1.154) max|S_K − S_128| = 0.442

Computed The numbers of Figure 2 were computed on this machine (partial sums with 128 fixed random numbers ξ_k). At $K = 1, 16, 64, 128$, $\sum \xi_k^2 = 2.13,\ 13.4,\ 56.5,\ 132.8$, $\max_t |S_K - S_{128}| = 0.589,\ 0.202,\ 0.090,\ 0$. If JavaScript is disabled, it can be read from these numbers and Table 2.

Seen on a lattice — quadratic variation

The same thing is seen when cut into a lattice. The sum of squares of the increments over N equal parts (the **quadratic variation**) approaches 1 (Durrett Exercise 7.1.6)

Known , while for smooth functions it approaches 0, so the discretised energy $\sum(\Delta B)^2/\Delta t$ diverges in proportion to N (Table 2). The $e^{-\frac{1}{2}\int \dot{x}^2 dt}\prod_t dx(t)$ of the path integral (2-07) is a formal way of writing this measure. To be exact, neither the exponent nor $\prod_t dx(t)$ has a meaning on its own (§ 01); as a pair they denote the Wiener measure.

Table 2 One path made with 2^{20} points, viewed divided into N equal parts Computed

N	$\sum(\Delta B)^2$ (fluctuation $\sqrt{2/N}$)	$\sum(\Delta B)^2/\Delta t$	$\sum(\Delta h)^2$ of $h = t^2$
16	0.689 (0.35)	11.0	0.083
4096	0.979 (0.02)	4010	3.3×10^{-4}
1048576	0.9999 (0.001)	1.05×10^{6}	1.3×10^{-6}

07

White noise — no value at a point

The formal derivative $\xi = \dot{B}$ is **white noise**. It has no pointwise values (the lower panel of Figure 2), so it is defined paired with functions. White noise is a Gaussian random variable $\xi(f)$ for each square-integrable function $f \in L^2[0,1]$, linear in f and satisfying the following (Hairer Example 4.63) Known .

$$E\big[\xi(f)\,\xi(g)\big] = \int_0^1 f(t)\,g(t)\,dt \qquad \big(\text{formally, } E[\xi(s)\xi(t)] = \delta(s-t)\big) \tag{9}$$

Example: if $\mathbf{1}_{[0,t]}$ is the function equal to 1 on $[0,t]$ and 0 outside, the covariance of $\xi(\mathbf{1}_{[0,s]})$ and $\xi(\mathbf{1}_{[0,t]})$ is $\int_0^1 \mathbf{1}_{[0,s]}(u)\,\mathbf{1}_{[0,t]}(u)\,du = \min(s,t)$, and this is Brownian motion (Hairer Exercise 4.64). The variance of the mean $\xi(\mathbf{1}_{[t,t+\varepsilon]})/\varepsilon$ over an interval of width ε is $1/\varepsilon$; for a box of side ε in d dimensions it is ε^{-d}, and it diverges when shrunk to a point.

White noise is the Gaussian whose covariance is the identity map (all $\sigma_k = 1$), and as in § 03 it is not in L^2. To measure, dimension by dimension, how far it misses, we now move the stage from $[0,1]$ to the d-dimensional **torus** $\mathbb{T}^d = [0,2\pi)^d$ (a box with opposite ends joined) (the integral of (9) is replaced by the integral over $\mathbb{T}^d$). A function on $\mathbb{T}^d$ is written $f(x) = \sum_{k\in\mathbb{Z}d} c_k\, e^{ik\cdot x}$, and c_k are called the **Fourier coefficients**. $-\Delta = -\sum_i \partial_i^2$ multiplies $e^{ik\cdot x}$ by $|k|^2$, so on coefficients it is the operation turning c_k into $|k|^2 c_k$. The Fourier coefficients of white noise are independent except for the pairs related by $c_{-k} = c_k$, and $E|c_k|^2$ is the same constant for all.

Let H^{-s} be the space of formal series $\sum_k c_k e^{ik\cdot x}$ given the length $\big(\sum_k (1+|k|^2)^{-s}|c_k|^2\big)^{1/2}$ (the larger s, the more lightly high frequencies are counted). For white noise the mean of the square of this length is a constant times $\sum_k (1+|k|^2)^{-s}$, and the shell of

radius R holds about R^{d-1} values of k, so it is finite when $s > d/2$. In fact, exactly then can white noise be realised in H^{-s} (Hairer Exercise 4.65) Known . Things that only give values when paired with functions are **distributions**; the general theory is in 1-11, and stochastic quantisation, which shakes a field with white noise, is in 7-09.

08

The free field and dimension — the variance at a point diverges

The free field of (1), considered on the torus, is white noise smoothed by multiplying its Fourier coefficients c_k by $(|k|^2 + m^2)^{-1/2}$. Writing this operation $(-\Delta + m^2)^{-1/2}$, the covariance is $(-\Delta + m^2)^{-1}$ (the form of the problem statement in § 04), and the variance of the coefficients is proportional to $1/(|k|^2 + m^2)$. By the count of § 07, the condition to lie in H^{-s} becomes $s > d/2 - 1$: one order smoother than white noise, but for $d \geq 2$ it is not a function. The free field on a bounded region can be made as samples in L^2 if $d = 1$, and cannot for $d = 2$ (the remark after Sheffield Proposition 2.7) Known .

On the lattice $a\mathbb{Z}^d$ with lattice spacing a, we replace $-\Delta$ by differences: the $(\Delta x)_j = x_{j+1} - 2x_j + x_{j-1}$ of 1-08 (10) added over coordinate directions and divided by a^2. $e^{ik\cdot x}$ are eigenfunctions of this difference Laplacian, and from $e^{ik_i a} - 2 + e^{-ik_i a} = -4\sin^2 \frac{k_i a}{2}$ the eigenvalues are $-\sum_i \frac{4}{a^2}\sin^2\frac{k_i a}{2}$. Hence the variance at a point is the following integral (k runs over the box that the lattice can distinguish).

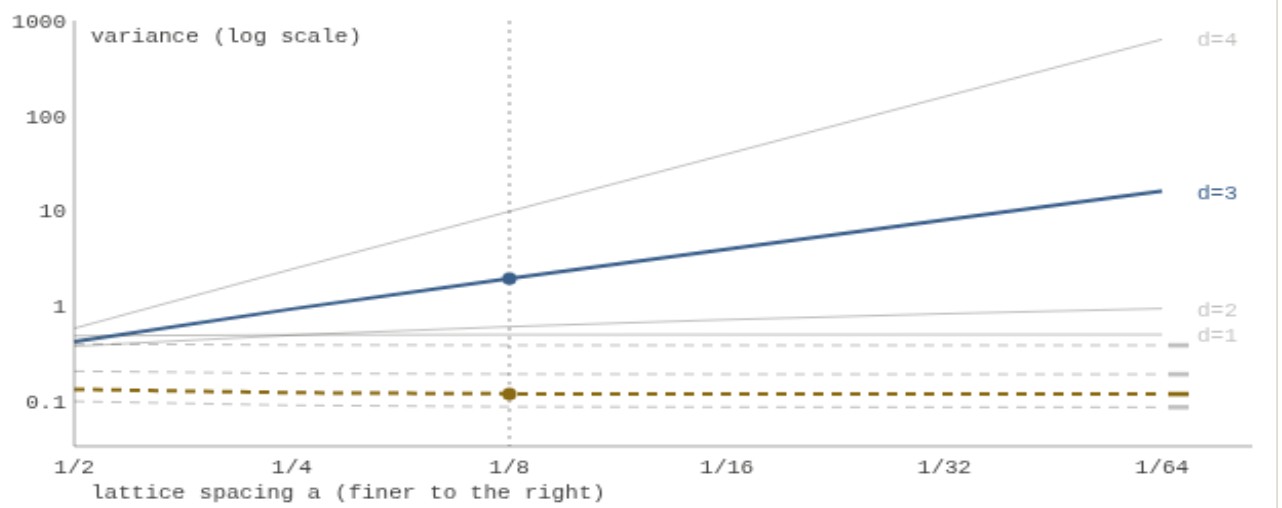
$$G_a(0) = \int_{[-\pi/a,\,\pi/a]^d} \frac{d^d k}{(2\pi)^d}\, \frac{1}{m^2 + \sum_{i=1}^{d} \frac{4}{a^2}\sin^2\frac{k_i a}{2}} \qquad (10)$$

For $a \to 0$ the contribution of large $|k|$ is about $\int^{1/a} k^{d-3}\,dk$, and it diverges as $d \geq 2$. Instead of the value at a point, look at the value $\varphi(\rho_\ell)$ paired with a Gaussian kernel $\rho_\ell(x) = (2\pi\ell^2)^{-d/2} e^{-|x|^2/(2\ell^2)}$ of standard deviation ℓ (formally $\int \varphi(x)\rho_\ell(x)\,dx$, the **smoothed value**); its variance is

$$\operatorname{Var}\varphi(\rho_\ell) = \int_{\mathbb{R}^d} \frac{d^d k}{(2\pi)^d}\, \frac{e^{-\ell^2|k|^2}}{|k|^2 + m^2}$$

(on the lattice, k is restricted to the same box as in (10) and the denominator replaced by that of (10)). The factor $e^{-\ell^2|k|^2}$ suppresses large $|k|$, so it is finite for every d.

Figure 3 Refining the lattice (a from $1/2$ to $1/64$ with the slider), the variance at a point $G_a(0)$ (solid line) keeps growing as $d \geq 2$, while the variance of the smoothed value (dashed) settles at the value marked on the right edge. Pressing a dimension button darkens the two lines for that d (vertical axis logarithmic, $m = 1$, $\ell = 1/4$)



a = 1/8, d = 3, variance at a point G_a(0) = 1.941 (a·G = 0.2426) , variance of the smoothed value 0.1193 (as a→0, 0.1183)

Table 3 The variance at a point $G_a(0)$ of the lattice free field ($m = 1$), and the variance of the value $\varphi(\rho\ell)$ smoothed by a Gaussian kernel of standard deviation $\ell = 1/4$ Computed . This is the authoritative table in this book for the dependence of the variance at a point on d (2-07 § 09 and 2-02 cite it)

a	$d = 1$	$d = 2$	$d = 3$	$d = 4$
1	0.4472	0.2540	0.1705	0.1271
1/2	0.4851	0.3769	0.4223	0.5816
1/4	0.4961	0.4932	0.9289	2.431
1/8	0.4990	0.6057	1.941	9.859
1/16	0.4998	0.7168	3.963	39.60
1/32	0.4999	0.8273	8.007	158.6
1/64	0.5000	0.9377	16.10	634.5
Growth ($a \to 0$)	→ 1/2	$\approx \frac{1}{2\pi}\log\frac{1}{a} + 0.276$	$\approx 0.2527/a$	$\approx 0.1549/a^2$
Smoothed value ($a \to 0$)	0.3852	0.1912	0.1183	0.0861

(10) has been checked against exact values for $d = 1$ and against direct sums on lattices for $d = 2, 3$. $d = 1$ converges to $1/(2m)$; for $d = 2$ the difference from $\frac{1}{2\pi}\ln\frac{1}{a}$ settles at 0.276; for $d = 3$, $a\,G_a(0)$ approaches about 0.253 (0.2515 at $a = 1/64$); and for $d = 4$, $a^2 G_a(0)$ approaches 0.155 Computed . These two destinations are the values at the origin of the lattice Green's functions for $\mathbb{Z}^3$, $\mathbb{Z}^4$, and come out as 0.2527 and 0.1549 from a different one-dimensional integral ($d = 3$ is half of Watson's integral) Computed . The variance of the smoothed value converges to a finite value for every d, and agrees with the limit to the third decimal place at $a = 1/32$. The value of the field is "infinite at a point, finite once smoothed". A figure setting this quantity, whose borderline is 2, beside quantities that switch at four is the dimension slider of 1-12 § 10.

Computed The numbers of Figure 3 and Table 3 were computed on this machine (the values for $a = 2^{-p}$, $p = 1, \ldots, 6$ are embedded in the figure as a table; Table 3 adds to them the row for $a = 1$ and the row of growth). If JavaScript is disabled, they can be read from Table 3.

09

Products of fields — without values at points, multiplication is in trouble

As in § 08, for $d \geq 2$ the mean of $\varphi(x)^2$ is infinite. So terms like φ^4 need treatment by subtraction (renormalisation, 4-04, 7-03), and products of distributions cannot in general be defined (1-11). The four-dimensional Yang–Mills field A is expected to be, at short distances, as rough as the four-dimensional free field (asymptotic freedom, 4-06) Physics . To be exact, the expectation is that correlation functions at short distances approach those of the free field up to logarithmic corrections. The problem statement explains asymptotic freedom as 'at short distances the field displays quantum behavior very similar to its classical behavior' (§ 1), and requires that the correlation functions of the theory to be constructed 'should agree at short distances with the predictions of asymptotic freedom' (§ 4) Known . The $A \wedge A$ of the field strength $F = dA + A \wedge A$ (1-07) is a product of such distributions (7-10).

10

Detour: mutually singular measures — scale, mass, interaction

A section you may skip.

Two Gaussian measures are either equivalent (§ 05) or mutually singular, with nothing in between (Feldman 1958, Hájek 1958; in textbooks it is said to be found in Bogachev Theorem 2.7.2 and elsewhere) Known . In finite dimensions, Gaussian measures with positive definite covariance are always equivalent. In infinite dimensions, as in the following examples, a slight difference makes them mutually singular.

Scale. Multiplying the covariance by c^2 makes $\frac{1}{N}\sum_{k \leq N} x_k^2/\sigma_k^2$ converge not to 1 but to c^2 (the law of large numbers). c can be read from a single sample, so if $c \neq \pm 1$ they are mutually singular (Hairer Proposition 4.43) Known .

Mass. Comparing free fields of masses m_1, m_2 (both positive) on the torus, the ratio of variances of the Fourier coefficients is $1 + \varepsilon_k$, $\varepsilon_k = (m_1^2 - m_2^2)/(|k|^2 + m_2^2)$. To compare two product measures of independent coordinates we use Kakutani's criterion. Let p_k, q_k be the two densities of the k-th coordinate (here both positive Gaussian densities). If the infinite product of the overlaps $\int \sqrt{p_k(x) q_k(x)}\, dx$ for each dimension is positive the two are equivalent, and if 0 they are mutually singular (Kakutani 1948) Known . For a real-valued field the real and imaginary parts of the Fourier coefficients are independent coordinates, but the ratio of variances is $1 + \varepsilon_k$ for both, so the condition below does not change. The overlap of two Gaussians with ratio of variances $r^2 = 1 + \varepsilon$ is $\sqrt{2r/(1+r^2)} = 1 - \varepsilon^2/16 + O(\varepsilon^3)$, and $\varepsilon_k \to 0$, so the product being positive is equivalent to $\sum_k \varepsilon_k^2 < \infty$.

$$\sum_{k \in \mathbb{Z}^d} \varepsilon_k^2 = (m_1^2 - m_2^2)^2 \sum_{k \in \mathbb{Z}^d} \frac{1}{(|k|^2 + m_2^2)^2} < \infty \quad \Longleftrightarrow \quad d \leq 3 \tag{11}$$

Looking at the partial sums $\sum(|k|^2+1)^{-2}$ of $|k| \le R$ for $m_2 = 1$ with $R = 16, 32, 64, 128$: for $d = 3$ they are 9.21, 9.61, 9.80, 9.90, the increments for each doubling of R halve as 0.39, 0.20, 0.10, and they converge near 10.0 (the remainder is about $4\pi/R$; a separate integral gives 9.998 for the whole sum). For $d = 4$ the increments do not halve, and grow by $2\pi^2 \ln 2 \approx 13.7$ for each doubling of R Computed . Hence, if $m_1 \neq m_2$, the measures of two free fields of different masses on the torus are equivalent if $d \le 3$ and mutually singular if $d \ge 4$ Known (Schröfl–Floerchinger 2025 § 4.1; the cases of Dirichlet, Neumann and periodic boundaries on bounded regions). (11) is the same condition as the Feldman–Hájek criterion (with C_i the covariance of mass m_i, $C_1^{-1/2} C_2 C_1^{-1/2} - I$ is a Hilbert–Schmidt operator; here, the sum of squares of the eigenvalues ε_k is finite).

Interaction. The construction that multiplies a Gaussian measure by a density e^{-V} breaks down when V cannot be defined for almost every sample. Measures made as limits in general have no density (Sheffield § 3.3), and the φ^4 measure on the three-dimensional torus is mutually singular with the free field (Barashkov–Gubinelli 2020) Known . Constructive field theory makes measures as limits of cut-off measures (7-01, 7-05).

What this chapter can say

Said	Not said
Known The theorem of § 02 (proved in the text). The other theorems are cited from the literature.	Proofs of the theorems of Kolmogorov, Cameron–Martin, Kakutani and Feldman–Hájek; the general theory of constructing measures on spaces of distributions.
Known The criterion by (11) for free fields of different masses on the torus ($m_1 \neq m_2$) (the same conclusion as Schröfl–Floerchinger 2025 § 4.1).	No machine checking. The whole of $\mathbb{R}^d$ is not treated.
Computed Tables 1–3, Figures 1–3, the numbers and checks in the text.	The limiting values (0.253, 0.155, 10.0 and so on) are readings from finite a, R and numerical integration, without rigorous error estimates.
Physics The expectation that correlation functions of the four-dimensional Yang–Mills field approach those of the free field at short distances (up to logarithmic corrections).	Not a theorem. The existence of interacting measures in four dimensions (including Yang–Mills) is unsolved, and this site does not show it either (the construction published in 2026-11 did not make a measure either; 6-04).

Sources and reproduction

Item	Kind	Source / tools
Definitions and theorems of § 03– § 10 (numbers as in the text)	Known	M. Hairer, "An Introduction to Stochastic PDEs", arXiv:0907.4178 (v2, 2023) § 4. Text checked.
Theorem 2.1.21 (Kolmogorov's extension theorem), Theorems 7.1.5, 7.1.6 and the surrounding attributions (Wiener 1923, Paley–Wiener–Zygmund 1933), Exercise 7.1.6	Known	R. Durrett, *Probability: Theory and Examples*, 5th ed., Cambridge Univ. Press (2019). Text checked in the author's public version.
Proposition 2.7 and the remark after it, § 3.3	Known	S. Sheffield, "Gaussian free fields for mathematicians", Probab. Theory Relat. Fields 139 (2007) 521–541, arXiv:math/0312099. Text checked.
Original papers (Wiener measure, non-differentiability, Cameron–Martin, Kakutani, the dichotomy of Gaussian measures)	Known	N. Wiener, J. Math. Phys. 2 (1923) 131–174. R. E. A. C. Paley, N. Wiener, A. Zygmund, Math. Z. 37 (1933) 647–668 (bibliographic data from Durrett's bibliography). R. H. Cameron, W. T. Martin, Ann. of Math. 45 (1944) 386. S. Kakutani, Ann. of Math. 49 (1948) 214–224. J. Feldman, Pacific J. Math. 8 (1958) 699–708 (text checked: the alternative of equivalence or orthogonality for measures of Gaussian processes). J. Hájek, Czechoslovak Math. J. 8 (1958) 610–618 (in Russian; bibliographic data checked in DML-CZ, text bibliographic data only). The others bibliographic data only (Kakutani's statement checked in secondary sources).
The dichotomy of Gaussian measures (textbooks)	Known	V. I. Bogachev, *Gaussian Measures*, AMS (1998) Theorem 2.7.2. G. Da Prato, J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge Univ. Press (2014) Theorem 2.25. All bibliographic data only (mentions checked in secondary sources).
The criterion for free fields of different masses in § 10 (bounded region; equivalent for $d < 4$, singular for $d \geq 4$)	Known	M. Schröfl, S. Floerchinger, "Relative entropy and mutual information in Gaussian statistical field theory", Ann. Henri Poincaré 26 (2025) 3233–3319, arXiv:2307.15548. Text checked (§ 4.1, Theorem 3; periodic boundaries are the cube case of the same section).
The values at the origin of the lattice Green's functions for $\mathbb{Z}^3$ (§ 08)	Known, Computed	G. N. Watson, "Three triple integrals", Quart. J. Math. Oxford 10 (1939) 266–276. Bibliographic data only. The value 0.2527 was computed on this machine from a different one-dimensional integral.
The singularity of φ^4 on the three-dimensional torus	Known	N. Barashkov, M. Gubinelli, "The Φ^4_3 measure via Girsanov's theorem", arXiv:2004.01513 (2020). Abstract checked.

Item	Kind	Source / tools
Quotations from the problem statement in §01, §04, §09	Known	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory" (the problem description of the Clay Mathematics Institute), §1, §2, §4, §6. Text checked.
Tables, figures, numbers in the text	Computed	Python 3.12 on this machine (NumPy 1.26, SciPy 1.11): dims.py, cameron_martin.py, brownian.py, gff.py, watson.py, fig3.py. The functions of the figures were checked against the tables with fig_check.js (node).

Next chapter: 1-10 Hilbert spaces and self-adjoint operators — putting the ℓ^2, covariance operators and eigenfunction expansions of this chapter in order in the language of spectra.

Hilbert spaces and self-adjoint operators — measuring "the lowest" and "the next" in infinite dimensions

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — Spaces in which functions are treated as vectors whose length can be measured, and unbounded operators such as differentiation. That the spectrum is wider than the set of eigenvalues, and that boundary conditions are part of the operator. That the variational principle measuring the lowest value and the gap is the same thing as the best constant of the Poincaré inequality.

Prerequisite chapters — 1-01 (inner products, Hermitian matrices, diagonalisation). Sections 02, 05, 08 and 09 touch on 1-08 and 1-09 (sets of length 0, Brownian motion, Gaussian distributions).

Said first — We introduce the entrance to functional analysis, and omit the details of domains and the proofs of theorems. The Hamiltonians of quantum field theory are much harder to handle (4-09), and this chapter shows nothing about the Yang–Mills mass gap.

Order of this chapter

1. Why infinite dimensions — the lower eigenvalues stay, the upper ones run away (Figure 1, Table 1)
2. Hilbert spaces — limits stay inside the space
3. Operators and self-adjointness — boundary conditions are part of the operator
4. The spectrum — some values have no eigenvalue
5. The spectral theorem — the gap decides the rate of decay
6. The variational principle — upper bounds are easy, lower bounds need other information (Table 2, box "The asymmetry of upper and lower bounds")
7. The Poincaré inequality — the best constant is the inverse of the gap

8. The weighted Poincaré inequality — $1/\kappa$ if convex, large if there are two valleys (Figure 2, Table 3)
9. With many variables — products, Gaussians, the free field, and two gaps
10. Detour: counting eigenvalues — Weyl's law and the shape of a drum

---

01

Why infinite dimensions — the lower eigenvalues stay, the upper ones run away

In finite dimensions the smallest eigenvalue and the gap to the next can always be computed (1-01). What is asked is whether the gap stays positive in the limit of more and more degrees of freedom. As an example, discretise on a lattice of N equal parts the rule assigning $-u''$ to a function u on $[0,1]$ vanishing at both ends (the same form as the Hessian matrix of 2-01 § 05).

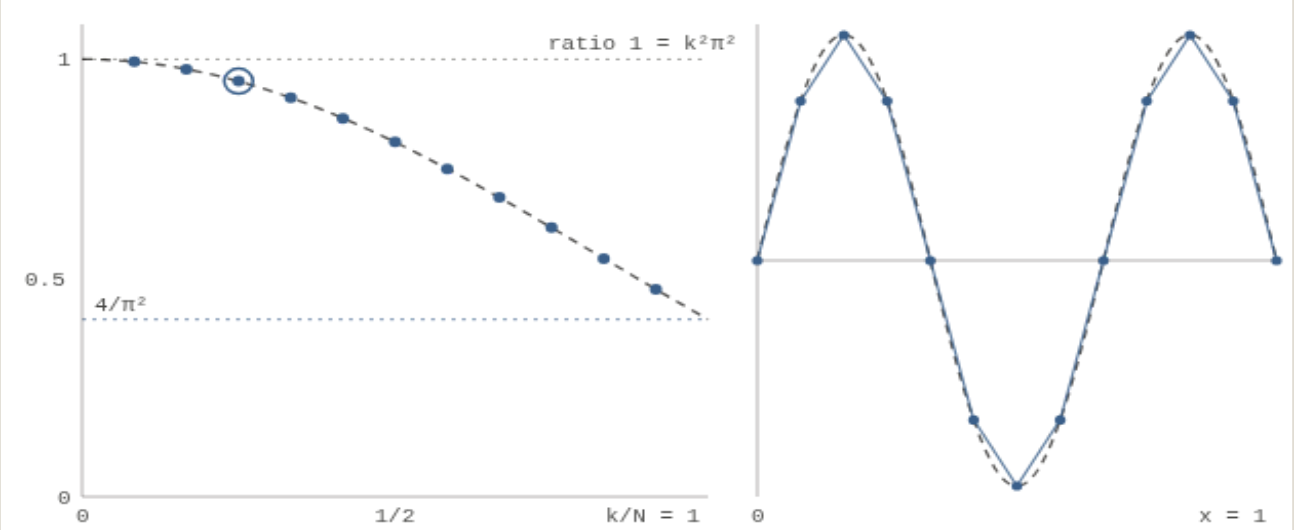
$$(A_N u)_j = N^2\,(-u_{j-1} + 2u_j - u_{j+1}), \quad u_0 = u_N = 0; \qquad \lambda_k(N) = 4N^2 \sin^2 \frac{k\pi}{2N} \quad (1 \le k$$

The eigenvectors are $u_j = \sin(k\pi j/N)$ (this can be checked by substitution). Fixing k and letting $N \to \infty$ gives $\lambda_k(N) \to k^2\pi^2$. This is an eigenvalue of $-u'' = \lambda u,\ u(0) = u(1) = 0$. The top eigenvalue grows to near $4N^2$ and has no limit.

Computed **Table 1** Eigenvalues of the matrix of (1) (the difference between a tridiagonal eigenvalue computation and (1) is at most 10^{-15} relative).

N	λ_1	λ_2	top
10	9.78870	38.19660	390.2
40	9.86453	39.39731	6390.1
160	9.86929	39.47334	102390.1
continuum	π^2 = 9.86960	$4\pi^2$ = 39.47842	(none)

Figure 1 Left: the ratios of eigenvalues of (1) λ_k(N)/k²π² (dots, k = 1 ... N − 1) plotted against t = k/N. The dots lie on the curve (sin(πt/2)/(πt/2))² (dashed) regardless of N, and increasing N brings the dot for number k (circle) towards the ratio 1. The ratio 4/π² at the right end means that the top end levels off at 4N². Right: the k-th eigenvector (dots) and sin kπx (dashed)



Computed The numbers in the figure are computed on the spot from (1) (checked against Table 1).

The ratio is fixed by $t = k/N$ alone: $\lambda_k(N)/k^2\pi^2 = \big(\sin(\pi t/2)/(\pi t/2)\big)^2$ (the dashed line on the left of Figure 1). At $k = N/2$, $8/\pi^2 \approx 0.81$; at the right end $t \to 1$, $4/\pi^2 \approx 0.41$. Fixing k and increasing N gives $t \to 0$, and the ratio approaches 1. An eigenvector with k near N is, at the lattice points, exactly the values of $\sin k\pi x$, but its oscillation is as fine as the spacing of the points, so joining the points makes it look like a different shape. In fact $\sin(k\pi j/N) = -(-1)^j \sin((N-k)\pi j/N)$: a slow oscillation multiplied by a sign that alternates point by point (visible on the right of Figure 1 by bringing k near $N - 1$). What remains in the limit is the lower eigenvalues and the gap, and the limiting object is an operator unbounded above. In the continuum limit of lattice gauge theory (5-05) too, what is read from the lattice is the lower end of the spectrum (the lightest state) Physics . In four dimensions, that the scaling limit of lattice gauge theory exists was shown by the construction published in 2026-11 (the committee's assessment is that it is proved in Lean; 6-00). The lower end of the spectrum of that limiting theory, that is, the mass gap, has not been shown (6-04).

02

Hilbert spaces — limits stay inside

To treat functions as "vectors with infinitely many components" we need an inner product (1-01 § 05) and **completeness**. Complete means that sequences whose terms come arbitrarily close to one another (Cauchy sequences) have limits inside the space. A complex vector space with these two is called a **Hilbert space**.

$$\ell^2 = \Big\{(c_1, c_2, \dots) : \sum_k |c_k|^2 < \infty\Big\}, \qquad L^2(0,1) = \Big\{f : \int_0^1 |f|^2 dx < \infty\Big\}, \qquad \langle f, g\rangle = \int_0^1$$

In L^2, functions that differ only on a set of length 0 are regarded as the same (1-08). Continuous functions alone are not complete: $\tanh(n(x-\frac{1}{2}))$ come close to one another, yet their destination is a step function. To take the destination of "a sequence making the ratio smaller and smaller" in section 06, the complete L^2 is needed Known .

$e_k = \sqrt{2}\sin k\pi x$ $(k \geq 1)$ is an orthonormal basis of $L^2(0,1)$, and for $c_k = \langle e_k, f\rangle$, $\|f\|^2 = \sum_k |c_k|^2$ (Parseval's identity) Known (1-11). Example. For $f = x(1-x)$, $c_k = 4\sqrt{2}/(k\pi)^3$ for odd k and 0 for even. $\sum_{k\ \text{odd}} 32/(k\pi)^6 = 1/30$ agrees with $\int_0^1 x^2(1-x)^2 dx = 1/30$. The ℓ^2 of 1-09 connects to spaces of functions in this way.

03

Operators and self-adjointness — domains and boundary conditions

An operator whose supremum of stretching factors $\|A\| = \sup_{\psi\neq 0}\|A\psi\|/\|\psi\|$ is finite is called bounded. $-d^2/dx^2$ multiplies e_k by $k^2\pi^2$, so it is not bounded, and there are functions whose second derivatives are not in L^2. So the subspace of functions it may be applied to, the **domain** $D(A)$, is specified as part of the operator. If $\langle \phi, A\psi\rangle = \langle A\phi, \psi\rangle$ on $D(A)$, it is **symmetric**. Self-adjointness is one step stronger: it says that the partners ϕ of this identity do not increase even when sought outside $D(A)$. To be exact, the set of ϕ for which there is a η with $\langle \phi, A\psi\rangle = \langle \eta, \psi\rangle$ for all $\psi \in D(A)$ is taken as the domain $D(A^\dagger)$ of the adjoint $A^\dagger$, and $A^\dagger\phi = \eta$ is defined. If symmetric, $D(A) \subset D(A^\dagger)$, and when $D(A^\dagger) = D(A)$ it is **self-adjoint**. The spectral theorem and time evolution e^{-itA} (2-05) presuppose self-adjointness Known . For the momentum $P = -i\,d/dx$, integration by parts gives the following.

$$\langle \phi, P\psi\rangle - \langle P\phi, \psi\rangle = -i\,\big[\,\phi\,\psi\,\big]_0^1 \quad (3)$$

(a) If $D = \{\psi(0) = \psi(1) = 0\}$, it is symmetric, but the partner ϕ needs no condition, and $P^\dagger e^{izx} = ze^{izx}$ for every complex number z, so it is not self-adjoint. (b) If $D_\theta = \{\psi(1) = e^{i\theta}\psi(0)\}$, the partner satisfies the same condition and it is self-adjoint, with eigenvalues $2\pi n + \theta$ $(n \in \mathbb{Z})$. The phase difference $q\Phi/\hbar$ from a magnetic flux Φ threading the ring plays the role of θ (the Aharonov–Bohm effect of 2-03). For $-d^2/dx^2$ too, with $u = 0$ at the ends (Dirichlet) the spectrum is $\{k^2\pi^2\}_{k\geq 1}$, and with $u' = 0$ (Neumann) it is $\{k^2\pi^2\}_{k\geq 0}$. **Boundary conditions are part of the operator**, and change the lowest value.

04

The spectrum — values for which no inverse can be made

In infinite dimensions there are cases where a map is one-to-one but its inverse is not bounded. So the **spectrum** is defined as follows.

$$\sigma(A) = \{\lambda \in \mathbb{C} : \ (A-\lambda)^{-1} \text{ does not exist as a bounded operator}\}, \qquad \|(A-\lambda)^{-1}\| = \frac{1}{\text{dis}}$$

The identity on the right holds when A is self-adjoint and $\lambda \notin \sigma(A)$, and then $\sigma(A)$ is a closed subset of the real line Known . Example. The operator M multiplying by x on $L^2(0,1)$ has no eigenvalues, since $xf = \lambda f$ implies $f = 0$. Even so, for $\lambda \in [0,1]$, $1/(x-\lambda)$ is not bounded, and $\sigma(M) = [0,1]$. A function of length 1 that is constant only on $[\lambda, \lambda + \varepsilon]$ has $\|(M-\lambda)f\| \le \varepsilon$: there are things arbitrarily close to eigenvectors (Weyl's criterion of 2-05 § 03). Discretising makes the eigenvalues lattice points, which fill up $[0,1]$ (the lower eigenvalues of section 01 stay apart).

Isolated eigenvalues of finite multiplicity are called the **discrete spectrum**, and the rest the **essential spectrum**. The mass gap is the claim that the vacuum eigenvalue 0 is isolated and the rest of the spectrum lies at $\Delta > 0$ or above (4-11).

05

The spectral theorem — the gap decides the rate of decay

A self-adjoint operator can be carried by a suitable unitary to "an operator multiplying by a real-valued function", like M (the spectral theorem) Known . It is the infinite-dimensional version of the diagonalisation of matrices (1-01), and to be exact it includes parts that cannot be written with eigenvalues alone. Functions g can be substituted, and if there is a basis of eigenvectors it takes the following form.

$$g(A)\,\psi = \sum_{k} g(\lambda_k)\,\langle e_k, \psi\rangle\, e_k, \qquad \left\| e^{-tA}\psi - \langle e_0,\psi\rangle e_0 \right\| \le e^{-\Delta t}\left\| \psi - \langle e_0,\psi\rangle e_0 \right\| \quad (5)$$

The right-hand side is for the case $A \ge 0$, with the eigenspace of 0 spanned by e_0 of length 1 and the rest of the spectrum in $[\Delta, \infty)$; it follows by using $e^{-t\lambda} \le e^{-\Delta t}$ on each component. **The gap is the slowest rate at which the components other than the lowest disappear** (for the reading in imaginary time see 2-05 § 05). Example 1. For the heat equation $\partial_t u = \partial_x^2 u$ on a rod insulated at both ends, A is the Neumann $-d^2/dx^2$, $\Delta = \pi^2$, and the deviation of the temperature from its mean is at most $e^{-\pi^2/2} \approx 0.0072$ times at $t = 0.5$. Example 2. The inverse of the Dirichlet $-d^2/dx^2$ is $(Gf)(x) = \int_0^1 G(x,y) f(y)\,dy$, $G(x,y) = \min(x,y) - xy$ ($-(Gf)'' = f$ and the zeros at both ends can be checked). Making $X(t) = B(t) - tB(1)$ from the Brownian motion B of 1-09 gives zeros

at both ends (the **Brownian bridge**), and its covariance $E[X(s)X(t)]$ is exactly $G(s,t)$. The eigenvalues of G are $(k\pi)^{-2}$, and the finiteness of the sum $\sum_k (k\pi)^{-2} = 1/6$ is the condition of 1-09 (4) for samples of the Brownian bridge to lie in L^2.

06

The variational principle — easy from above, hard from below

The lower end of the spectrum of a self-adjoint operator bounded below is the infimum of the **Rayleigh quotient**. The n-th too can be written in the form "take the maximum over a n-dimensional subspace, and the minimum over subspaces" (the min-max principle; for the eigenvalues below the essential spectrum, in increasing order counted with multiplicity) Known .

$$\inf \sigma(A) = \inf_{0 \neq \psi \in D(A)} \frac{\langle \psi, A\psi \rangle}{\langle \psi, \psi \rangle}, \qquad \lambda_n = \min_{\dim V = n} \ \max_{0 \neq \psi \in V} \frac{\langle \psi, A\psi \rangle}{\langle \psi, \psi \rangle} \tag{6}$$

Expanded in eigenvectors, the Rayleigh quotient is a weighted average of eigenvalues $\sum \lambda_k |c_k|^2 / \sum |c_k|^2$, and does not fall below λ_1. The eigenvalues of the matrix restricted to a subspace V are **upper bounds** for $\lambda_1, \lambda_2, \ldots$, and decrease as V is enlarged (the Rayleigh–Ritz method). Example. Putting $x(1-x)$ into the Dirichlet $-d^2/dx^2$ gives the quotient $(1/3)/(1/30) = 10 \geq \pi^2$. For estimates from below there is Temple's inequality. If R is the Rayleigh quotient of $\psi \in D(A)$, and we know a lower bound μ for the next eigenvalue ($R < \mu \leq \lambda_2$), the following holds Known .

$$\lambda_1 \ \geq \ R - \frac{\|A\psi\|^2/\|\psi\|^2 - R^2}{\mu - R} \tag{7}$$

If $\psi = x(1-x)$, then $A\psi = 2$, $\|A\psi\|^2/\|\psi\|^2 = 120$. Here $\lambda_2 = 4\pi^2$ is known exactly (section 03), so we can put $\mu = 39$, giving $\lambda_1 \geq 10 - 20/29 = 9.310$. The Ritz value for $39.50\ldots$ appearing below Table 2 is an upper bound for λ_2, and is no basis for μ. **The variational principle alone gives only upper bounds; holding it down from below needs other information.** In Temple's inequality it is the lower bound μ for the next level (the convexity $V'' \geq \kappa$ of section 08 is another example). This is, with finitely many degrees of freedom, the shape of why the proof of the mass gap is hard (2-05 § 04, 4-11).

Computed **Table 2** Rayleigh–Ritz values on $V_n = \{x(1-x)p(x) : \ p \text{ of degree at most } n-1 \}$ and the lower bound from (7) ($\mu = 39$). The matrices were built exactly with fractions.

n	Upper bound for λ_1	Lower bound for λ_1 (7)	Upper bound for λ_3
1	10.000000000	9.310344828	—
3	9.869749621	9.866602308	102.1303
5	9.869604435	9.869602524	89.1736
exact	9.869604401		88.82644

The antisymmetric functions added for even n improve only λ_2, and the upper bound comes down to $42,\ 39.50155,\ 39.47847$ ($4\pi^2 = 39.47842$). At $n = 5$, π^2 is caught in an interval of width 2×10^{-6} (the lower end depends on $\mu = 39 \le 4\pi^2$).

The asymmetry of upper and lower bounds — one from above, all from below

To hold a gap or the lowest eigenvalue **from above**, it is enough to exhibit **one** trial vector. Every value of the quotient (6) is an upper bound as it stands. To hold it **from below**, one must show that the quotient does not fall below it for **every** vector, and information outside the variational principle is needed (the μ of Temple's inequality, the convexity of section 08, the decomposition of the whole by the spectral theorem) Known . The Poincaré constant C_P (section 07) is the inverse of the gap, so the directions swap: one trial function gives a lower bound for C_P, and an upper bound needs every function.

The mass gap $\Delta > 0$ is a **lower bound** on the energy of states orthogonal to the vacuum (4-11 § 01). This book puts it in one sentence: **a proof of the mass gap is a proof of a lower bound.** Trial states and measurements of finitely many correlations give upper bounds for Δ, and say nothing about the gap being positive (4-11 § 02, 18-01 § 08). As a metaphor: looking for the lowest valley in the mountains, finding one valley lets you say "at most this low", but "there is no valley lower than this" needs the whole range to be examined. To be exact: evaluating the infimum of (6) with one ψ gives an upper bound, and holding the infimum itself from below needs every ψ.

07

The Poincaré inequality — the best constant is the inverse of the gap

The Poincaré inequality is "a function with small slope does not stray far from its mean" put in numbers. On the interval $[0, 1]$ it takes the following form, and the constant $1/\pi^2$ cannot be made smaller Known .

$$\int_0^1 |f - \bar{f}|^2 dx \ \le\ \frac{1}{\pi^2} \int_0^1 |f'|^2 dx, \qquad \bar{f} = \int_0^1 f\, dx \quad (8)$$

Proof. Writing $f-\bar f=\sum_{k\ge 1}a_k\sqrt2\cos k\pi x$ with the Neumann eigenfunctions $1,\ \sqrt2\cos k\pi x$, the left-hand side is $\sum|a_k|^2$ and the integral on the right is $\sum k^2\pi^2|a_k|^2\ge\pi^2\sum|a_k|^2$. Equality holds only when $f=a+b\cos\pi x$. On an interval of length L the constant is $(L/\pi)^2$, **proportional to the square of the size** (the version on a circle is (4) of 17-06).

This is the variational principle of section 06 used for "the next after 0" of the Neumann operator. In general, if for a self-adjoint $A\ge 0$ only constants satisfy $Af=0$ and the gap is Δ, then the best constant of $\operatorname{Var}(f)\le C_P\langle f,Af\rangle$ is $C_P=1/\Delta$, and by (5) the smoothing motion settles at rate $1/C_P$. The Dirichlet version $\int|f|^2\le\pi^{-2}\int|f'|^2$ ($f(0)=f(1)=0$) has the same number, but equality is at $\sin\pi x$, and the content is different. Example. For $f=x$ the left-hand side is $1/12$ and the right-hand side $1/\pi^2$, with ratio 0.82.

08

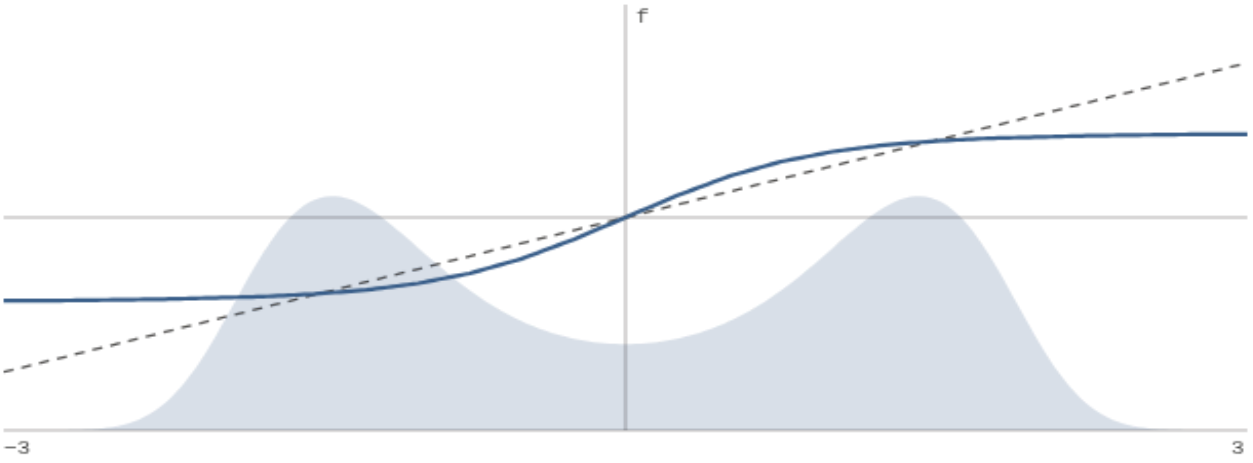
The weighted Poincaré inequality — 1/κ if convex, large if there are two valleys

What field theory needs is the form for a probability measure $\mu(dx)=e^{-V(x)}dx/Z$ (Z is the constant making the total 1). Write the inner product as $\langle f,g\rangle_\mu=\int f\,g\,d\mu$, the space of functions of finite length in this inner product as $L^2(\mu)$, and the variance as $\operatorname{Var}_\mu(f)=\int\bigl|f-\int f\,d\mu\bigr|^2d\mu$.

$$\operatorname{Var}_\mu(f)\le C_P\int|f'|^2\,d\mu;\qquad \operatorname{Var}_\mu(x)\ \le\ C_P\ \le\ \frac{1}{\kappa}\quad(V''\ge\kappa>0)\ \text{(9)}$$

The best C_P is the inverse of the gap of $Lf=-f''+V'f'$ after 0. L is self-adjoint on $L^2(\mu)$, integration by parts gives $\langle f,Lf\rangle_\mu=\int|f'|^2d\mu$, and only constants satisfy $Lf=0$ (the same form as section 07; from the side of the ground state, 2-05 § 08). The inequality on the left follows by putting in $f=x$. The one on the right is said to be a corollary of the variance inequality of Brascamp–Lieb (1976), and also follows from the logarithmic Sobolev inequality of Bakry–Émery (1985) (an inequality stronger than the Poincaré inequality) Known . In one dimension the following line suffices. Differentiating $Lf=\lambda f$, $g=f'$ gives $Lg+V''g=\lambda g$; the inner product with g gives $\int|g'|^2d\mu+\int V''g^2d\mu=\lambda\int g^2d\mu$, hence $\lambda\ge\kappa$ (assuming there is an eigenfunction and the boundary terms vanish). For the normal distribution $V=x^2/(2\sigma^2)$ all three are σ^2 (Chernoff 1981).

Figure 2 Moving c (from −2 to 4) splits the weight exp(−V_c) (shaded) into two peaks, and the f giving equality (solid line) changes from the straight line x/√Var(x) (dashed) to a step



c = 2.0, barrier c²/4 = 1.00, Var(x) = 1.6655 ≤ C_P = 2.6714 (there is no κ > 0, so there is no upper bound on the right)

Computed The numbers in the figure were computed on this machine (the best f and C_P were found with python3 and embedded; the weight e^{-V_c} is computed on the spot).

Computed **Table 3** C_P for $V_c = x^4/4 - cx^2/2$ (found by turning L into an operator of Schrödinger type and using three-point differences; the method is in the table of sources; halving the step agrees to 4 digits). At $c < 0$, $\kappa = -c$.

c	1/κ	Var(x)	C_P
−2	0.500	0.3451	0.3503
0	—	0.6760	0.7307
1	—	1.0418	1.2625
2	—	1.6655	2.6714
3	—	2.5853	7.6577
4	—	3.6707	33.580

With two valleys ($c > 0$, height of the hill between the valleys $c^2/4$), C_P grows much faster than $\mathrm{Var}(x)$, and at $c = 4$ it is 9 times as large. The best f becomes a step close to -1 in the left valley and $+1$ in the right, and its slope matters only on the hill, where there is almost no weight. **A narrowing of the weight makes C_P large.**

For how this book uses this, see 23-14 § 02 (the distance to the nearest known theorem, condition by condition).

09

With many variables — products, Gaussians, the free field, and two gaps

When there are many variables $x = (x_1, \dots, x_n)$, replace the $|f'|^2$ of (9) by the squared length of the gradient $|\nabla f|^2 = \sum_j |\partial_j f|^2$. The basics are the following two Known .

(i) The CP of a product measure of independent variables is the maximum of the CP of the factors, and does not depend on the number of variables. L becomes the sum of the L for each variable, the eigenfunctions are products $f_1(x_1)\cdots f_n(x_n)$ and the eigenvalues sums, so the value after 0 is the minimum of the gaps of the factors (assuming a basis of eigenfunctions).

(ii) The CP of the Gaussian measure with mean 0 and covariance matrix C is the largest eigenvalue of C. Taking the eigenvectors of C as orthogonal coordinates, the measure splits into a product of independent one-dimensional Gaussians whose variances are the eigenvalues of C (1-08), and $|\nabla f|^2$ does not change under a change of orthogonal coordinates. The rest uses (i) and the one-dimensional $CP = \sigma^2$ (section 08).

For the free field on the ring of 1-08 (10), $C = (-\Delta_{\text{lat}} + m^2)^{-1}$. The eigenvalues of $-\Delta_{\text{lat}}$ are at least 0, and 0 is the constant direction (the direction shifting every x_j by the same amount), so the smallest eigenvalue of C^{-1} is m^2, hence $CP = 1/m^2$. Combined with the correlation length $\xi \approx 1/m$ (1-08 (11)) this gives $CP \approx \xi^2$ (this CP is measured with the gradient in the field variables x_j. In lattice gauge theory, taking the angles of the links as variables changes the constant factor, so we always state which is used. 23-01).

Here we distinguish two gaps. For the free field, the inverse m^2 of CP is the gap of L (the rate at which e^{-tL} removes the deviation from the mean in (5)). On the other hand, the gap of the Hamiltonian when 1-08 (10) is read as an imaginary-time path integral is $1/\xi$ in units of the lattice spacing (1-08 (11); $\approx m$ when m is small). The squared relation $CP \approx \xi^2$ is a free-field computation; whether it holds in general in interacting theories is the question of 23-01, and has not been proved (the same form on a group is in 5-11).

10

Detour: counting eigenvalues — Weyl's law and the shape of a drum

A section you may skip. Kac (1966) relates that at the end of October 1910, Lorentz, invited to give the Wolfskehl lectures in Göttingen, posed at the end of his fourth lecture the problem that "the number of high-frequency standing waves in a cavity should be proportional to its volume regardless of its shape", and Weyl, who was in the audience, proved it in less than two years. For the Dirichlet $-\Delta$ on a planar region Ω it takes the following form Known .

$$N(\lambda) = \#\{k:\ \lambda_k \le \lambda\} \ \sim\ \frac{|\Omega|}{4\pi}\lambda \quad (\lambda\to\infty); \qquad d \text{ dimensions give } N(\lambda) \propto |\Omega|\,\lambda^{d/2} \quad (10)$$

Computed For the unit square (eigenvalues $\pi^2(j^2+k^2)$), $N = 79269$ at $\lambda = 10^6$. Adding to the leading term 79577.5 the correction $-\sqrt{\lambda}/\pi$ for the perimeter 4 gives 79259.2. Kac discussed that

the area and the perimeter can be "heard" from the sequence of eigenvalues (to be exact, are determined by the sequence of eigenvalues alone), and asked about the shape itself. The answer is no: Gordon, Webb and Wolpert (1992) constructed planar regions of different shapes with the same sequence of eigenvalues. Reading λ as the squared momentum Λ^2, the number of modes up to Λ is proportional to Λ^4 in four dimensions. This abundance of short-distance degrees of freedom is one of the starting points for thinking about divergences (4-04, 1-12).

What this chapter can say

Content	Status
The eigenvalues of (1), the convergence of the lower eigenvalues and the divergence of the top end	Said Known (Table 1, Figure 1 Computed)
Reading the lower end of the spectrum from the lattice in the continuum limit (end of section 01)	At the level of Physics . The scaling limit in four dimensions was made by the construction published in 2026-11 (at the level of the committee's assessment; 6-00). Its lower end (the mass gap) has not been shown (6-04)
Hilbert spaces, self-adjointness, the spectrum, the spectral theorem, Weyl's law (10)	Definitions and examples given, theorems cited Known
The variational principle (6), Temple's lower bound (7), Table 2	Said Known , Computed . $\mu = 39$ is a value we could put because we know $\lambda_2 = 4\pi^2$
The Poincaré inequality (8), $C_P = 1/\Delta$, (9), the C_P of products and of Gaussian measures, the $C_P = 1/m^2$ of the free field	Said Known . The one-dimensional line for (9) and the line for (i) of section 09 assume the existence of eigenfunctions. Table 3 and Figure 2 are Computed
The relation between C_P and the correlation length in interacting theories	Not treated (23-01; not proved). $C_P \approx \xi^2$ is a free-field computation
The spectrum of the Hamiltonian of quantum field theory, the Yang–Mills mass gap	Not treated (4-09, 4-11). Nothing is shown

Sources and reproduction

Item	Kind	Source / tools
Tables 1–3, Figures 1–2, the numbers of section 10	Computed	python3 on this machine (NumPy 1.26, SciPy 1.11). Table 3 was found by putting $f = \psi e^{V/2}$, turning L into $-d^2/dx^2 + V'^2/4 - V''/2$ (the $\psi = f\phi_0$ of 2-05 § 08 with $\phi_0 = e^{-V/2}$), and using three-point differences on $[-7, 7]$. The figures are plain JavaScript, with the readouts checked against the tables with node
(4)–(7), the spectral theorem, Weyl's criterion	Known	G. Temple, Proc. R. Soc. Lond. A 119 (1928) 276–293 / W. Ritz, J. reine angew. Math. 135 (1909) 1–61 / J. von Neumann, Math. Ann. 102 (1930) 49–131 (regarded as a representative paper on the spectral theorem in infinite dimensions) / M. Reed, B. Simon, *Methods of Modern Mathematical Physics I* (Academic Press, 1972). All bibliographic data only
(9), the C_P of products and of Gaussian measures in section 09	Known	H. J. Brascamp, E. H. Lieb, J. Funct. Anal. 22 (1976) 366–389 (bibliographic data only; that the inequality on the right of (9) follows as its corollary is not checked) / D. Bakry, M. Émery, Sém. Probab. XIX (1985) 177–206 (the summary at the start checked) / H. Chernoff, Ann. Probab. 9 (1981) 533–535 (abstract checked) / D. Bakry, I. Gentil, M. Ledoux, *Analysis and Geometry of Markov Diffusion Operators* (Springer, 2014; bibliographic data only)
Weyl's law, Lorentz's problem, drums	Known	H. Weyl, Math. Ann. 71 (1912) 441–479 (title checked, bibliographic data only) / M. Kac, Amer. Math. Monthly 73 (1966) no. 4, part II, 1–23 (text checked) / C. Gordon, D. L. Webb, S. Wolpert, Bull. AMS 27 (1992) 134–138, arXiv:math/9207215 (abstract checked)

Next chapter: 1-11 Fourier analysis and distributions. For use in quantum mechanics, see 2-05.

Fourier analysis and distributions — handling things without values at points by waves and averages

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That splitting into frequencies turns differentiation into multiplication. Defining objects without values at points (distributions) by their values when paired with smooth functions. That the propagator diverges near by and decreases as e^{-mr} far away, that its square becomes just barely non-integrable in four dimensions, and that a product of distributions is not determined in general, but is determined if the total roughness is positive.

Prerequisite chapters — 1-10 (Hilbert spaces, orthonormal bases). § 01, § 06 and § 08 use the white noise and free field of 1-09 and the counting of pairings of 1-08 § 06 (formulas restated).

Said first — Theorems are cited with proofs omitted. Products of interacting fields are only touched at the entrance (7-09, 7-10). This chapter shows nothing about the Yang–Mills mass gap.

Order of this chapter

1. Why this chapter — measuring averages instead of values at points
2. Fourier series — smoothness shows in the decay of the coefficients (Figure 1)
3. The Fourier transform — differentiation becomes multiplication, equations become division
4. Test functions and distributions — measuring by testing
5. The Fourier transform of distributions — δ contains all frequencies equally
6. Rulers of roughness — Sobolev spaces and the scaling exponent ((9a)(9b))
7. The propagator — divergent near by, e^{-mr} far away (Figure 2, Table 1)
8. Counting singularities — four dimensions is the borderline (Table 2, Figure 3)
9. The trouble with products — the way of smoothing remains in the answer (Figure 4)

10. When they can be made — if the total roughness is positive

01

Why this chapter — averages instead of values at points

The free field in two or more dimensions has no value at a single point, because the variance at a point is infinite (1-09 § 08). What has meaning is the value averaged with a smooth weight f, and its variance can be written as a double average of the two-point correlation G.

$$\varphi(f) = \int_{\mathbb{R}^d} \varphi(x)\, f(x)\, d^d x, \qquad E\big[\varphi(f)^2\big] = \iint f(x)\, G(x-y)\, f(y)\, d^d x\, d^d y \quad (1)$$

G is called the **propagator**, and formally $E[\varphi(x)\varphi(y)] = G(x-y)$. G is infinite at $x = y$, but the right-hand side of (1) is finite. **Example:** the variance of the value averaged over a box of side ε is, for $d \geq 3$, proportional to ε^{2-d} by $G \approx c\, r^{2-d}$ (§ 07); in four dimensions it is 4 times larger each time the box is halved. Infinite at a point, finite when averaged.

02

Fourier series — splitting a periodic function into waves

Split a function of period 2π into a superposition of waves e^{ikx}. $e^{ikx}/\sqrt{2\pi}$ ($k \in \mathbb{Z}$) is an orthonormal basis of $L^2(-\pi, \pi)$ (the complex version of 1-10 § 02), and the coefficients are extracted by inner products Known .

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\, e^{-ikx}\, dx, \qquad f = \sum_{k \in \mathbb{Z}} c_k\, e^{ikx}, \qquad \frac{1}{2\pi} \int_{-\pi}^{\pi} |f|^2\, dx = \sum_{k \in \mathbb{Z}} |c_k|^2 \quad (2)$$

The last is Parseval's identity. Integrating by parts, differentiation is replaced by multiplying the coefficients by ik.

$$c_k[f'] = ik\, c_k[f] \qquad (f \text{ periodic and smooth}) \quad (3)$$

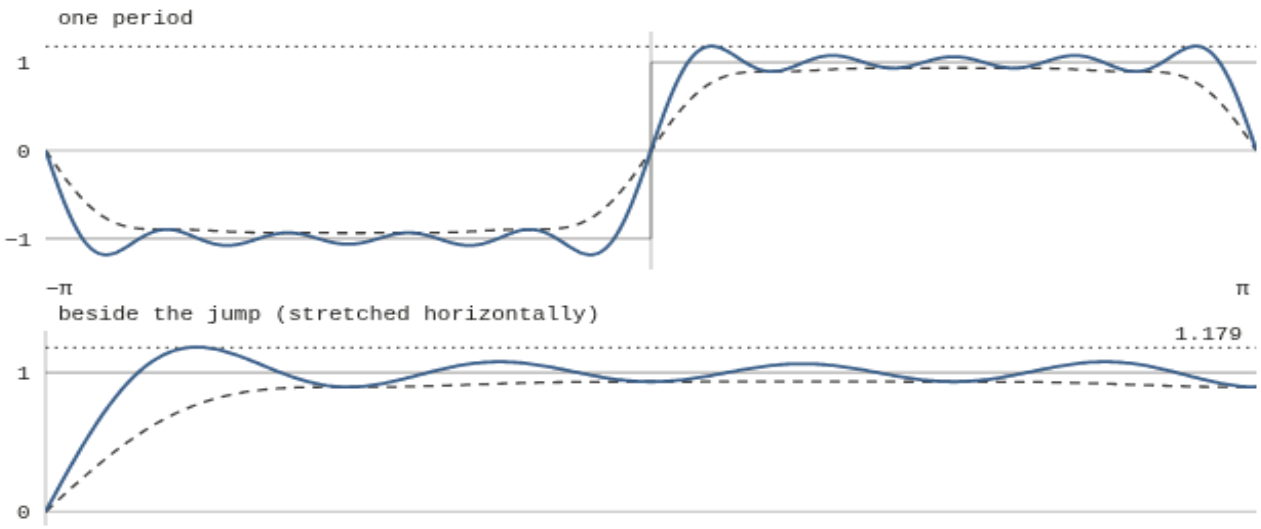
So **smoothness shows in the decay of the coefficients**. If the function is still square-integrable after n derivatives, $\sum_k k^{2n} |c_k|^2 < \infty$. With a jump, the coefficients decrease only as $1/k$.

Example: the partial sums of the square wave $f(x) = \operatorname{sign} x$ are $S_N(x) = \frac{4}{\pi} \sum_{k \text{ odd} \leq N} \frac{\sin kx}{k}$. The mean-square error goes to 0 as $\frac{8}{\pi^2} \sum_{k \text{ odd} > N} k^{-2} \approx \frac{4}{\pi^2 N}$ by Parseval, yet

the maximum overshoots to about 1.179 beside the jump, and the ratio $\frac{1}{\pi}\int_0^{\pi}\frac{\sin t}{t}dt-\frac{1}{2}$ $=0.08949\cdots$ to the jump size 2 remains regardless of N. This is the **Gibbs phenomenon**; Wilbraham (1848) is said to have studied it first, Gibbs's letter of 1899 to have made it widely known, and Bôcher (1906) to have named it Known .

Computed The maximum of S_N (the value at $x=\pi/(N+1)$) is 1.27324, 1.18233, 1.17898 at $N=1,9,999$ (limit $\frac{2}{\pi}\,\mathrm{Si}(\pi)$), and the mean-square error is 0.18943, 0.04040, 0.000405 (Parseval's formula and numerical integration agree; for $N=999$, 0.0004052846 agrees to 10 digits). It converges in mean square, and not uniformly. "Convergence" has meanings of different strength. The **Fejér mean** $\sigma_N=\frac{1}{N+1}(S_0+S_1+\cdots+S_N)$, which averages the partial sums, smooths with a positive kernel, so no overshoot occurs ($|\sigma_N|\le 1$) Known . "Smooth first, then take the limit" is the basic move of approximation in § 04.

Figure 1 Partial sums S_N (solid line) and Fejér means σ_N (dashed) of the square wave (thin line). Above, one period; below, the region beside the jump $0<x<8\pi/(N+1)$ ($0<x<\pi$ at $N<7$) stretched horizontally. Increasing N, above S_N approaches the square wave, yet below the height of the peak stays at about 1.179 (dotted line), and σ_N does not exceed 1. "Play" increases N



N = 9, maximum of partial sum 1.18233 (x = π/(N+1), limit 1.17898), overshoot / jump 0.09116, maximum of Fejér mean 0.93573, mean-square error 0.040395 (Parseval)

Computed The numbers in this figure were computed on this machine (the partial sums of the series are added on the spot; the readout agrees with the computation above to 5 digits at $N=1,3,5,9,19,49,99,199$).

03

The Fourier transform — differentiation becomes multiplication

For functions without a period the frequencies are arranged continuously, and sums are replaced by integrals. The conventions of this book are as follows. Taking the measure as $d^dk/(2\pi)^d$ matches (10) of 1-09.

$$\hat f(k) = \int_{\mathbb{R}d} f(x)\, e^{-ik\cdot x}\, d^d x, \qquad f(x) = \int_{\mathbb{R}d} \hat f(k)\, e^{ik\cdot x}\, \frac{d^d k}{(2\pi)^d} \tag{4}$$

The same rules as for series hold Known . Length is preserved (Plancherel), differentiation becomes multiplication, and convolution $(f*g)(x) = \int f(y) g(x-y)\, d^d y$ becomes a product.

$$\int |f|^2\, d^d x = \int |\hat f|^2\, \frac{d^d k}{(2\pi)^d}, \qquad \widehat{\partial_j f} = ik_j\, \hat f, \qquad \widehat{f*g} = \hat f\, \hat g \tag{5}$$

Then $\widehat{-\Delta f} = |k|^2 \hat f$, and a differential equation with constant coefficients becomes division frequency by frequency. If $m > 0$ the denominator is never 0, and the solution is determined uniquely.

$$(-\Delta + m^2)\, u = f \quad\Longleftrightarrow\quad \hat u(k) = \frac{\hat f(k)}{|k|^2 + m^2} \tag{6}$$

Example: the transform of a Gaussian of width σ is a Gaussian of width $1/\sigma$ (the uncertainty relation of 2-05). The transform of $e^{-m|x|}$ is $\frac{2m}{k^2+m^2}$: the corner at $x=0$ shows up as the slow decrease $1/k^2$ (the one-dimensional propagator of § 07).

04

Test functions and distributions — measuring by testing

Objects without values at points are defined by their values when paired with smooth functions. The partners are called **test functions**, taken to be functions that can be differentiated any number of times and are 0 outside a bounded set (the whole set is $\mathcal{D}$. Example: $e^{-1/(1-|x|^2)}$ on $|x|<1$, 0 outside).

A **distribution** is a linear and continuous rule assigning a number $\langle T, \phi\rangle$ to a test function ϕ (continuous meaning that $\langle T, \phi_n\rangle \to 0$ for sequences ϕ_n whose supports (the closure of the set of points where the function is non-zero) lie in one bounded set and all of whose derivatives approach 0 uniformly) Known . "Measuring by testing" in the title is a metaphor; to be exact, it means regarding this rule $\phi \mapsto \langle T, \phi\rangle$ itself as the object instead of values at points. The English word is distribution, and this is the "distributions" of the chapter title (not the "distribution" of probability). The idea of distributions is said to begin with Sobolev (1936) and to have been systematised by Schwartz Known .

Examples: (a) a function f integrable on bounded sets becomes a distribution by $\langle f, \phi\rangle = \int f\phi\, d^d x$. (b) The δ of $\langle \delta, \phi\rangle = \phi(0)$ cannot be written as the integral of any function. Differentiation is defined by promoting integration by parts to a definition.

$$\langle \partial_j T, \phi \rangle = -\langle T, \partial_j \phi \rangle \quad (7)$$

The right-hand side always makes sense, so every distribution can be differentiated any number of times. For the step function H (1 for $x > 0$, 0 for $x < 0$), $\langle H', \phi \rangle = -\int_0^\infty \phi'\, dx = \phi(0)$, that is, $H' = \delta$.

Approximation: if $\rho \geq 0$ is smooth with $\int \rho = 1$ and $\rho_\varepsilon(x) = \varepsilon^{-d}\rho(x/\varepsilon)$, then $\langle \rho_\varepsilon, \phi \rangle \to \phi(0)$, that is, $\rho_\varepsilon \to \delta$. $\sin nx$ does not converge pointwise, but integration by parts gives $\frac{1}{n}$, so as a distribution it converges to 0.

05

The Fourier transform of distributions — δ contains all frequencies equally

For transforms we use partners wider than $\mathcal{D}$. On the set $\mathcal{S}$ of functions that can be differentiated any number of times and whose values and derivatives decrease faster than any power of $|x|$ (the Schwartz space), the transform maps $\mathcal{S}$ one-to-one onto $\mathcal{S}$ Known . The transform of a continuous linear rule on $\mathcal{S}$ (a **tempered distribution**) is defined to match $\int \hat{f}\phi = \int f\hat{\phi}$.

$$\langle \hat{T}, \phi \rangle = \langle T, \hat{\phi} \rangle \qquad (\phi \in \mathcal{S}) \quad (8)$$

Example: since $\langle \hat{\delta}, \phi \rangle = \hat{\phi}(0) = \int \phi\, d^d x$, $\hat{\delta} = 1$. Something concentrated at one point contains all frequencies with the same weight (the origin of the "white" in white noise).

Conversely, the transform of the constant 1 is $(2\pi)^d\delta$: something of the same size everywhere contains only the frequency 0. The rules of (5) also extend as they are, giving $\widehat{\partial_j T} = ik_j\hat{T}$ Known . This is why in § 07 the propagator can be written $\hat{G} = 1/(|k|^2 + m^2)$.

06

Rulers of roughness — Sobolev spaces and the scaling exponent

Smoothness shows in the decay of the coefficients or of $\hat{f}$ ((3)(5)). We set up two rulers to measure this with numbers. The first is a length that puts weights on frequencies, and the set of objects for which it is finite is called the Sobolev space H^s. The larger s, the smoother the things admitted; the more negative s, the rougher the things admitted.

$$\|f\|_{H^s}^2 = \int (1 + |k|^2)^s\, |\hat{f}(k)|^2\, \frac{d^d k}{(2\pi)^d} \quad (9a)$$

The second is behaviour under rescaling. The $T(\lambda\,\cdot)$ of a distribution is defined by $\langle T(\lambda\,\cdot),\phi\rangle = \lambda^{-d}\langle T,\phi(\cdot/\lambda)\rangle$, promoting the change-of-variables formula to a definition.

$$T(\lambda\,\cdot) = \lambda^{\alpha}\,T \quad \text{(for all } \lambda>0\text{)} \quad \text{: then } \alpha \text{ is called the scaling exponent} \tag{9b}$$

"Ruler" is a metaphor; to be exact, it means the two numbers s of (9a) and α of (9b). Since $\hat{\delta}=1$, $\|\delta\|_{H^s}$ is finite only when $s<-d/2$. The scaling exponent is $-d$ for δ and $-a$ for $|x|^{-a}$.

For a random object T, when $T(\lambda\,\cdot)$ and $\lambda^{\alpha}T$ have the same probability distribution, we take the exponent to be α. For white noise ξ ($E[\xi(f)\xi(g)]=\int fg\,d^dx$, the $\mathbb{R}^d$ version of 1-09 § 07), the covariance of the values of $\xi(\lambda\,\cdot)$ paired with test functions ϕ,ψ is $\lambda^{-2d}\int\phi(x/\lambda)\psi(x/\lambda)\,d^dx=\lambda^{-d}\int\phi\psi\,d^dx$, and a Gaussian with mean 0 is fixed by its covariance, so $\xi(\lambda\,\cdot)$ has the same probability distribution as $\lambda^{-d/2}\xi$. The exponent is $-d/2$. The massless free field ($m=0$, $d\ge 3$) has covariance $G(\lambda x)=\lambda^{2-d}G(x)$ (the near-side formula of § 07 is exact for $m=0$), so its exponent is $1-d/2$. For $d=2$, G is logarithmic, so the same holds up to constants Known . Measuring white noise in spacetime, including time, by the parabolic scaling that multiplies time by λ^2 and space by λ, the exponent is $-(d+2)/2$ (Exercise 2.3 of Hairer's lectures) Known . This exponent decides in § 10 whether products can be made.

07

The propagator — a field responding to δ

The solution of the equation with δ on the right-hand side of (6) is the propagator. "Responding to δ" is a metaphor; to be exact, it means satisfying the first formula of (10) as an identity of distributions, and such a solution is called a **fundamental solution**. From $\frac{1}{|k|^2+m^2}=\int_0^\infty e^{-t(|k|^2+m^2)}dt$ and the transform of the Gaussian, there also comes a representation as the time integral of the fundamental solution $(4\pi t)^{-d/2}e^{-|x|^2/4t}$ (the **heat kernel**) of the heat equation $\partial_t u=\Delta u$.

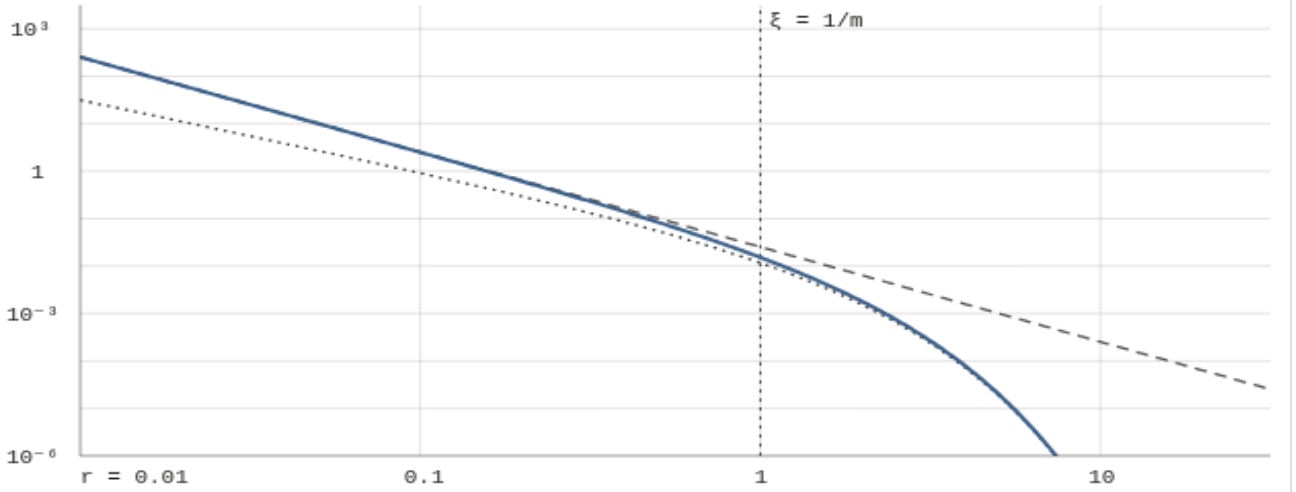
$$(-\Delta+m^2)\,G=\delta, \qquad \hat{G}(k)=\frac{1}{|k|^2+m^2}, \qquad G(x)=\int_0^\infty \frac{e^{-|x|^2/4t-m^2t}}{(4\pi t)^{d/2}}\,dt \tag{10}$$

G is the covariance of the free field, and in physics the Euclidean propagator of a particle of mass m (4-02, 4-03). With $r=|x|$, it is $e^{-mr}/2m$ for $d=1$, $e^{-mr}/4\pi r$ for $d=3$, and in general $(2\pi)^{-d/2}(m/r)^{d/2-1}K_{d/2-1}(mr)$ using the modified Bessel function $K_\nu(s)=\int_0^\infty e^{-s\cosh t}\cosh\nu t\,dt$ Known . What matters are its two faces.

$$G(r) \approx \begin{cases} \dfrac{\Gamma(d/2-1)}{4\pi^{d/2}}\, r^{2-d} & (r \to 0,\ d \geq 3) \\ \dfrac{1}{2\pi} \ln \dfrac{1}{r} & (r \to 0,\ d = 2) \end{cases} \qquad G(r) \approx \frac{1}{2m} \left(\frac{m}{2\pi r} \right)^{\frac{d-1}{2}} e^{-mr} \quad (r \to \infty) \quad (11$$

On the **near face** m disappears and it diverges as $d \geq 2$. On the **far face** it decreases as e^{-mr}. "Face" is a metaphor; to be exact, it means that as $r \to 0$ the ratio of $G(r)$ to the left formula of (11) approaches 1, and as $r \to \infty$ the ratio to the right formula approaches 1. Calling the ξ for which correlations decrease as $e^{-r/\xi}$ the correlation length (2-08), $\xi = 1/m$, and "rate of decrease of correlations = mass" is the free-field version of the definition of the mass gap (4-11).

Figure 2 A log–log graph of the propagator $G(r)$ (solid line). Moving d and m, the near side follows the left formula of (11) (dashed; for $d = 2$ with the constant $(\ln(2/m) - \gamma)/2\pi$ added, for $d = 1$ the value at the origin $1/(2m)$), the far side follows the right formula (dotted), and the bend comes at $\xi = 1/m$ (vertical line). "Play" moves m back and forth



d = 4, m = 1.00, G(0.01) = 2.532e+2 (ratio to the near-side formula 0.9997), G(10ξ) = 4.724e−8 (ratio to the far-side formula 1.036)

Computed The numbers in this figure were computed on this machine (the integral representation of K_ν by the trapezoidal rule; agreeing with Table 1 to 7 digits).

Computed **Table 1** G for $m = 1$ and its ratio to the formulas of (11) (the near side for $d = 2$ adds the constant $\ln 2 - \gamma$). The Bessel function formula and numerical integration of (10) agree to 15 digits.

d	r	G(r)	Ratio to near side	Ratio to far side
2	0.01	0.7514	1.0000	0.3805
4	0.01	253.2	0.9997	8.057
4	1	0.01525	0.6019	1.306
4	10	4.724×10^{-8}	0.0002	1.036

The lattice propagator

The propagator on a lattice of spacing a is the integrand of (10) of 1-09 multiplied by $e^{ik\cdot x}$ and integrated ((10) of 1-09 is its value at $x=0$). For $d=3$, $m=1$ at distance 1, the difference from the continuum G shrinks in proportion to a^2: 0.9% at $a=1/8$ and 0.2% at $a=1/16$ Computed (5-01).

08

Counting singularities — four dimensions is the borderline

Even if it diverges at the origin, it is a distribution by (a) of § 04 if it can be integrated near by. For powers it is a comparison with the dimension ($|S^{d-1}|$ is the area of the unit sphere).

$$\int_{\varepsilon<|x|<1} |x|^{-a}\, d^d x = |S^{d-1}| \int_\varepsilon^1 r^{d-1-a}\, dr \xrightarrow[\varepsilon\to 0]{} \begin{cases} \text{finite} & (a<d) \\ \ln(1/\varepsilon) \text{ divergence} & (a=d) \\ \varepsilon^{d-a} \text{ divergence} & (a>d) \end{cases} \tag{12}$$

$G\sim r^{2-d}$ is $a=d-2<d$, so it is a distribution in every dimension. Its n-th power is $a=n(d-2)$, integrable when $n(d-2)<d$; for the square when $d<4$, for the cube when $d<3$.

Powers of the propagator come from products of fields. The fourth moment of a Gaussian is a sum over ways of pairing (1-08 § 06). In $:\varphi^2: = \varphi^2 - E[\varphi^2]$, with the infinite $E[\varphi^2]=G(0)$ subtracted (the Wick product), only the pairings joining the φ of x with the φ of y remain (the pairings among the x and among the y cancel with what was subtracted) Known .

$$E\big[:\varphi^2:(x)\,:\varphi^2:(y)\big] = 2\,G(x-y)^2, \qquad E\big[:\varphi^2:(f)^2\big] = 2\iint f(x)\,G(x-y)^2\,f(y)\,d^dx\,d^dy \tag{13}$$

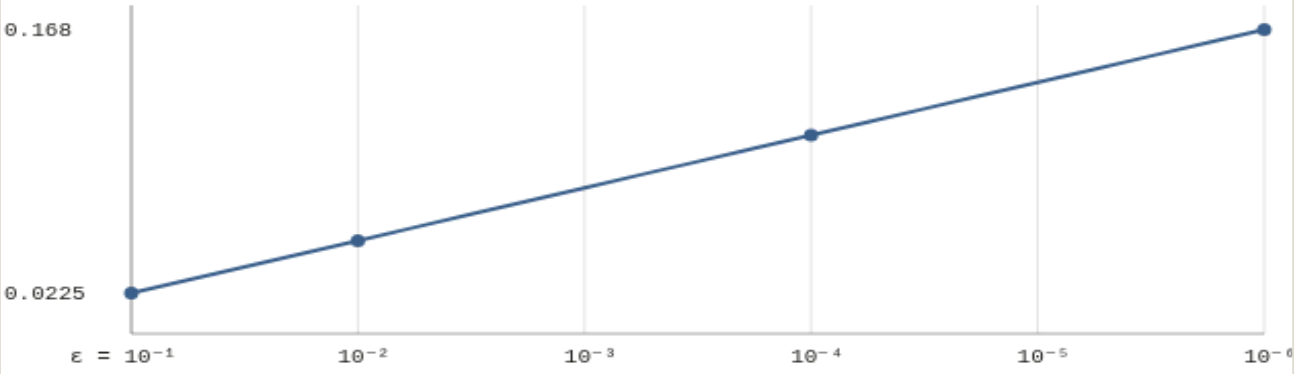
The variance of $:\varphi^2:(f)$ is finite when G^2 is integrable, $d<4$; for the cube, $6G^3$, when $d<3$. A figure setting this beside other computations that switch at four is the dimension slider of 1-12 § 10.

Computed **Table 2** $I_n(\varepsilon)=\int_{\varepsilon<|x|<1} G(x)^n\,d^dx$ ($m=1$, numerical integration in the radius). Pressing a row draws it in Figure 3.

d	n	ε = 0.1	0.01	10^{-4}	10^{-6}	ε → 0
2	3	0.01079	0.01408	0.01426	0.01426	finite
3	2	0.02719	0.03362	0.03440	0.03440	finite
3	3	0.00565	0.01866	0.04763	0.07679	diverges logarithmically

d	n	ε = 0.1	0.01	10^{-4}	10^{-6}	ε → 0
4	1	0.1851	0.1876	0.1876	0.1876	finite
4	2	0.02248	0.05143	0.1097	0.1681	diverges logarithmically
4	3	0.01404	1.598	16041	1.604×10^{8}	diverges as ε^{-2}

Figure 3 Pressing a row of Table 2 (on the keyboard, Tab and Enter) draws $I_n(\varepsilon)$ for that (d, n) against $\log_{10}(1/\varepsilon)$. If finite it levels off, if it diverges logarithmically it is a straight line, and if it diverges as a power it rises steeply (a straight line with a logarithmic vertical axis). The circles are the values in the columns of Table 2



d = 4, n = 2 (diverges logarithmically), increase from ε = 10⁻⁵ to 10⁻⁶ 0.0292, ratio 1.21

Computed The numbers in this figure were computed on this machine (a table embedding to 4 digits the same numerical integration in the radius as Table 2, done at $\log_{10}(1/\varepsilon) = 1, 1.25, \ldots, 6$).

$d = 4, n = 2$ grows by 0.02916 each time ε is divided by 10, agreeing with $\ln 10/8\pi^2$ from (11)(12) Computed . Four dimensions is the borderline where G^2 just stops being integrable, and the coupling constant being dimensionless in four dimensions (4-04) is the same counting of scales Known (1-12).

09

The trouble with products — smoothing before multiplying leaves the way of smoothing behind

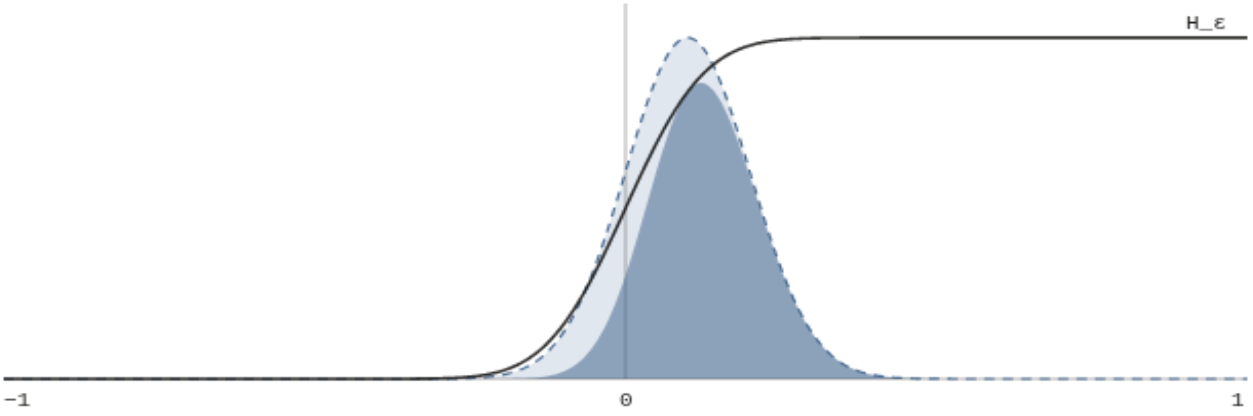
The product of a smooth g and a distribution T is determined by $\langle gT, \phi\rangle = \langle T, g\phi\rangle$. "Smoothing, multiplying and removing the smoothing" for two distributions makes three things happen. **(i) It diverges:** the area of ρ_ε^2 goes to infinity as $\varepsilon^{-d}\int \rho^2$. **(ii) The way of smoothing remains:** smooth H with ρ and δ with ψ (both probability densities), and replace by $x = \varepsilon X$.

$$\int (H * \rho_\varepsilon)\,\psi_\varepsilon\,\phi\,dx = E\big[\mathbf{1}\{Y \le X\}\,\phi(\varepsilon X)\big] \xrightarrow[\varepsilon\to 0]{} c\,\phi(0), \qquad c = P(Y \le X) \quad (Y \sim \rho,\ X$$

The limit is $c\,\delta$, but c is determined by the way of smoothing. Taking ρ to be the standard Gaussian and ψ a Gaussian centred at s gives $c = \Phi(s/\sqrt{2})$ (Φ is the

distribution function of the standard normal distribution), which can be any value between 0 and 1.

Figure 4 The smoothed H (solid line), the shape of the smoothed δ (dashed; height to scale), and their product (dark shading). Moving s changes the area of the product between 0 and 1, and shrinking ε does not change it. The area of the product of δ with itself grows as $1/\varepsilon$. "Play" shrinks ε



ε = 0.1000, s = 1.0, area of the product of H and δ = 0.76025, area of the product of δ with itself = 2.197

Computed The numbers in this figure were computed on this machine (closed formulas; checked against the numerical integration below).

Computed In numerical integration, the area of the product of H and δ is 0.078650, 0.500000, 0.760250, 0.921350 at $s = -2, 0, 1, 2$ (the same at $\varepsilon = 0.1$ and 0.01, agreeing with $\Phi(s/\sqrt{2})$ to 6 digits), and that of δ with itself is 1.038, 2.821, 2.197, 1.038 at $\varepsilon = 0.1$, and 10 times that at $\varepsilon = 0.01$.

(iii) Associativity breaks. For the principal value $\langle \mathrm{p.v.}\frac{1}{x}, \phi\rangle = \lim_{\varepsilon\to 0}\int_{|x|>\varepsilon}\frac{\phi(x)}{x}dx$, as products "smooth function $\times$ distribution", $x\cdot\delta = 0$ and $x\cdot\mathrm{p.v.}\frac{1}{x} = 1$.

$$(\delta\cdot x)\cdot\mathrm{p.v.}\frac{1}{x} = 0, \qquad \delta\cdot\left(x\cdot\mathrm{p.v.}\frac{1}{x}\right) = \delta \quad (15)$$

A product keeping all the properties of the product of functions cannot be put on the whole space of distributions. To be exact, it is said that there is no associative product on distributions (containing at least the continuous functions and δ) that agrees with the ordinary product on continuous functions and for which differentiation satisfies the Leibniz rule $(fg)' = f'g + fg'$ (Schwartz 1954) Known .

10

When they can be made — if the total roughness is positive

The condition for products to be made can be written with the scaling exponent of § 06. Let C^α be the following space. For $0 < \alpha < 1$, the set of Hölder continuous functions of order α ($|f(x) - f(y)| \le C|x-y|^\alpha$, $|x-y| \le 1$); for $\alpha > 1$, the same form of

condition imposed on derivatives; for $\alpha < 0$, the set of distributions whose scaling exponent in the sense of § 06 is locally uniformly at least α. The larger α, the smoother. If the total roughness is positive, products can be made (Proposition 4.14 of Hairer 2014) Known .

$$f \in C^{\alpha},\ \ g \in C^{\beta},\ \ \alpha + \beta > 0 \quad \Longrightarrow \quad f \cdot g \in C^{\min(\alpha,\beta)} \text{ is defined continuously } (16)$$

"Defined continuously" means that extending the product of smooth functions by limits gives a single result independent of how the limit is approached. The same proposition states that, if α is not a natural number, $\alpha + \beta > 0$ is also necessary.

Example: the step function H is bounded but has a jump, and its roughness exponent is 0 ($H(\lambda x) = H(x)$; it lies in $\alpha < 0$ for every C^{α}). The one-dimensional δ is -1. The total is not positive, so (16) cannot be used, consistent with the phenomenon of (ii) in § 09. The free field is slightly rougher than $1 - d/2$ (it lies in C^{α} for $\alpha < 1 - d/2$), so its square can be made as the determined product of (16) only for $d < 2$ Known . In § 08 the variance of $:\varphi^2:(f)$ became finite for $d = 2, 3$ because the divergence of (i) in § 09 was subtracted (renormalised) first, and the fact that it takes a form independent of the way of smoothing depends on probabilistic cancellation (the counting of Gaussian pairings).

The Yang–Mills connection A is expected to be as rough as the free field at short distances Physics , and the $A \wedge A$ of the field strength $F = dA + A \wedge A$ (1-07) is a product that does not satisfy (16) in two or more dimensions. In two and three dimensions the analysis of products of this kind is being advanced by stochastic quantisation (7-09), and in four dimensions the construction published in 2026-11 made the continuum theory as a limit of the lattice (the committee's assessment is that it is proved in Lean; 6-00). There, local products are defined as insertions renormalised on the lattice (8-04). What had been reached before that is in 7-10.

What this chapter can say

Content	Status
The rules of Fourier (2)–(6), the Gibbs overshoot, the Fejér mean	Said Known . The numbers and Figure 1 are Computed
Distributions, differentiation (7), transform (8), rulers (9a)(9b), the scaling exponents of white noise and the free field	Definitions and examples Known . The two exponents have a one-line derivation in § 06

Content	Status
The propagator (10)(11), the correlation length $= 1/m$, the condition for G^n to be integrable, the variance of the Wick product (13)	Said Known . Tables 1 and 2 and Figures 2 and 3 are Computed
The trouble with products (14)(15), the condition for products to be made (16)	The computation of (14) is shown. Figure 4 and the numerical integration are Computed . Schwartz's theorem bibliographic data only; the statement of (16) checked (the proof not followed)
The lattice error a^2 (one point at $d = 3$) / the roughness of the Yang–Mills connection / interacting fields, the mass gap	Observation Computed / expectation Physics / not treated; nothing is shown

Sources and reproduction

Item	Kind	Source / tools
Tables, figures, numbers in the text	Computed	python3 on this machine (NumPy 1.26, SciPy 1.11). The figures are plain JavaScript: the partial sums of Figure 1, the integral representation of K_ν in Figure 2 and the closed formulas of Figure 4 are computed on the spot, and Figure 3 embeds a table computed in python. The readouts were checked against the tables with node
The history of the Gibbs phenomenon	Known	E. Hewitt, R. E. Hewitt, Arch. Hist. Exact Sci. 21 (1979) 129–160 (text checked: Part II) / H. Wilbraham, Cambridge and Dublin Math. J. 3 (1848) 198–201, J. W. Gibbs, Nature 59 (1899) 606 (bibliographic data only)
The beginning of distributions and the impossibility of products	Known	S. Soboleff, Rec. Math. [Mat. Sbornik] N.S. 1(43) (1936) 39–72, L. Schwartz, *Théorie des distributions* (Hermann), L. Schwartz, C. R. Acad. Sci. Paris 239 (1954) 847–848 (all bibliographic data only). The attribution of the beginning of the idea of distributions, and the year of *Théorie des distributions* (said to be 1950–51), not checked (from secondary sources)
(16)	Known	M. Hairer, Invent. Math. 198 (2014) 269–504, arXiv:1303.5113 (text checked: § 1 and Proposition 4.14, including the second half, that $\alpha + \beta > 0$ is also necessary if $\alpha \notin \mathbb{N}$)
The scaling exponent	Known	The exponent $-(d+2)/2$ of spacetime white noise under parabolic scaling: M. Hairer, *An Introduction to Stochastic PDEs*, arXiv:0907.4178 (text checked: Exercise 2.3). The $-d/2$ of spatial white noise and the $1 - d/2$ of the massless free field are from the one-line derivation in § 06

Item	Kind	Source / tools
The transform, $\mathcal{S}$, tempered distributions, Sobolev spaces, the formulas for the propagator	Known	M. Reed, B. Simon, *Methods of Modern Mathematical Physics II* (Academic Press, 1975) and other standard textbooks (bibliographic data only)

Next chapter: 1-12 Detour: what happens only in four dimensions. For uses of the propagator see 4-03; for the axioms of fields with values in distributions, 4-09.

Detour: what happens only in four dimensions — rereading eight cases from the Yang–Mills side

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That rotations in four dimensions split into two copies of SU(2), and the field strength splits into self-dual and anti-self-dual parts. That only in four dimensions is the action independent of scale, so that a field can collapse to a point while keeping its action. That integers appear because S^3 at infinity and SU(2) have the same shape.

Prerequisite chapters — None (it is easier if you have read 1-02 (SU(2)), 1-03 (Lie algebras), 1-06 ($dx_1 \wedge dx_2$ and the Hodge star) and 1-07 (gauge transformations and pure gauge)). For the counting of dimensions, see 2-04.

Said first — This is a detour, so you may skip it. It is a guide to known facts, and shows nothing about the mass gap or the construction of the continuum limit. The main treatment is in 3-05, 3-06 and 4-05.

Order of this chapter

1. A sketch map — the eight of the article and this book
2. Multiplying on the left and on the right
3. The field strength splits in two (Figure 1)
4. Changing the scale does not change the action (Figure 2)
5. Collapsing to a point (Table 1)
6. The sphere at infinity and SU(2)
7. Planes meet in a point (Detour: knots and Wilson loops)
8. The 24-cell is a finite subgroup of SU(2) (Table 2)
9. The area of the sphere, and integrals that begin to diverge in four dimensions (Table 3, Figure 3)
10. The dimension slider (Figure 4)

01

A sketch map — the eight of the article, and where they matter in this book

The starting point of the article "What happens only in four dimensions" was the count that, while there are n directions to move in, the number of planes one can rotate in is

$$\dim SO(n) = \binom{n}{2} = \frac{n(n-1)}{2} \quad (1)$$

In three dimensions planes correspond one-to-one with perpendicular axes, and a rotation can be described by its axis. In four dimensions what is perpendicular to each of the 6 planes is again a plane, and this 6 splitting into 3 + 3 is the starting point of many sections. The way it splits is written concretely in sections 02 and 03, as multiplication of quaternions on the left and on the right.

The table shows where the eight cases of the article appear in this book. The left column links to the headings of the article, the right column to sections of this chapter and to the chapters with the main treatment. Topics not in the article (scale and concentration, the winding number at infinity) are added in sections 04–06. The numbers on the left of the table are the numbers of the cases in the article, separate from the section numbers of this chapter. This chapter itself is a detour, so you may pick out and read only the rows that interest you.

Case in the article	Its face in this book	Section / chapter
01 The axis disappears from rotation	The rotation group splits into two copies of SU(2), the field strength into two	02, 03, 3-05
02 The regular polytopes increase (the 24-cell)	The vertices form a finite subgroup of SU(2)	08, 5-10
03 Knots come untied	Wilson loops do not see knots	07
04 A left hand becomes a right hand	A mirror swaps self-dual and anti-self-dual	03
05 Planes meet in a point / 06 Exotic $\mathbb{R}^4$	Intersection forms; Yang–Mills sees topology	07, 3-06
07 Difficulty is not monotone	The dimension at the borderline where the coupling becomes dimensionless	04, 05, 7-10
08 The ball is largest in five dimensions	Moving the dimension continuously in the same formula	09, 4-05
(not in the article)	Winding number, instanton number	06, 3-05

02

Multiplying on the left and on the right — rotations in four dimensions are two copies of SU(2)

Read a point of four dimensions as a single number. Quaternions are numbers made by adding the symbols i, j, k to the real numbers, and the rules are the following one line. Multiplication is associative but not commutative ($ij = k$, $ji = -k$).

$$x = x_1 + x_2 i + x_3 j + x_4 k, \qquad i^2 = j^2 = k^2 = ijk = -1, \qquad \bar{x} = x_1 - x_2 i - x_3 j - x_4 k,$$

The quaternions of length 1 form S^3, and under multiplication they are the same group as SU(2) (that SU(2) is S^3 is in section 05 of 1-02) Known . Multiplying by q, p of length 1 on the left and on the right gives a rotation of four dimensions, and every rotation can be written in this form. Only the pairs (q, p) and $(-q, -p)$ give the same rotation Known . Unlike $SO(3) \cong SU(2)/\{\pm 1\}$ in three dimensions (1-02), there are two copies of SU(2).

$$x \longmapsto q\, x\, \bar{p} \quad (|q| = |p| = 1), \qquad SO(4) \cong \big(SU(2) \times SU(2)\big)/\{\pm(1,1)\} \ (3)$$

Example. If $q = \cos a + i \sin a$, $p = \cos b + i \sin b$, the (x_1, x_2) plane turns by the angle $a - b$ and the (x_3, x_4) plane by the angle $a + b$. The two sliders of case 01 (the speeds of the xy plane and the zw plane) are the difference and the sum of the left and right angles. We checked with 200 random pairs that (3) is an orthogonal matrix of determinant 1 Computed .

03

The field strength splits in two — the left turns only F^+, the right only F^-

The field strength F is at each point an antisymmetric array $F_{\mu\nu} = -F_{\nu\mu}$, with 6 independent components attached one to each plane. For electromagnetism, these 6 components are the 3 of the electric field E and the 3 of the magnetic field B ((13) of 1-06). In Yang–Mills each component is an element of a Lie algebra (1-03), but in this section we take them to be real numbers. The Hodge star $*$ is the operation taking each plane to the perpendicular plane; with the totally antisymmetric symbol of $\varepsilon_{1234} = 1$,

$$(*F)_{\mu\nu} = \frac{1}{2} \sum_{\rho,\sigma} \varepsilon_{\mu\nu\rho\sigma} F_{\rho\sigma}, \qquad (*F)_{12} = F_{34}, \quad (*F)_{13} = F_{42}, \quad (*F)_{14} = F_{23} \ (4)$$

The $*$ in n dimensions sends 2-forms to $(n-2)$-forms, so only for $n = 4$ does it return to 2-forms. Since $** = 1$ (1-06), it can be split as $F^{\pm} = \frac{1}{2}(F \pm *F)$ ($|F|^2 = \sum_{\mu<\nu} F_{\mu\nu}^2$), and the density of the action and the density of topology (section 06) can be written with the same two. $F = \pm * F$ are called self-dual and anti-self-dual.

$$|F|^2 = |F^+|^2 + |F^-|^2, \qquad \frac{1}{4}\sum \varepsilon_{\mu\nu\rho\sigma} F_{\mu\nu} F_{\rho\sigma} = |F^+|^2 - |F^-|^2 \quad (5)$$

Using (4), $F^{\pm}$ has three independent components each, and arranged so that the squared length is $|F^{\pm}|^2$ they take the following form. The three axes of Figure 1 are in this order.

$$F^+ = \tfrac{1}{\sqrt{2}}(F_{12}+F_{34},\ F_{13}+F_{42},\ F_{14}+F_{23}), \qquad F^- = \tfrac{1}{\sqrt{2}}(F_{12}-F_{34},\ F_{13}+F_{24},\ F_{1}$$

The key point is the relation with (3). The rotation $x \mapsto qx\bar{p}$ turns the 3 components of F^+ by q only and the 3 components of F^- by p only, and does not mix the two Known .

$$\Lambda^2\mathbb{R}^4 = \Lambda^2_+ \oplus \Lambda^2_- \ (3+3), \qquad q \text{ turns } \Lambda^2_+ \ ,\ p \text{ turns } \Lambda^2_- \quad (6)$$

Which one turns F^+ is fixed by convention. Changing the orientation or the assignment of coordinates to quaternions swaps left and right. This book follows (2) and $\varepsilon_{1234} = 1$. In this convention, the matrix multiplying by i on the left is proportional to $dx1 \wedge dx2 + dx3 \wedge dx4$ and self-dual, and the matrix multiplying on the right is proportional to $dx_1 \wedge dx_2 - dx_3 \wedge dx_4$ and anti-self-dual Computed .

Example. $q = \cos a + i \sin a$ turns F^+ about the first axis by the angle $2a$, and returns to the start at $a = 180°$ ($q = -1$). The double cover of 1-02 is visible on F^+. We checked with 200 random elements that (5) holds, that the components mixing the two are at most 2.2×10^{-16}, and that under left multiplication F^- is fixed while the mirror $x_4 \mapsto -x_4$ swaps the two Computed .

Figure 1 Moving the left angle a turns only F⁺, and moving the right angle b turns only F⁻. "Reflect in the mirror" reflects the initial field by $x_4 \to -x_4$, swapping F⁺ and F⁻, before turning (the roles of a and b do not change). Lengths do not change

```
a = 0°, b = 0°, mirror no, F⁺ = (0.7071, 0.0000, 0.6364), F⁻ = (0.7071, 0.0000, 0.2121)
```

Computed The initial field is $F = dx_1 \wedge dx_2 + 0.6\, dx_1 \wedge dx_4 + 0.3\, dx_2 \wedge dx_3$. The figure computes (2)(4) on the spot, agreeing with the python3 values (including the rows with the mirror). The dotted lines are orbits.

The mirror and the Minkowski signature

A mirror changes the sign of the right-hand side of (5) (case 04). The mirror image of an instanton is an anti-instanton, and this is also why the θ term (the term added to the action by multiplying the integral of the right-hand side of (5) by the coefficient θ) changes sign under a mirror (3-05) Known . With the Minkowski signature, $** = -1$, so there are no real self-dual fields (One sign different); this is one of the reasons for thinking in imaginary time (4-02).

> 04

Changing the scale does not change the action — the dimension in which the coupling is dimensionless

Keeping the shape of the field and multiplying only its size by $1/\lambda$ gives $A_\lambda(x) = \lambda A(\lambda x)$, $F_\lambda(x) = \lambda^2 F(\lambda x)$ (a gauge field has the dimension of inverse length). Replacing $y = \lambda x$,

$$S_d[A_\lambda] = \frac{1}{2g^2}\int_{\mathbb{R}^d} \lambda^4 |F(\lambda x)|^2 \, d^d x = \lambda^{4-d}\, S_d[A] \quad (7)$$

so only in four dimensions does the action not depend on size. This means the classical theory has no preferred length, and it is another face of the $[g^2] = 4 - d$ of 2-04 Known . The problem statement's 'the mass gap is not classically visible' (Jaffe–Witten § 6.6) is also related to this (4-07).

One consequence. For a solution (stationary point) of finite action, the action does not change at first order under deformation of size either.

$$\frac{d}{d\lambda} S_d[A_\lambda]\Big|_{\lambda=1} = (4-d)\, S_d[A] = 0 \quad (8)$$

Hence if $d \neq 4$ then $F = 0$, and non-trivial solutions of finite action exist only in four dimensions (assuming enough decay for the deformation to be allowed; an argument of the same kind is attributed to Derrick 1964) Known .

Example. Seen on a lattice it is the same. On a n-dimensional lattice of spacing ϵ, the sum of quantities attached to the small squares (plaquettes) of the lattice plays the role of the action, and it is multiplied by the inverse coupling β (the coefficient in front of the action, corresponding to $1/g^2$; the constant factor depends on normalisation, and in Wilson's convention $\beta = 2N/g^2$) (5-01). The plaquette sum is approximated by $\frac{\epsilon^{4-n}}{4}$ $S_{\mathrm{YM}}(A)$ (S_{YM} is the integral of the action of (7) with $1/2g^2$ removed). So classically β need only keep $\beta\,\epsilon^{4-n}$ constant to compensate for the shrinking of the sum as ϵ^{4-n} ($\beta \propto \epsilon^{n-4}$), and at $n = 4$ it does not move. In three dimensions $\beta \propto 1/\epsilon$, agreeing with $[g^2] = 1$. In physics the four-dimensional β is believed to move like $\log(1/\epsilon)$, but there are doubts too, and in mathematics it is unsolved, as Chatterjee (arXiv:1803.01950 § 3, 2018) writes Physics . The construction published in 2026-11 (6-00) includes in its statement that the ratio of the bare coupling of the chosen family of lattices to log(1/ε) approaches a fixed value (10-06; the committee's assessment is that it is proved in Lean; a statement about the one-loop leading part only). This running is asymptotic freedom (4-06).

Figure 2 Move the size ρ and the dimension d. On the left, a cross-section of the action density (half the action inside the dotted circle); on the right, S(ρ)/S(1) = ρ^(d−4) on a log–log plot, horizontal only for d = 4. "Play" moves ρ

back and forth

```
ρ = 1.000, d = 4, S(ρ)/S(1) = 1.000, [g²] = 0, central density 48/ρ⁴ = 48.00, total action 8π² ≈ 78.957
```

Computed The left is (9), the right the ρ^{d-4} of (7) ($\rho = 1/\lambda$), computed on the spot (agreeing with the ratios of integrals in python3). $d \neq 4$ is the ratio when the same shape is shrunk, not a solution.

05

Collapsing to a point — the borderline dimension

From here on we deal with fields whose values are in a Lie algebra, so we fix $|F|^2$. Write $F_{\mu\nu} = \sum_a F^a_{\mu\nu} t_a$ with $t_a = -iT_a$ ($\mathrm{tr}(t_a t_b) = -\frac{1}{2}\delta_{ab}$), the T_a of 1-04 multiplied by $-i$, and

$$|F|^2 = \sum_{\mu<\nu} \sum_a \left(F^a_{\mu\nu}\right)^2, \qquad S = \frac{1}{2g^2}\int |F|^2\, d^4x = -\frac{1}{2g^2}\int \sum_{\mu,\nu} \mathrm{tr}\, F_{\mu\nu} F_{\mu\nu}\, d^4x$$

Since t_a is anti-Hermitian, $\operatorname{tr} F^2$ is negative, so a minus sign is attached. (5) is also read as summed over each a.

The simplest self-dual solution (the BPST instanton, 3-05) has its action density gathered roundly at the origin. The action density $s = \frac{1}{2}|F|^2$ in units of $1/g^2$ and the fraction of the action inside radius R are as follows for the solution of size ρ ($d^4x = 2\pi^2 r^3 dr$; $2\pi^2$ is the area of the unit S^3) Known . The form of s was recomputed from the BPST connection with the normalisation above Computed .

$$s(x) = \frac{48\,\rho^4}{(|x|^2+\rho^2)^4}, \qquad \int_{\mathbb{R}4} s\,d^4x = 8\pi^2, \qquad \frac{1}{8\pi^2}\int_{|x|<R} s\,d^4x = 1 - \frac{3}{(1+u)^2} + \frac{2}{(1+u)^3}$$

Computed **Table 1** Numerical integration of (9). The total does not depend on ρ, and half always lies inside radius ρ.

ρ	Central density 48/ρ⁴	Total action / 8π²	Inside radius ρ	Radius containing 90%
4	0.1875	1.000000000000	0.500000	8.1065
1	48	1.000000000000	0.500000	2.0266
1/16	3.146 × 10⁶	1.000000000000	0.500000	0.1267

Example. As $\rho \to 0$, the density collapses to a point while keeping the action at $8\pi^2/g^2$. In three dimensions or fewer, shrinking increases the action (Figure 2), so concentration stops, but in four dimensions it does not stop. That handling this concentration is the crux of the analysis of classical Yang–Mills in four dimensions is said to have been known since Uhlenbeck's work of 1982 Known . § 2 of the problem statement writes 'Asymptotic freedom ensures a decisive regularity in the case when classical Sobolev inequalities are borderline.', stating flatly that asymptotic freedom "ensures" a decisive regularity. This chapter reads it as meaning that asymptotic freedom makes up for the classical borderline seen in (7) and Table 1 (estimates falling just short in four dimensions) (an interpretation).

06

The sphere at infinity and SU(2) — where integers come from

The far region of $\mathbb{R}^d$ is a large sphere S^{d-1}. If the action is finite and the decay sufficient, far away A approaches a form that can be written by a gauge transformation alone (pure gauge; 1-07), and the map of the gauge transformation remains (the same mechanism as the monopole of 1-07). To distinguish it from the coupling constant g, in this section we write that map as h. The classes of maps, with maps that can be continuously deformed into each other regarded as the same, form the homotopy group π_{d-1}.

$$h:\ S^3 \to SU(2) \cong S^3, \qquad k = \deg h \in \pi_3(S^3) = \mathbb{Z}, \qquad S \geq \frac{8\pi^2}{g^2}\,|k| \quad (10)$$

In four dimensions both the far region and SU(2) are S^3, and maps between spheres of the same dimension have an integer, "how many times it covers" (the degree). To be exact, choose one general point of the target S^3, attach $+1$ to each point of its preimage where h preserves orientation and -1 where it reverses it, and add them up (the same value whichever point is chosen). This is the instanton number k.

The inequality follows from (5). With the normalisation of section 05 one can write $k = \frac{1}{16\pi^2}\int (|F^+|^2 - |F^-|^2)\,d^4x$, so $S = \frac{1}{2g^2}\int (|F^+|^2 + |F^-|^2)\,d^4x \geq \frac{8\pi^2}{g^2}|k|$, with equality when self-dual or anti-self-dual. The BPST solution has $k = 1$ (3-05) Known .

Example. According to the tables of homotopy groups, the classes $\pi_{d-1}(S^3)$ into SU(2) = S^3 are $0, 0, \mathbb{Z}, \mathbb{Z}/2$ for $d = 2, 3, 4, 5$, and the classes into U(1) = S^1 are $\mathbb{Z}$ only for $d = 2$ (the number of vortices; the same mechanism as the monopole of 1-07). For a compact simple Lie group π_3 is always $\mathbb{Z}$, and SU(3) too has k Known .

07

Planes meet in a point — intersection forms and exotic $\mathbb{R}^4$

As in case 05, two planes in four dimensions usually meet in one point. Two closed surfaces in a closed four-dimensional manifold also meet in finitely many points, and the sum of ± 1 with orientations included does not change when the surfaces are moved. This is the intersection form (a symmetric matrix of integers).

$$\dim(\Sigma_1 \cap \Sigma_2) = 2 + 2 - 4 = 0, \qquad Q(\Sigma_1, \Sigma_2) = \sum_{p \in \Sigma_1 \cap \Sigma_2} (\pm 1) \quad (11)$$

Example. The intersection form of $\mathbb{CP}^2$ is (1), and that of $S^2 \times S^2$ is $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ Known . Using the space of solutions for anti-self-dual connections ($F^+ = 0$ in (6); the same as self-dual connections with the orientation reversed, and for the convention of the primary sources see 3-06) as a tool, Donaldson arrived at the theorem that "the definite intersection form of an oriented closed smooth four-dimensional manifold can be diagonalised". He is said to have shown the simply connected case in 1983 and later extended it to a form without the assumption of simple connectivity (details in 3-06). Contrasted with Freedman's topological theorem, spaces homeomorphic but not diffeomorphic to $\mathbb{R}^4$ (exotic $\mathbb{R}^4$) appear, and Taubes is credited with first showing that there are continuum-many of them Known . The main treatment is in 3-06.

Detour: knots and Wilson loops

A subsection you may skip.

A Wilson loop is the trace of the holonomy of a closed path (1-07), and in three-dimensional Chern–Simons theory its expectation value gives a knot invariant (the Jones polynomial) (Witten 1989) Known . In four dimensions every loop comes untied (case 03), and Yang–Mills is not topological either, so the expectation value is determined by the size and shape of the path, and in theories without matter fields an area law is taken as the sign of confinement Physics (5-02).

08

The 24-cell is a finite subgroup of SU(2)

Multiplying the vertices of the 24-cell (two coordinates ± 1, two 0) on the left by $(1+i)/\sqrt{2}$ and scaling by $1/\sqrt{2}$ gives the following 24 unit quaternions Computed . They form a finite subgroup of SU(2) closed under multiplication (the binary tetrahedral group; its image in SO(3) is the 12 rotations of the regular tetrahedron) Known .

$$\bar{T} = \left\{ \pm 1,\ \pm i,\ \pm j,\ \pm k,\ \tfrac{1}{2}(\pm 1 \pm i \pm j \pm k) \right\}, \qquad |\bar{T}| = 24 \ \text{(12)}$$

The 120 vertices of the 600-cell likewise form the binary icosahedral group $\bar{I}$.

Example. In lattice gauge theory (5-01) elements of a finite group can be placed on the links. Petcher–Weingarten 1980 computed with $\bar{T}$, $\bar{O}$, $\bar{I}$, and according to the abstract, for $\bar{I}$ the mean plaquette, Wilson loops and estimates of the string tension agreed with SU(2) over a wide range of couplings. At sufficiently weak coupling it is said to behave differently from SU(2) because of the gaps between elements Physics (sources and details in 5-10).

Computed **Table 2** All three checked to be closed under products. "Angle to the farthest point" is the maximum (an estimate) of the angle from 200,000 random points on S^3 to the nearest element.

Group	Order	Nearest-neighbour distance	Number of edges	Angle to the farthest point
$\bar{T}$ (24-cell)	24	1	96	44.7°
$\bar{O}$	48	$\sqrt{2-\sqrt{2}} \approx 0.7654$	144	31.2°
$\bar{I}$ (600-cell)	120	$1/\varphi \approx 0.6180$	720	22.2°

09

The area of the sphere, and integrals that begin to diverge in four dimensions

The volume of the ball in case 08 comes from the formula for the area of the unit sphere. This formula makes sense even when d is not an integer, and the integrals of the form first met in perturbation theory (4-03) also become functions of d.

$$|S^{d-1}| = \frac{2\pi^{d/2}}{\Gamma(d/2)}, \qquad I(d) = \int_{\mathbb{R}d} \frac{d^d k}{(k^2+1)^2} = |S^{d-1}| \int_0^\infty \frac{k^{d-1} dk}{(k^2+1)^2} = \pi^{d/2}\,\Gamma\!\left(2-\frac{d}{2}\right) \quad (13$$

Γ is the gamma function, the extension of the factorial to real numbers ($\Gamma(n) = (n-1)!$, $\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$, $\Gamma(x+1) = x\Gamma(x)$). The integrand is k^{-4} far away and the volume $k^{d-1}\,dk$, so it converges only for $d < 4$, and four dimensions is the borderline where divergence begins, logarithmically. In the dimension where the coupling is dimensionless, the divergence is exactly logarithmic (4-04). Computing at $d = 4-\epsilon$ (dimensional regularisation, 't Hooft–Veltman 1972, Bollini–Giambiagi 1972) and cutting off the momentum at K give

$$I(4-\epsilon) = \frac{2\pi^2}{\epsilon} + O(1), \qquad \int_{|k|<K} \frac{d^4 k}{(k^2+1)^2} = 2\pi^2 \ln K - \pi^2 + O(K^{-2}) \quad (14)$$

and $1/\epsilon$ and $\ln K$ correspond with the same coefficient $2\pi^2 = |S^3|$.

Computed **Table 3** The formula (13) and numerical integration.

d	I(d) formula	I(d) numerical integration	(4−d)·I(d)
2	3.141593	3.141593	6.283185
3	9.869604	9.869604	9.869604
3.99	1957.039052	–	19.570391
3.999	19722.225249	–	19.722225

Example. The last column approaches $2\pi^2 = 19.739209$, and for the cut-off, $I_K - 2\pi^2 \ln K = -9.869604$ at $K = 10^4$ ($-\pi^2$). The volumes of balls in the article ($V_5 = 5.2638$ and so on) also agree with the same formula Computed .

Using $\Gamma(x+1) = x\Gamma(x)$, $(4-d)\,I(d) = 2\pi^{d/2}\Gamma(3-\frac{d}{2})$, and the right-hand side is smooth across $d = 4$. The divergence is only in the factor $1/(4-d)$, and the remaining coefficient is $2\pi^2\Gamma(1) = 2\pi^2$ at $d = 4$. For the cut-off, $I_K = \pi^2\big(\ln(1+K^2) + \frac{1}{1+K^2} - 1\big)$, and the growth $2\pi^2 K^4/(K^2+1)^2$ for each increase of $\ln K$ by 1 approaches the same $2\pi^2$. Figure 3 sets these two side by side.

Figure 3 Bringing d towards 4 with the left slider, I(d) (dotted) diverges but (4−d)·I(d) (solid) approaches 2π². Enlarging the cut-off K with the right slider, the growth of I_K per ln K approaches the same 2π²

```
d = 3.0000, I(d) = 9.8696, (4−d)·I(d) = 9.8696 | K = 10.00, I_K = 35.7775, I_K − 2π² ln K =
-9.6737 (2π² = 19.7392)
```

Computed The left computes $I(d) = \pi^{d/2}\Gamma(2 - \frac{d}{2})$ and $(4-d)I(d) = 2\pi^{d/2}\Gamma(3 - \frac{d}{2})$, the right $2\pi^2 K^4/(K^2+1)^2$ and I_K, on the spot (the gamma function by Lanczos's approximation; agreeing with the python3 values). The left slider moves the u of $4 - d = 10^{-u}$.

10

The dimension slider — rows that line up at four and rows that do not

We line up the d-dependences scattered through the book on a single slider. Each row divides by the sign of a power $p(d)$, and at the borderline $p = 0$ the power vanishes, leaving a logarithm or invariance. The six rows of the four of spacetime reach their borderline at the same $d = 4$. The remaining four rows switch somewhere other than four, or for a different reason. Known To be exact, the six rows line up because of the same counting of dimensions (the 2 of the propagator $1/k^2$, twice), and this does not mean that construction in four dimensions is impossible.

Figure 4 Moving d changes the sentences of all eleven rows at once. Green = below, thick line = borderline, ochre = beyond, grey = a d the source does not treat

Computed The borderlines were computed on this machine; the sentences are copies from each chapter. If JavaScript is disabled, see Table 4.

Table 4 The rows of Figure 4.

Quantity (chapter)	Label	Borderline	Below (small d)	At the borderline	Beyond (large d)
[g^2] and the type of divergence (4-04 Table 2)	four of spacetime	4	finitely many divergent diagrams	finitely many types, at all orders; the coupling runs logarithmically	types increase without limit
Trace coefficient 1−d/4 (3-03 (8))	four of spacetime (classical)	4	non-zero	identically 0 (no scale in the classical theory)	non-zero (opposite sign)
Instanton action ρ^(d−4) ((7), 3-05)	four of spacetime (classical)	4	spreading out reduces the action	independent of ρ	collapsing reduces the action
Wilson–Fisher, Σg_k^2 (4-04 (11), 7-05)	four of spacetime	4	fixed point $\hat{u}^*>0$, Σg_k^2 finite	$\hat{u}$ goes to 0 at a logarithmic rate, Σg_k^2 grows as the logarithm of the number of steps	$\hat{u}$ goes to 0 as a power
Room in stochastic quantisation (7-09 Table 2)	four of spacetime	4	positive room (locally subcritical)	room 0, the symbols do not fit into finitely many	negative room
Delta well, analogy with D=d−2 (4-07 Table 5)	four of spacetime (analogy)	4	D=1: g settles at 2	D=2: 1/g is proportional to ln K	D=3: g is tuned as a power
Variance at a point $G_a(0)$ (1-09 Table 3)	contrast (two)	2	finite	diverges as ln(1/a)	diverges as a^(2−d)
Variance of the Wilson loop (7-06 Table 2)	contrast (three)	3	finite (area)	proportional to ln(1/a)	proportional to 1/a
Overlap of vacua, D=d−1 (4-09 § 07)	four of space	5	finite	diverges logarithmically (spacetime of five dimensions)	—

Quantity (chapter)	Label	`Borderline`	Below (small d)	At the borderline	Beyond (large d)
Complete families (5-11, Lean `no_perfectZ_of_five`)	four by coincidence	—	exist	only for integer d	do not exist
Steps of construction for YM (7-01 Table 2)	steps 1–6	—	d=2: done, done, done, done, partial, —; d=3: done, done (lattice), partial, not yet, not yet, not yet; d=4: done, done (lattice), done, partial, done, not yet (the third to fifth are by the construction published in 2026-11; the volume limit only along a subsequence. The four-dimensional row of 7-01 Table 2 is what the literature before 2026 had reached (done, done (lattice), partial, not yet, not yet, not yet), and here the published construction is laid over it. The committee's assessment is that it is proved in Lean; 6-01)		

What this chapter can say

Content	Status
(3)–(8), the swap under a mirror	Known . (8) assumes decay. Numerical checks are Computed (the surjectivity of (3) is not checked)
The running of $\log(1/\epsilon)$ for β on the lattice / the difference between finite groups and SU(2) at weak coupling	Physics . For the former, the source states plainly that it is unsolved in mathematics. The source for the latter is in 5-10
(9), Table 1, the normalisation of $\lvert F\rvert^2$ / the reading of the problem statement in section 05	The form of the density and the integrals are Computed (that it is a solution, 3-05) / this chapter's interpretation
(10)(11), the theorems of section 07	Known (original papers bibliographic data only)
(12)–(14), Tables 2 and 3, Figure 3	Computed (the covering angle is an estimate from samples)
Figure 4, Table 4	Copies from each chapter. The arrangement is this book's
The mass gap, the continuum limit	This chapter shows nothing

Sources and reproduction

Item	Kind	Source / tools
Tables 1–4, the normalisation of section 05, the reference values of Figures 1–3	Computed	python3 (NumPy, SciPy). The figures are JavaScript in the browser (checked by running the same functions with node)
The eight cases, the volume of the ball, the 24-cell	This site	What happens only in four dimensions
The lattice and β, quotations from the problem statement	Known	S. Chatterjee, arXiv:1803.01950 § 3 / A. Jaffe, E. Witten, Quantum Yang–Mills theory (the problem description of the Millennium Prize Problem of the Clay Mathematics Institute, PDF at claymath.org) § 2, § 6.6 (text checked for both)
Section 08	Known	D. Petcher, D. H. Weingarten, Phys. Rev. D 22 (1980) 2465 (abstract checked; the source for the difference at weak coupling is in 5-10)
(3), the π_{d-1} of section 06, the claims and attributions of section 07	Known	Wikipedia "Rotations in 4-dimensional Euclidean space", "Homotopy groups of spheres", "Donaldson's theorem" (which takes the 1983 version to assume simple connectivity and the extension to general fundamental groups to be the 1987 paper), "Exotic R4" (secondary sources)
Bibliographic data only (text not read)	Known	Conway–Smith, *On Quaternions and Octonions* (2003) / Belavin–Polyakov–Schwartz–Tyupkin, Phys. Lett. B 59 (1975) 85 / Derrick, J. Math. Phys. 5 (1964) 1252 / Uhlenbeck, Commun. Math. Phys. 83 (1982) 11, 31 / Donaldson, J. Differential Geom. 18 (1983) 279 / Donaldson, ibid. 26 (1987) 397 / Freedman, ibid. 17 (1982) 357 / Taubes, ibid. 25 (1987) 363 / Witten, Commun. Math. Phys. 121 (1989) 351 / 't Hooft–Veltman, Nucl. Phys. B 44 (1972) 189 / Bollini–Giambiagi, Nuovo Cim. B 12 (1972) 20

Next chapter: 2-01 Lagrangians and the action. It continues in 3-05, 3-06, 4-07 and 7-10.

Lagrangians and the action — motion can be written as a stationary point of "one number attached to the whole path"

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That the equations of motion can be written as the stationarity condition of the action. That the Euler–Lagrange equations come out of it. That continuous symmetries give conserved quantities (Noether's theorem), and local symmetries give constraints.

Prerequisite chapters — 1-01 Vector spaces and linear maps (rotation matrices, eigenvalues of real symmetric matrices). Otherwise, multivariable calculus and Newton's equation of motion.

Said first — We restrict to finitely many degrees of freedom (fields are in 2-02, the Yang–Mills action in 3-03). The proofs stay within smooth paths and variations with fixed ends. We do not answer why the action principle holds.

Order of this chapter

1. Why start from the action — three advantages
2. The action, a single number — the integral of kinetic energy − potential energy
3. Shifting the path a little — discretised, a function of many variables (Figure 1)
4. The Euler–Lagrange equations — one integration by parts; no choice of coordinates
5. Not necessarily a minimum — eigenvalues and conjugate points (with computations)
6. Noether's theorem — conserved quantities from symmetries (Figure 2)
7. Detour: the discretised action has Noether too — why angular momentum is conserved exactly
8. Local symmetries give constraints — a foretaste of gauge theory

01

Why start from the action

Newton's equation of motion $m\ddot{q} = F$ is a local rule that fixes the acceleration at each moment. In this chapter we write the same motion in another form. Fix the start and the arrival, and give the **whole path** between them a single number S (the **action**). The actual motion is chosen as the path for which S does not change at first order when the path is shifted slightly. This is the **action principle** (also called the principle of least action).

There are three advantages. It **does not depend on coordinates** (§ 04). **Conservation laws come out automatically** from symmetries (§ 06). **It is the language of field theory**: Yang–Mills theory is defined by an action, the integral of $\mathrm{tr}\, F^2$ (3-03), and the path integral too sums over paths with the action in the exponent (2-07).

As a metaphor, the equation of motion is "step-by-step directions", and the action principle is the view of "comparing whole routes on a map". To be exact, for systems with conservative forces both give the same equation (§ 04), and the only difference is what is easy to see.

02

The action, a single number

For a particle on a line, let $q(t)$ be its position at time t, $\dot{q} = dq/dt$ its velocity, m its mass and $V(q)$ the potential energy. The **Lagrangian** L is the kinetic energy **minus** the potential energy, a different quantity from the energy, which is their sum. The number obtained by integrating it along the path is the **action**.

$$L(q, \dot{q}) = \tfrac{1}{2} m \dot{q}^2 - V(q)$$

$$S[q] = \int_{t_0}^{t1} L\big(q(t), \dot{q}(t)\big)\, dt$$

The square brackets show that it is a map that takes a whole path and returns a number (a **functional**).

Example. Take $V = 0$, $m = 1$, and ends $q(0) = 0$, $q(1) = 1$. For the uniform path $q = t$, $S = 0.5$. For the detour $q = t + 0.3 \sin \pi t$ joining the same ends, $S = \tfrac{1}{2} \int_0^1 (1 + 0.3\pi \cos \pi t)^2\, dt = \tfrac{1}{2}(1 + 0.09\pi^2/2) \approx 0.722$. Taking a detour increases the action.

03

Shifting the path a little

An extremum of a function was a point where the first-order change under a small change of the variable is 0. For the action we move the path. Taking a function $\eta(t)$ that vanishes at the ends, $S(\varepsilon) = S[q + \varepsilon\eta]$ is an ordinary function of the real number ε. The **action principle** is the requirement "$S'(0) = 0$ for every η with fixed ends", and a path satisfying it is called a **stationary point**.

Discretised, it becomes a function of finitely many variables

Dividing time into N equal parts (step h), making the path a broken line and approximating the integral by the trapezoidal rule, the action becomes a function of many variables, the intermediate positions $q_1, \dots, q_{N-1}$ (the ends q_0, q_N fixed).

$$S_N = \sum_{k=0}^{N-1} \Big[\frac{m(q_{k+1} - q_k)^2}{2h} - \frac{h}{2}\big(V(q_k) + V(q_{k+1})\big) \Big]$$

$$\frac{\partial S_N}{\partial q_k} = -h\Big[m\,\frac{q_{k+1} - 2q_k + q_{k-1}}{h^2} + V'(q_k) \Big]$$

The action principle becomes "the gradient is 0", and this is a difference version of $m\ddot{q} = -V'(q)$ (used again in § 07).

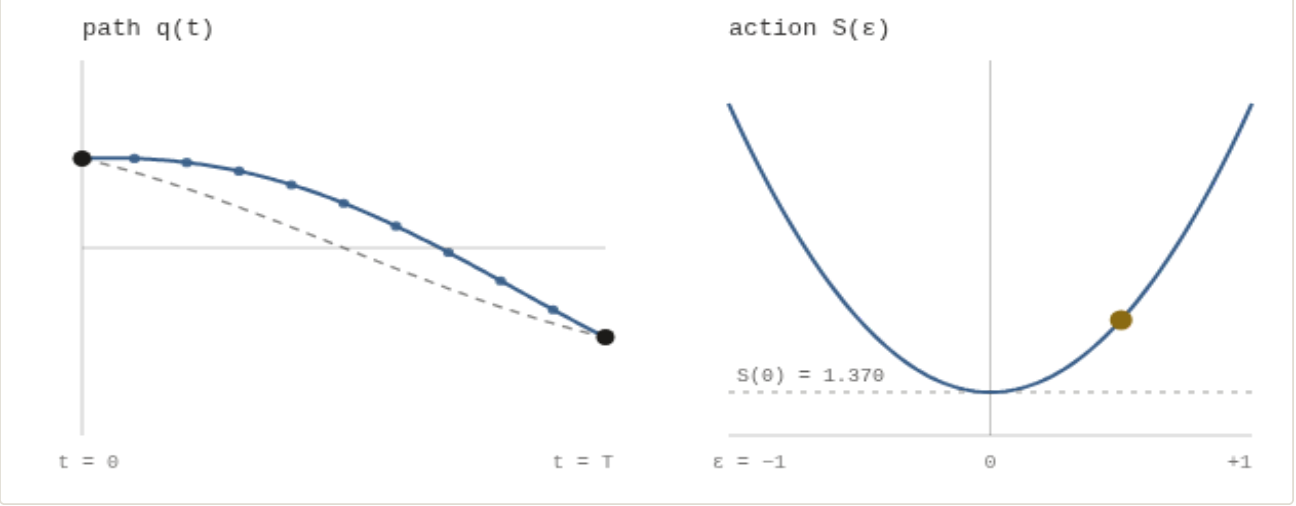
Example: a particle on a spring

Take $V = \frac{1}{2}m\omega^2 q^2$, $m = 1$, time length $T = 1$, and ends $q(0) = 1$, $q(1) = -1$. The solution of the equation of motion is $q_{\mathrm{cl}}(t) = \sin\big(\omega(\tfrac{1}{2} - t)\big)/\sin(\omega/2)$ (cl stands for "classical path"). For the shift $\eta_n = \sin n\pi t$ ($n = 1, 2, 3$), the action is a quadratic in ε.

$$S(\varepsilon) = \omega\cot\frac{\omega}{2} + \frac{\varepsilon^2}{4}\big((n\pi)^2 - \omega^2\big)$$

The absence of a first-order term is what "stationary" means. The second-order term comes from $\frac{1}{2}\int(\dot{\eta}^2 - \omega^2\eta^2)\,dt$. Moving ε in Figure 1 moves the action along a parabola, with slope 0 at $\varepsilon = 0$.

Figure 1 Moving ε, the shape n and ωT changes the shifted path (left) and the action S(ε) (right)S(ε) = 1.764, coefficient of the second-order term +1.579, opens upwards = a minimum in this direction



Computed The curves draw the formula above in the browser, and the formula was checked against numerical integration on this machine (Table 1). The dotted line is the classical path, and the dots are the variables when discretised with $N = 10$.

Table 1 Values of the action ($m = T = 1$, $q(0) = 1$, $q(1) = -1$)

$\omega T/\pi$	n	$S(0)$	$S(1)$	Coefficient of the second-order term
0.5	1	1.5708	3.4213	+1.851
1.0	1	0.0000	0.0000	0
1.5	1	−4.7124	−7.7966	−3.084
1.5	2	−4.7124	−0.3944	+4.318

Computed The difference between numerical integration by Simpson's rule (20000 intervals) and the formula is at most 4×10^{-15}. The first-order coefficient is also at most 5×10^{-13}.

04

The Euler–Lagrange equations

We compute the condition of § 03 without discretising. Differentiate with respect to ε and integrate the second term by parts (the boundary term vanishes because $\eta = 0$ at the ends).

$$S'(0) = \int_{t0}^{t1} \left[\frac{\partial L}{\partial q}\,\eta + \frac{\partial L}{\partial \dot{q}}\,\dot{\eta} \right] dt$$

$$= \int_{t0}^{t1} \left[\frac{\partial L}{\partial q} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}} \right] \eta \, dt$$

If this is 0 for every η, the bracket is 0 at each time (the bracket is continuous, so if it is positive somewhere it is positive nearby, and taking a η positive only nearby makes the integral positive: the **fundamental lemma of the calculus of variations**). This gives the **Euler–Lagrange equations**. With several degrees of freedom there is one for each component.

$$\frac{d}{dt}\frac{\partial L}{\partial \dot q} - \frac{\partial L}{\partial q} = 0 \quad \text{(EL)}$$

Example. If $L = \tfrac{1}{2}m\dot q^2 - V(q)$, then $m\ddot q = -V'(q)$: Newton's equation for a conservative force itself.

No choice of coordinates

This derivation does not use what q represents. It may be an angle (**generalised coordinates**), and equations of the same form hold Known . This is because the same path gets the same action whatever the coordinates (1-05). Writing a pendulum with the angle θ, $L = \tfrac{1}{2}ml^2\dot\theta^2 + mgl\cos\theta$ gives $ml^2\ddot\theta = -mgl\sin\theta$. The tension does not appear.

Adding $dG(q,t)/dt$ to L does not change the equations (the action changes only by a quantity fixed by the end values) Known . The **θ term** of Yang–Mills theory is the four-dimensional version of this, with the form of a four-dimensional divergence (a **total divergence**) instead of a total time derivative. To be exact, the θ term does not change the classical equations, but it matters in quantum theory through the topology of the field (3-05).

05

Not necessarily a minimum

The action principle asks only that the first-order change be 0; whether it is a minimum is decided by the second-order term. In the example of § 03 it is the sign of the coefficient $((n\pi)^2 - \omega^2)/4$. If $\omega T < \pi$ it is positive for every n, and a general η, expanded in a sine series, splits into a sum over each n, so the classical path really is a minimum (because the action is quadratic).

When ωT exceeds π, the coefficient of $n = 1$ becomes negative, and in that direction the action **decreases** (a **saddle point**). Moving $\omega T/\pi$ either side of 1 in Figure 1 turns the parabola of $n = 1$ over. At exactly $\omega T = \pi$, moving in the direction of $n = 1$ does not change the action (it increases in the direction of $n \geq 2$), and the solutions joining the same ends form a family $\cos\pi t + c\sin\pi t$. In general, a time at which a family of nearby solutions leaving the same point at time 0 gathers again at one point is called a **conjugate point** of 0 Known . The conjugate points in this example are $t = \pi/\omega, 2\pi/\omega$, …, and $\omega T = \pi$ is the case where the final time falls exactly on the first conjugate point.

The second-order term is an eigenvalue problem for a matrix

Discretised, the second-order term is the Hessian matrix $H_{jk} = \partial^2 S_N / \partial q_j \partial q_k$. It is real symmetric, so it can be diagonalised (1-01), and a local minimum means all eigenvalues are positive. The number of negative eigenvalues is the number of directions in which the action can be reduced, called the **Morse index** (here a name for the discretised matrix). For the harmonic oscillator H is tridiagonal (diagonal $2m/h - m\omega^2 h$, neighbours $-m/h$); its eigenvalues divided by h approach $m((n\pi/T)^2 - \omega^2)$, and $\lfloor \omega T/\pi \rfloor$ of them are negative.

Table 2 Hessian matrix of the discretised action ($m = T = 1$, $N = 200$)

$\omega T/\pi$	Number of negative eigenvalues	Smallest eigenvalue $/\,h$	$\pi^2 - (\omega T)^2$
0.50	0	7.4020	7.4022
1.05	1	−1.0118	−1.0116
2.50	2	−51.8156	−51.8154
3.20	3	−91.1953	−91.1951

Computed All eigenvalues found with numpy's `eigvalsh`. With $N = 20$ the numbers are the same, and the smallest eigenvalues differ by about 0.02.

Without discretising too, the directions that reduce the action in this example are those among $\sin n\pi t$ with $n\pi < \omega T$, and their number equals the number of conjugate points $k\pi/\omega$ in $(0, T)$. This is an instance of the **Morse index theorem** (the number of negative directions of the second-order term equals the number of conjugate points in $(0, T)$, counted with multiplicity) Known .

This view is extended to infinite dimensions in 1-10. The smallest eigenvalue here (of the second-order term of the action) and the mass gap of 4-11 (the gap in the spectrum of the Hamiltonian) are different things.

06

Noether's theorem

If L does not contain the coordinate q, the Euler–Lagrange equation says that $p = \partial L/\partial \dot q$ (the **conjugate momentum**) is conserved. $L = \frac{1}{2}m(\dot r^2 + r^2\dot\varphi^2) - V(r)$, a central force written in polar coordinates, does not contain φ, and the angular momentum $mr^2\dot\varphi$ is conserved. Noether's theorem is this detached from coordinates.

For transformations $q \mapsto \Phi_s(q)$ labelled continuously by a real number s and equal to the identity at $s = 0$ (translations, rotations by the angle s), let $K(q) = \frac{d}{ds}\Phi_s(q)\big|_{s=0}$ be the direction of motion. **Noether's theorem**: if $L\big(\Phi_s(q), \frac{d}{dt}\Phi_s(q)\big) = L(q, \dot q)$ for every path, then the following Q is constant along the motion Known .

$$Q = \sum_i \frac{\partial L}{\partial \dot{q}_i} K_i(q) \quad \text{(N)}$$

Proof. Differentiating the assumption with respect to $s = 0$ gives $\sum_i (\partial L/\partial q_i) K_i + \sum_i (\partial L/\partial \dot{q}_i)\, \dot{K}_i = 0$. Replacing $\partial L/\partial q_i$ by $\frac{d}{dt}(\partial L/\partial \dot{q}_i)$ using the Euler–Lagrange equation, the left-hand side is the derivative of a product dQ/dt. ∎

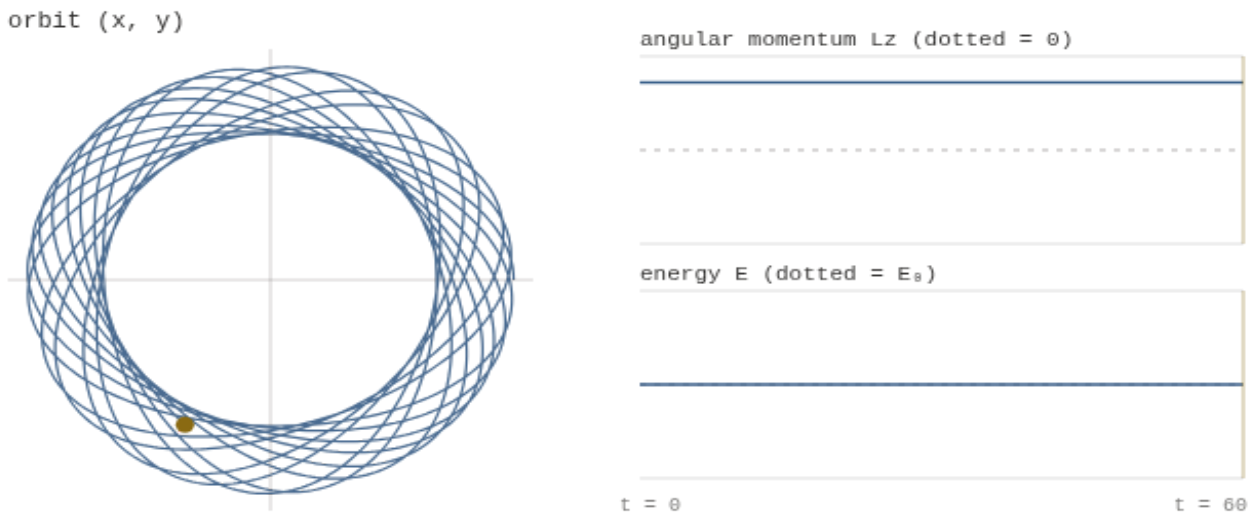
Transformation that does not change the action	K	Conserved quantity Q
Translation $\boldsymbol{x} \mapsto \boldsymbol{x} + s\boldsymbol{e}$	$\boldsymbol{e}$	Component of momentum $m\dot{\boldsymbol{x}} \cdot \boldsymbol{e}$
Rotation in the plane $\boldsymbol{x} \mapsto R(s)\boldsymbol{x}$	$(-y, x)$	Angular momentum $L_z = m(x\dot{y} - y\dot{x})$
Translation in time ($\partial L/\partial t = 0$)	—	Energy $E = \sum_i \dot{q}_i\, \partial L/\partial \dot{q}_i - L$

The third row follows by differentiating directly, from $dE/dt = -\partial L/\partial t$ ($E = \frac{1}{2} m \dot{q}^2 + V$).

Checking with a figure: breaking the symmetry removes the conserved quantity

Give a particle in the plane ($m = 1$) the potential $V = \frac{1}{2} r^2 + \frac{1}{4} r^4 (1 + \delta \cos 4\theta)$ in polar coordinates. If $\delta = 0$ it is rotationally symmetric, and L_z is conserved. With $\delta > 0$ the rotations that leave V unchanged are only multiples of 90°, and the symmetry drops to a discrete one. There are no infinitesimally small transformations for the theorem to use, so the theorem says nothing, and in fact L_z is not conserved. V does not depend on time, so E is conserved.

Figure 2 Moving δ makes the angular momentum Lz start to fluctuate, while the energy E hardly fluctuates (about 10^{-5})δ = 0.00range of Lz 1.4e−14, max|E−E₀| 2.0e−5, t = 60.0



Computed Initial position $(1, 0)$, velocity $(0, 0.9)$. The orbit is computed by the browser with the velocity Verlet method ($h = 0.01$), and Table 3 gives the values from running the same procedure on this machine (1.4e−14 means 1.4×10^{-14}). The two bands on the right have the same vertical scale.

Table 3 Changes of angular momentum and energy ($0 \le t \le 60$)

δ	h	Minimum of L_z	Range of L_z	$\max\lvert E - E_0 \rvert$
0	0.01	0.9000	1.4e−14	2.0e−5
0	0.005	0.9000	1.6e−14	5.0e−6
0.1	0.01	0.8581	0.080	2.8e−5
0.3	0.01	−1.0152	2.03	4.3e−5
0.5	0.01	−1.0938	2.19	5.8e−5

For $\delta = 0.3$ and 0.5 even the sign of L_z changes, and the particle reverses its direction of rotation partway. The deviation of E is about 10^{-5}, and looks flat in the figure.

07

Detour: the discretised action has Noether too

A section you may skip.

Figure 2 is computed step by step with step size h. Here position q_k and velocity v_k are vectors in the plane ($q_k \in \mathbb{R}^2$; the q_k of § 03 were numbers), and they are advanced by the **velocity Verlet method** using the force $F = -\nabla V$.

$$v_{k+1/2} = v_k + \frac{h}{2m} F(q_k), \qquad q_{k+1} = q_k + h\, v_{k+1/2}, \qquad v_{k+1} = v_{k+1/2} + \frac{h}{2m} F(q_{k+1})$$

Even so, in Table 3 at $\delta = 0$, L_z is constant to the level of rounding error, while the energy deviates by 10^{-5}. There is a reason.

Eliminating the velocity gives $m(q_{k+1} - 2q_k + q_{k-1}) = h^2 F(q_k) = -h^2 \nabla V(q_k)$. This is exactly the stationarity condition of the discretised action of § 03 carried over to the plane Known (Hairer–Lubich–Wanner 2003, § 1.6; this form is called the Störmer–Verlet method, and the velocity Verlet method is one way of writing it). If V is rotationally symmetric, the discretised action too is unchanged by a simultaneous rotation, the computation of § 06 goes through with differences, and the following J is constant regardless of k.

$$J = \frac{m}{h}\,(q_k \times q_{k+1})_z$$

Here $(a \times b)_z = a_x b_y - a_y b_x$. It can also be checked directly. Multiplying the stationarity condition by $q_k \times$ on the left, for a central force $F(q_k)$ is parallel to q_k, so the right-hand side vanishes and $q_k \times q_{k+1} = q_{k-1} \times q_k$ remains. J is $2m/h$ times the area of the triangle made by the origin and two neighbouring points, so the area swept in each step is constant.

This J is exactly the $L_z = m(q_k \times v_k)_z$ of Table 3. The first two update formulas give $q_{k+1} = q_k + h\, v_k + \frac{h^2}{2m} F(q_k)$. For a central force $F(q_k)$ is parallel to q_k, so $q_k \times q_{k+1} = h\, q_k \times v_k$, hence $J = L_z$. Theorem 3.5 of the same paper shows in general that quadratic conserved quantities such as angular momentum are preserved exactly by this method Known . And § 1.3, citing Feynman's lectures, notes that the argument by which the "Principia" showed Kepler's second

law has the same form as one way of writing this method (half a step at constant velocity, a kick towards the centre, half a step at constant velocity).

On the other hand, time has been discretised, so it can be shifted only by integer multiples of h. The theorem does not work for discrete symmetries, and energy is not conserved exactly. However, if the formula for the energy (a function of position and velocity) is smooth and the numerical solution stays within some bounded region, the following has been shown for any positive integer N Known (ibid. Theorem 5.1; C and C_N do not depend on t and h).

$$|E_k - E_0| \le Ch^2 + C_N h^N t \qquad (0 \le t = kh \le h^{-N})$$

The second term is at most $C_N h^2$ if $t \le h^{2-N}$. Taking N large, the error stays of order h^2 up to very long times, high powers of h^{-1}. It is not a theorem about infinite time. In Table 3, halving h makes the deviation about 1/4.

"Keep the symmetry you want to keep exactly, as a symmetry of the discretised action" is also the starting point of Wilson's method in lattice gauge theory (5-01). What is kept exactly there is gauge symmetry, and what is broken are the continuous rotations and translations of spacetime (four dimensions turned into imaginary time; 4-02). The 90° rotations of the lattice directions and translations by the lattice spacing remain.

08

Local symmetries give constraints

The s of § 06 was a constant (a **global** symmetry). A symmetry under which the action is unchanged even when transforming by a $s(t)$ that differs from time to time is a **local** symmetry, and it is at the centre of gauge theory (3-01). We look at a toy with two variables, $x(t)$ and $\lambda(t)$.

$$L = \tfrac{1}{2} m(\dot{x} - \lambda)^2, \qquad x \mapsto x + s(t), \quad \lambda \mapsto \lambda + \dot{s}(t)$$

For any $s(t)$, $\dot{x} - \lambda$ does not change. λ appears without time derivatives, so its equation is a **constraint**.

$$(\lambda)\ \ m(\dot{x} - \lambda) = 0, \qquad (x)\ \ \frac{d}{dt}\big[m(\dot{x} - \lambda)\big] = 0$$

The second is the derivative of the first, and only one is independent. x remains as an arbitrary function that the equations do not fix. Noether's **second theorem** (1918) states that when the action is unchanged under transformations containing arbitrary functions, some of the Lagrange equations follow from the rest Known . To be exact, as many identities as there are arbitrary functions hold among the left-hand sides of the Lagrange equations and their derivatives. In this toy the arbitrary function is the one $s(t)$, and the identity is also one: "the left-hand side for (x) is the time derivative of the

left-hand side for (λ)". The $Q = m(\dot{x} - \lambda)$ made by (N) from a constant s is, far from conserved, constrained to be 0.

In electromagnetism the scalar potential ϕ plays the role of λ, the vector potential $\mathbf{A}$ the role of x, and Gauss's law plays the role of the equation for λ (2-03, 4-01). To be exact, in electromagnetism $s(t)$ becomes a function at each point $\chi(t, \mathbf{x})$, $\mathbf{A}$ shifts by $\nabla\chi$ and ϕ by $-\partial\chi/\partial t$ (the formula for gauge transformations in 2-03; the sign of the shift of ϕ is opposite to the λ of the toy). The degrees of freedom the equations do not fix are a redundancy of description, and handling them is one of the big problems of quantising Yang–Mills theory (4-08).

The toy itself has no physical content (the only solution is $\lambda = \dot{x}$, with x arbitrary).

What this chapter can say

Matter	What is said	What is not said
The Euler–Lagrange equations	Known Derived for smooth paths with fixed ends	Variations with moving ends, the general theory of constraints
Stationary and minimum	Computed For the harmonic oscillator, a minimum if $\omega T < \pi$, a saddle point beyond (Tables 1 and 2)	The general theory of conjugate points (Jacobi's condition, the proof of the Morse index theorem)
Noether's first theorem	Known Proved for finitely many degrees of freedom when L is exactly invariant	The case where L changes by a total derivative, the form in field theory (2-02)
Numerical observation	Computed In one example ($t \leq 60$), the fluctuation and conservation of L_z and the error $\propto h^2$ of E were observed (Table 3)	Not a general claim. Long-time estimates rely on theorems in the literature
Noether's second theorem	The redundancy of equations and constraints seen in one toy	The general proof and the form in field theory (3-01)
The origin of the action principle	—	Physics 2-07

Sources and reproduction

Item	Kind	Source / tools
The Euler–Lagrange equation, generalised coordinates, (N)	Known Textbook	L. D. Landau, E. M. Lifshitz, *Mechanics*, 3rd ed., Pergamon Press, 1976. Bibliographic data only
Noether's two theorems	Known Primary source	E. Noether, Invariante Variationsprobleme, Nachr. Ges. Wiss. Göttingen, Math.-phys. Kl. (1918) 235–257. English translation M. A. Tavel, Invariant variation problems, Transport Theory and Statistical Physics 1(3) (1971) 186–207, doi:10.1080/00411457108231446, arXiv:physics/0503066 (Theorems I and II of § 1 checked)
Second variation, conjugate points	Known Textbook	D. Liberzon, *Calculus of Variations and Optimal Control Theory*, Princeton University Press, 2012, doi:10.1515/9781400842643 (Jacobi's condition and conjugate points checked in § 2.6.2 of the pre-publication version). I. M. Gelfand, S. V. Fomin, *Calculus of Variations*, 1963. Bibliographic data only
The Morse index theorem	Known Textbook	J. Milnor, *Morse Theory*, Annals of Mathematics Studies 51, Princeton University Press, 1963 (§ 15 Theorem 15.1 checked; what is stated is the case of geodesics)
On the Störmer–Verlet method of § 07	Known Review	E. Hairer, C. Lubich, G. Wanner, Acta Numerica 12 (2003) 399–450, doi:10.1017/S0962492902000144 (§ 1.3, § 1.6, Theorems 3.5 and 5.1 checked)
Variational integrators	Known Review	J. E. Marsden, M. West, Acta Numerica 10 (2001) 357–514. Bibliographic data only
Tables 1, 2, 3	Computed	python3 + numpy (the record of this book's study). `action_scan.py` (Simpson's rule), `hessian.py` (`eigvalsh`), `noether_verlet.py` (the velocity Verlet method; Figure 2 runs the same procedure in the browser)

Next chapter: 2-02 The idea of a field (replacing q_i by a field $\phi(x)$, the degrees of freedom become infinitely many).

The idea of a field — when the number of beads becomes infinite, the coordinates of the system become a function

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That a field is the limit of a chain of beads as the number of beads goes to infinity.
That a string is a collection of infinitely many independent simple oscillations, and that with a mass term the frequencies have a lower bound, which remains when the box is enlarged.
The price: the value of a field in thermal equilibrium is not determined at a point when space has two or more dimensions.

Prerequisite chapters — 2-01 (the Euler–Lagrange equations, Noether's theorem). Eigenvalues of symmetric matrices (1-01). Fourier series are built in section 04 as far as needed.

Said first — Only classical fields are treated (quantisation is in 2-06). The "lower bound on frequencies" of section 06 is a property of classical waves, and is not itself the mass gap of the problem statement (a gap in the spectrum of quantum theory, 4-11). This chapter shows nothing about the mass gap.

Order of this chapter

1. A chain of beads — N degrees of freedom
2. Normal modes — splitting into independent simple oscillations
3. The continuum limit — the wave equation comes out (Figure 1)
4. Solutions as waves — travelling waves and standing waves (Figure 2)
5. Energy — conservation and distribution over modes
6. The mass term — a lower bound on frequencies (Figure 3)
7. What infinitely many degrees of freedom means — a hot field is not determined at a point (Figure 4)

8. Where this goes in the book — it split because it was linear (Detour: the string of confinement also vibrates by this equation)

01

A chain of beads — a system with N degrees of freedom

The shortest way into fields is to start from finitely many beads. Place N beads at spacing $a = \ell/(N+1)$ on an interval of length ℓ, join neighbours with taut threads, and fix both ends to walls. Let $q_j(t)$ be the sideways displacement of bead j (the walls are $q_0 = q_{N+1} = 0$), $m = \rho a$ the mass of a bead (ρ is the linear density), and σ the tension.

If the difference Δq from a neighbour is small, the thread between them stretches by $\sqrt{a^2 + \Delta q^2} - a \approx \Delta q^2/(2a)$. We let the beads move only sideways, and discard the next term $-\Delta q^4/(8a^3)$. This makes the following equations linear (where linearity matters is section 08). Tension $\times$ stretch is the potential energy, so the Lagrangian (2-01) and the Euler–Lagrange equations take the following form.

$$L = \sum_{j=1}^{N} \frac{m}{2}\dot{q}_j^2 - \sum_{j=0}^{N} \frac{\sigma}{2a}(q_{j+1} - q_j)^2$$

$$m\ddot{q}_j = \frac{\sigma}{a}\left(q_{j+1} - 2q_j + q_{j-1}\right)$$

Only neighbouring beads appear on the right-hand side. This **locality** survives all the way to gauge theory. **Example.** If $N = 1$, it is a simple oscillation with $m\ddot{q} = -(2\sigma/a)q$, and when $\ell = \rho = \sigma = 1$, $\omega = \sqrt{8} \approx 2.828$.

02

Normal modes — untangling the oscillations

Writing everything together as $\ddot{\mathbf{q}} = -K\mathbf{q}$, K is the symmetric matrix with diagonal entries $2\sigma/(ma)$, $-\sigma/(ma)$ on either side, and 0 elsewhere. Being symmetric, it has N orthogonal eigenvectors, and its eigenvalues are real (1-01). The eigenvalues are not negative either: $\mathbf{q} \cdot K\mathbf{q}$ is $2/m$ times the potential energy, a sum of squares. For an eigenvector $\mathbf{v}$, $\mathbf{v} \cdot K\mathbf{v} = \lambda|\mathbf{v}|^2$, so $\lambda \geq 0$. Both ends stay at the walls, so the potential energy is 0 only when $\mathbf{q} = 0$, and $\lambda = 0$ does not occur either. So we write the eigenvalues as ω_n^2 ($\omega_n > 0$).

Taking the eigenvectors as new axes, each component obeys $\ddot{Q}_n = -\omega_n^2 Q_n$. **$N$ tangled beads split into N mutually unrelated simple oscillations**. These are the normal modes; each way of swinging is called a "mode", and n the number of the mode. For this matrix a closed formula can be written for $n = 1, \dots, N$.

$$v_n(j) = \sin \frac{jn\pi}{N+1}$$

$$\omega_n = 2\sqrt{\frac{\sigma}{ma}}\, \sin \frac{n\pi}{2(N+1)}$$

It can be checked by substitution using $\sin(A+B) + \sin(A-B) = 2\sin A \cos B$.
Example. If $\ell = \rho = \sigma = 1$ then $\omega_n = 2(N+1)\sin \frac{n\pi}{2(N+1)}$, and when n is much smaller than N, $\omega_n \approx n\pi$: the frequencies of the string in the next section.

03

The continuum limit — the field and the wave equation

Increase the number of beads and bring the spacing a to 0. ℓ, ρ, σ are fixed, and the mass of a bead $m = \rho a$ goes to 0. **What to fix when taking the limit** returns as a question of the same form in lattice gauge theory (5-09). Reading the number j as the position $x = ja$ and writing $q_j(t) = u(t, x)$, we get $(q_{j+1} - q_j)/a \to \partial_x u$, $\sum_j a(\cdots) \to \int dx$, $(q_{j+1} - 2q_j + q_{j-1})/a^2 \to \partial_x^2 u$, and

$$L = \int_0^\ell \mathcal{L}\, dx, \quad \mathcal{L} = \frac{\rho}{2}(\partial_t u)^2 - \frac{\sigma}{2}(\partial_x u)^2$$

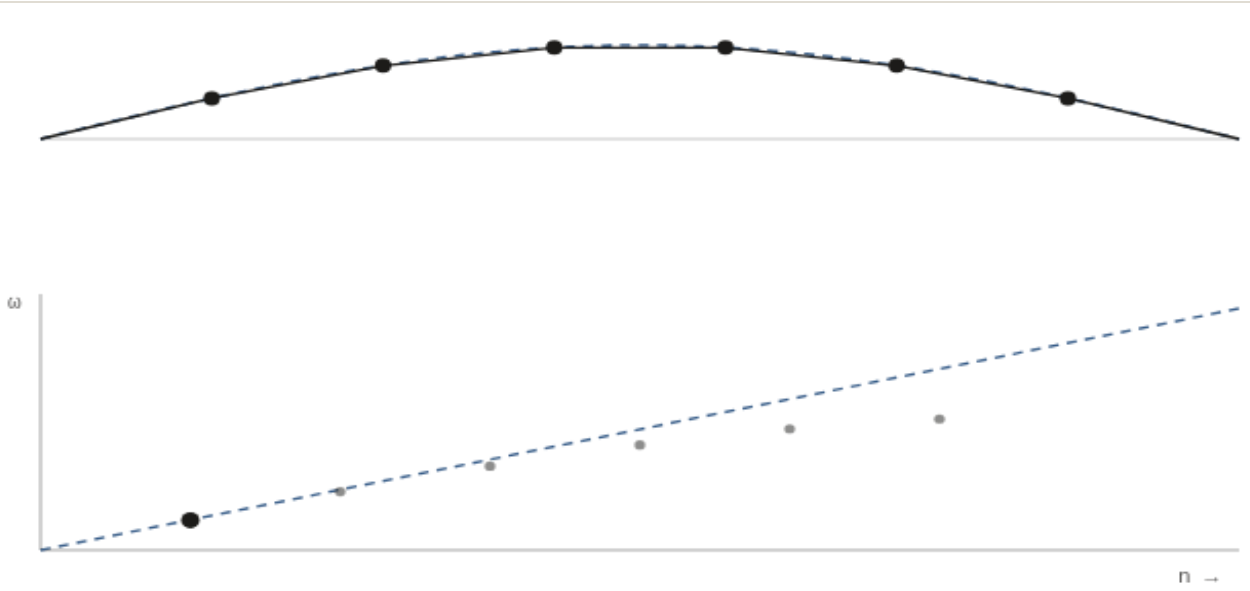
$$\partial_t^2 u = c^2 \partial_x^2 u, \quad c = \sqrt{\sigma/\rho}$$

$\mathcal{L}$ is called the **Lagrangian density**, and the formula below the **wave equation**. The same formula also comes from the Euler–Lagrange equation for fields $\partial_t \frac{\partial \mathcal{L}}{\partial(\partial_t u)} + \partial_x \frac{\partial \mathcal{L}}{\partial(\partial_x u)} - \frac{\partial \mathcal{L}}{\partial u} = 0$, and this form is used both for the electromagnetic field (2-03) and for the Yang–Mills field (3-03).

A **field** is a system in which the index j has been replaced by a continuous x. The "coordinates" of the system are not a set of numbers but a single function $u(t, \cdot)$. The normal modes of the string are $\sin(n\pi x/\ell)\cos\omega_n t$ with frequencies $\omega_n = n\pi c/\ell$ ($n = 1$, $2, \dots$), with no upper limit.

Figure 1 Changing N and n changes the vibration of the chain (solid line and dots) and of the string (dashed), and the mismatch of their frequencies. The lower panel shows the frequencies of all modes (dashed: string ω = nπ,

dots: chain). Pressing a number in Table 1 sets that N and n



N = 6, n = 1, ω(chain) = 3.115, ω(string) = 3.142, ratio = 0.9916

Computed The numbers in the figure are computed on the spot from the closed formulas in the text.

Computed **Table 1** Ratio $\omega_n/(n\pi)$ of the frequencies of the chain and the string ($\ell = \rho = \sigma = 1$). Values found numerically as eigenvalues of the matrix K; the relative difference from the closed formula was at most 3×10^{-12}.

N	n = 1	n = N/2	n = N
4	0.9836	0.9355	0.7568
16	0.9986	0.9114	0.6735
64	0.99990	0.9033	0.6464
256	0.999994	0.9011	0.6391

With n fixed the ratio approaches 1, but the column $n = N/2$ only approaches $\sin(\pi/4)/(\pi/4) \approx 0.900$, and the column $n = N$ only $2/\pi \approx 0.637$. The continuum limit is an approximation for **waves whose wavelength is much longer than the spacing**.

04

Solutions as waves — travelling waves and standing waves

The solutions can be read in two ways. **As travelling waves**: for any twice differentiable f, g, $u = f(x - ct) + g(x + ct)$ is a solution. f travels to the right and g to the left, keeping their shape, at speed c. With initial shape u_0 and initial velocity 0, $u = \frac{1}{2}[u_0(x - ct) + u_0(x + ct)]$, and the hump splits in two (with both ends fixed, extend u_0 to an odd function of period 2ℓ). Solutions of this form are said to go back to d'Alembert's paper of 1747.

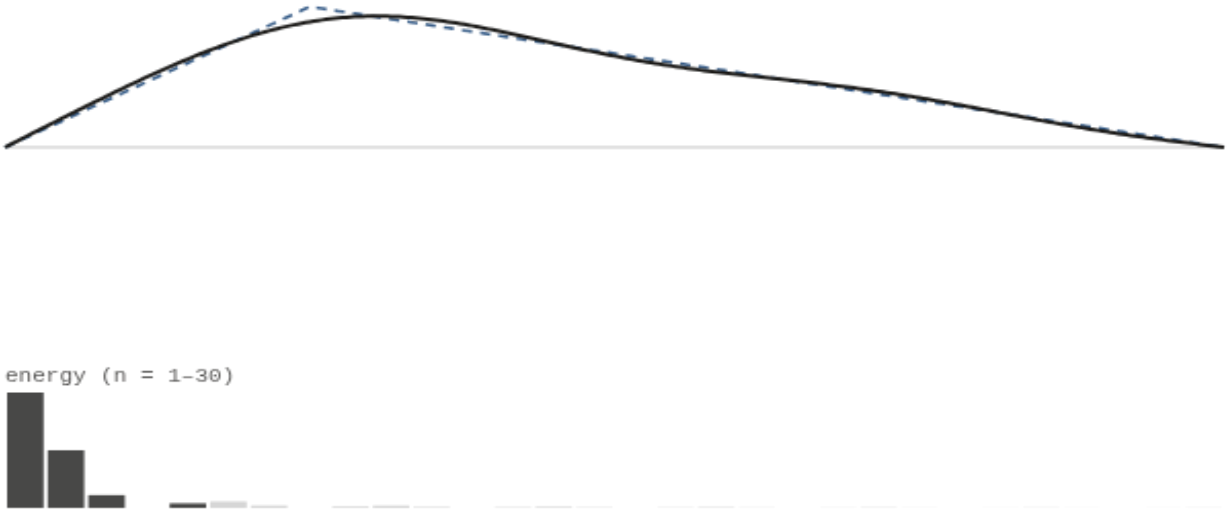
As superpositions of standing waves: the equation is linear, so sums of normal modes are also solutions.

$$u = \sum_{n \geq 1} b_n \sin \frac{n\pi x}{\ell} \cos \omega_n t$$

$$b_n = \frac{2}{\ell} \int_0^{\ell} u_0(x) \sin \frac{n\pi x}{\ell} \, dx$$

The formula for the coefficients comes from the orthogonality of the $\sin(n\pi x/\ell)$ (the integral of the product is 0 for $n \;/= n'$ and $\ell/2$ for $n = n'$). It is a Fourier sine series (1-11). **Example.** For the triangle obtained by pulling the position p to height h and releasing, $b_n = \frac{2h\ell^2}{n^2\pi^2 p(\ell - p)} \sin \frac{n\pi p}{\ell}$. The coefficients decrease only as $1/n^2$, and representing the corner needs modes up to high numbers.

Figure 2 Changing K and p changes how close the sum of standing waves (solid line) is to d'Alembert's solution (dashed), and the distribution of energy (bars). Played, the hump splits in two and turns over at the walls. Pressing a number in Table 2 sets that p and K



K = 5, p/ℓ = 0.25, energy of the first 5: 89.2%

Computed The numbers in the figure are computed on the spot from the partial sums of b_n and d'Alembert's formula ($\ell = c = h = 1$).

05

Energy — conservation, and distribution over modes

The sum E of kinetic and potential energy does not depend on time. Using the wave equation, the time derivative of the integrand becomes $\sigma\, \partial_x(\partial_t u\, \partial_x u)$, and since both ends do not move the boundary terms vanish.

$$E = \int_0^{\ell} \Big[\frac{\rho}{2} (\partial_t u)^2 + \frac{\sigma}{2} (\partial_x u)^2 \Big] dx$$

$$\frac{dE}{dt} = \sigma \big[\partial_t u \, \partial_x u \big]_0^{\ell} = 0$$

This is the field version of Noether's theorem (2-01). Splitting into normal modes, the cross terms vanish by orthogonality, $E = \sum_n E_n$, $E_n = \frac{\rho \ell}{4} (\dot{b}_n^2 + \omega_n^2 b_n^2)$. Each E_n is conserved separately. **A string is a collection of infinitely many mutually unrelated simple oscillations**.

Example. For a plucked string ($s = p/\ell$), the share of mode n is $f_n = \frac{2 \sin^2(n\pi s)}{n^2 \pi^2 s(1-s)}$.

Computed **Table 2** The fraction of the energy held by the first K modes. We also checked that the sum of all terms is 1 (to 5 digits or more with 2 million terms).

s = p/ℓ	K = 1	K = 3	K = 10	K = 100
0.5	0.8106	0.9006	0.9596	0.9959
0.25	**0.5404**	**0.8706**	**0.9508**	**0.9946**
0.1	0.2150	0.5732	0.8920	0.9887
0.02	0.0408	0.1217	0.3882	0.9489

Plucking in the middle gives $f_1 = 8/\pi^2 \approx 0.81$. Near the end ($s = 0.02$) the first 10 do not reach 40%, and the energy escapes to the short-wavelength side.

06

The mass term — a lower bound on frequencies

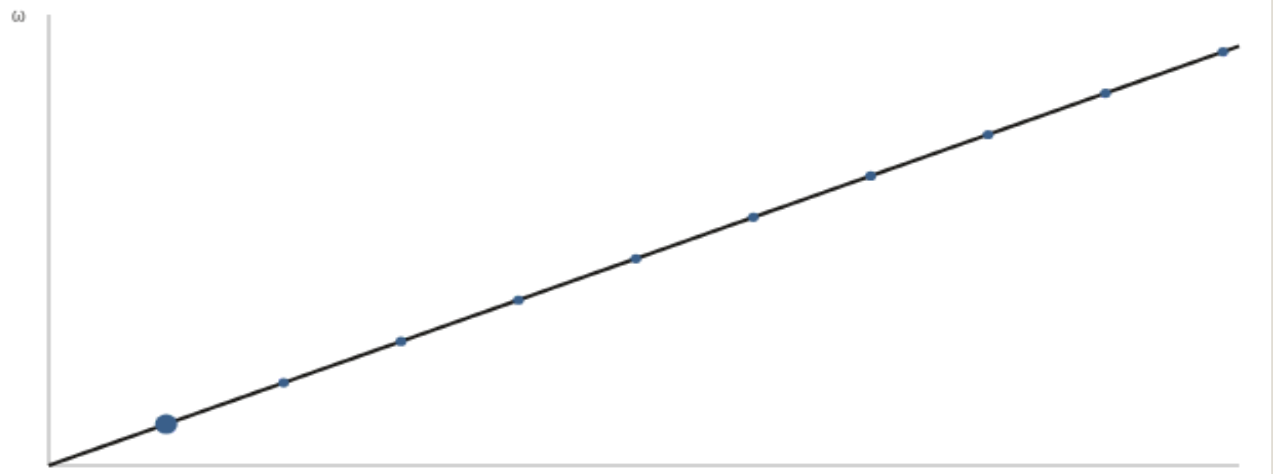
If each bead is also tied to the floor by a spring (constant κ) pulling it back to its original position, $\sum_j \frac{\kappa}{2} q_j^2$ is added to the potential energy. The continuum limit with $\Omega^2 = \kappa/m$ fixed is the Klein–Gordon equation on the left, and substituting a plane wave $\cos(kx - \omega t)$ gives the **dispersion relation** on the right. Here k is the **wave number** (the advance of phase per unit length; for wavelength λ, $k = 2\pi/\lambda$), a different quantity from the mode number n.

$$\partial_t^2 u = c^2 \partial_x^2 u - \Omega^2 u, \qquad \omega^2 = c^2 k^2 + \Omega^2$$

If $\Omega = 0$ then $\omega = c|k|$, and lengthening the wavelength brings the frequency towards 0. If $\Omega > 0$ then always $\omega \geq \Omega$: **the frequencies have a lower bound.** In a box of length ℓ the wave numbers are restricted to $k_n = n\pi/\ell$, and the lowest frequency is $\omega_1 = \sqrt{(c\pi/\ell)^2 + \Omega^2}$. In a finite box it is positive even for $\Omega = 0$, and **what is asked is**

whether the lower bound remains when the box is enlarged. This is why the definition of the mass gap is written in infinite volume (4-11).

Figure 3 Changing Ω and ℓ changes the dispersion relation (curve) and the waves the box allows (dots). The large dot at the left end is the lowest frequency



ℓ = 4.0, Ω = 0.00, lowest frequency ω_1 = 0.7891

Computed The numbers in the figure are computed on the spot from $\omega = \sqrt{k^2 + \Omega^2}$, $k_n = n\pi/\ell$ ($c = 1$).

Computed The lowest frequency ($c = 1$) for $\ell = 1, 10, 100, 1000$ is 3.142, 0.314, 0.0314, 0.00314 at $\Omega = 0$, and 3.297, 1.048, 1.0005, 1.000005 at $\Omega = 1$ (values of the closed formula). For $\ell \leq 100$ it agreed with the smallest eigenvalue of the matrix of a chain with spacing 0.05 to within a relative difference of 0.11% (largest at $\ell = 1$).

The classical lower bound, and the mass gap

Known After quantisation (2-06), the lowest excitation above the vacuum of a free field is $\hbar\Omega$ in infinite volume. Comparing with $E = mc^2$ we read it as $\Omega = mc^2/\hbar$, which is why this term is called the **mass term**. The Yang–Mills action has no mass term (gauge invariance does not allow it, 3-02). Physics Even so, in quantum theory a gap is thought to arise, and numerical computations on the lattice support this (5-08). Showing as mathematics, without adding a mass term, that a gap comes out of the theory itself — that is the content of the problem (0-03).

07

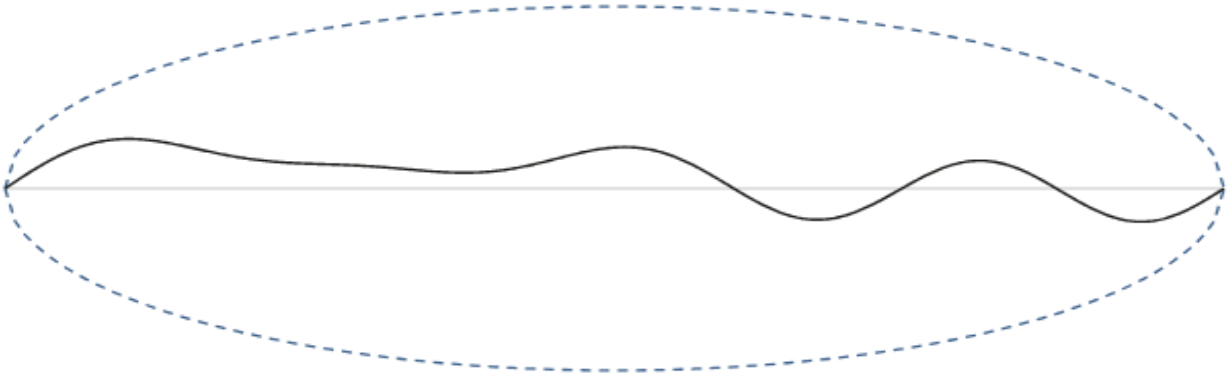
What infinitely many degrees of freedom means

A string has countably infinitely many degrees of freedom. The trouble comes when **all modes are shaken with the same weight**, the typical case being thermal equilibrium. A minimal definition (details in 2-08): in thermal equilibrium at

temperature T, the probability density of a configuration is proportional to $\exp(-V/k_BT)$ (k_B is Boltzmann's constant). If $V = \sum_n \frac{\kappa_n}{2} Q_n^2$, each Q_n follows an independent Gaussian distribution (1-08) with variance k_BT/κ_n, and the mean potential energy is $k_BT/2$ for each (**equipartition**). For the string, $\kappa_n = \sigma n^2\pi^2/(2\ell)$ for $u = \sum_n q_n \sin(n\pi x/\ell)$, and three things happen.

1. **The total energy is infinite**, because $k_BT/2$ is given to each of infinitely many modes.
2. **The displacement at a point is finite.** $\operatorname{Var} u(x) = \sum_n \frac{2\ell k_BT}{\sigma n^2\pi^2} \sin^2\frac{n\pi x}{\ell} = \frac{k_BT}{\sigma} \cdot \frac{x(\ell-x)}{\ell}$. This random curve is called the **Brownian bridge** (Known Sheffield 2007).
3. **The slope is not determined.** $\operatorname{Var} \partial_x u = \frac{2k_BT}{\sigma\ell} \sum_n \cos^2\frac{n\pi x}{\ell}$ diverges. The curve does not break off, yet it is all corners everywhere. To be exact, it means that at no point can the slope be defined as a random variable.

Figure 4 Doubling K again and again makes the sample of the string in thermal equilibrium finer. The overall course does not change, and it stays mostly within the dashed band $\pm 2\sqrt{x(1-x)}$ (twice the standard deviation) (the probability of being outside at each point is about 5%)



K = 8, variance at the centre = 0.2374 k_BTℓ/σ (limit 0.25), variance of the slope at the centre = 8 k_BT/(σℓ) (keeps growing with K)

Computed The figure computes $u = \sum_{n\le K} g_n \sqrt{2}\sin(n\pi x)/(n\pi)$ on the spot (g_n are normal random numbers with a fixed seed, $\ell = k_BT/\sigma = 1$).

Computed Raising the dimension of space d (a membrane, a field in a box), we count the variance at the centre of the box $[0,1]^d$ (in units of k_BT/σ) as a sum over modes with the number in each direction up to K (doubling from $K = 10$ to 160). The contribution of mode $n = (n_1, \dots, n_d)$ is proportional to $1/|n|^2$. The table and figure of the same quantity discretised on a lattice are in Table 3 and Figure 3 of 1-09, which this book takes as authoritative. A figure setting it beside quantities that switch at four is the dimension slider of 1-12 § 10.

$d = 1$ converges to $1/4$ (the value at the centre of the Brownian bridge). $d = 2$ increases by $\ln 2/(2\pi) \approx 0.110$ each time K is doubled (diverging as $\log K$); $d = 3$ doubles, and $d = 4$ quadruples. The number of modes in the shell of radius r is proportional to r^{d-1}, with $1/r^2$ for each, so the sum has the form $\sum r^{d-3}$. The divergence comes from **the short-wavelength side**.

Known Write a field in a box of dimension d as $\varphi(x)$ (the multidimensional version of the u of the string). For $d \geq 2$, the value $\varphi(x)$ of a hot field at a point has no meaning, and only the value $\int \varphi f \, dx$ obtained by multiplying by a smooth weight function f and integrating has meaning as a random variable. In this sense, the field takes values not in functions but in **distributions** (Sheffield 2007, 1-09, 1-11). Writing a quantum field in imaginary time treats time like one more direction of space, and a free field in four-dimensional spacetime has a sum of the same form as the $d = 4$ row (4-02).

The rougher the field, the harder it is to give meaning to products of fields (non-linear terms). To be exact, products of distributions cannot in general be defined (1-11). Divergences (4-04) are their naive appearance. Stating exactly how rough a field can be while non-linear terms can still be handled is the "subcritical / critical" of stochastic quantisation (7-09).

08

Where this goes in the book

Uses ahead: the potentials of the electromagnetic field and their redundancy (2-03), the Lorentz invariance of $\partial_t^2/c^2 - \partial_x^2$ (2-04, the article on spacetime; whether a wave leaves a tail is decided by the dimension of space: same article), the Fock space in which the modes become quantum oscillators (2-06), probability on fields (2-07).

In 3-02 one assumption breaks down. **The splitting into normal modes worked because the equation was linear**. The Yang–Mills field takes values in a Lie algebra at each point and interacts with itself, so the modes are no longer independent. Classical waves have no lower bound either. Physics Even so, in quantum theory a gap is thought to arise.

Detour: the string of confinement also vibrates by this equation

A subsection you may skip.

Physics When a pair of heavy charges is separated by a distance r, the Yang–Mills field between them is thought to become a tube like a string with tension σ (17-03). The tube is regarded as a relativistic string, in which case the linear density is $\rho = \sigma/c^2$ and the speed of transverse waves $\sqrt{\sigma/\rho}$ equals the speed of light c. The transverse vibrations of the tube are taken to obey the

wave equation; summing the zero-point energy of mode $\omega_n = n\pi c/r$ (the lowest energy $\hbar\omega_n/2$ of a quantum oscillator, 2-06) over the $d-2$ directions of vibration (here d is the dimension of spacetime, the same letter as in 17-03; the d of section 07 is the dimension of the space the field spreads over), and formally assigning $-1/12$ to the divergent $1+2+3+\cdots$ (4-12), gives $V(r) = \sigma r - \frac{(d-2)\pi\hbar c}{24 r} + \cdots$. This Lüscher term has been compared, as a prediction of effective string theory, with numerical results on the lattice (Lüscher 1981; review Hari Dass 2023). It is an outline, not a derivation.

What this chapter can say

Content	Status
The classical facts of sections 01–06 (the frequencies of the chain, the continuum limit, the two solutions as waves, the conservation and distribution of energy, the lower bound from the mass term)	Said Known (numbers also checked in Tables 1 and 2 and section 06 Computed)
That the variance at a point of a field in thermal equilibrium is finite for d = 1 and diverges for d ≥ 2	Said Known (the rate of divergence from numbers with K ≤ 160 and counting shells Computed)
Rigorous convergence of the continuum limit	Not said (the limit in this chapter is formal)
The correspondence between the classical lower bound on frequencies and the quantum gap	Said only for free fields Known (2-06)
The string of confinement and the Lüscher term / the Yang–Mills mass gap	The former is only an outline at the level of physics Physics . On the latter this chapter shows nothing

Sources and reproduction

Item	Kind	Source / tools
Tables 1 and 2, the numbers of sections 06 and 07, Figures 1–4	Computed	The tables with python3 and NumPy (eigenvalues, sums of series). The figures with JavaScript in the browser (the formulas in the text and partial sums)
The equation of the string and its general solution	Known	J. le R. d'Alembert, “Recherches sur la courbe que forme une corde tenduë mise en vibration”, Hist. Acad. R. Sci. Belles-Lettres Berlin 3 (for 1747) pp. 214–219. Bibliographic data only (year of publication not checked)
The Brownian bridge, fields becoming distributions	Known	S. Sheffield, “Gaussian free fields for mathematicians”, Probab. Theory Relat. Fields 139 (2007) 521–541, arXiv:math/0312099 (§ 2.5 checked)

Item	Kind	Source / tools
Lattice numbers supporting the gap of quantum Yang–Mills	Physics	Left to the table of sources of 5-08. One example: C. J. Morningstar, M. Peardon, “The glueball spectrum from an anisotropic lattice study”, Phys. Rev. D 60 (1999) 034509, arXiv:hep-lat/9901004 (bibliographic data only)
The Lüscher term	Physics	M. Lüscher, “Symmetry breaking aspects of the roughening transition in gauge theories”, Nucl. Phys. B180 (1981) 317–329 (bibliographic data only). The form of the formula checked in formula (1) of N. D. Hari Dass, “Effective String Theories (EST) of Yang-Mills Flux Tubes”, arXiv:2312.10629 (Chapter 22 of the book Strings to Strings, Springer 2023)

Next chapter: 2-03 Electromagnetism is a gauge theory. To go ahead to quantisation, 2-06.

Electromagnetism is a gauge theory — the freedom in writing the potentials becomes an element of U(1) chosen at each point

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That infinitely many potentials give the same electric and magnetic fields (gauge transformations).
That Gauss's law is a constraint, and that gauge invariance and the conservation of charge come as a pair.
That in quantum mechanics a gauge transformation turns the phase (an element of U(1)) point by point, and the phase factor of a closed path is observable.

Prerequisite chapters — 2-02, section 08 of 2-01 (local symmetries and constraints), 1-01. An introduction to vector analysis and electromagnetism is assumed; the minimum of quantum mechanics is set out in section 05.

Said first — The relativistic form is in 2-04, the non-abelian case in 3-02. The electromagnetic field without interactions has no mass gap (section 04). This chapter shows nothing about the mass gap of Yang–Mills.

Order of this chapter

1. The electric and magnetic fields — Maxwell's equations
2. Potentials — satisfying the source-free Maxwell equations automatically
3. Gauge transformations — the freedom in writing (Figure 1)
4. From the action — constraints and the conservation of charge
5. Quantum phase and U(1) — turning the phase point by point (phase factors)
6. The Aharonov–Bohm effect — the fringes move even with B = 0 (Figure 2, Table 1)
7. A particle on a ring — the levels are periodic functions of the flux
8. A lattice ring — putting elements of U(1) on the edges (Figure 3, Table 2)

9. From U(1) to SU(N) — towards the gauge principle

01

The electric and magnetic fields — four equations

The protagonists of electromagnetism are two fields carrying a three-dimensional vector at each point, the **electric field** $\mathbf{E}(t,\mathbf{x})$ and the **magnetic field** $\mathbf{B}(t,\mathbf{x})$. A particle of charge q and velocity $\mathbf{v}$ feels the **Lorentz force** $\mathbf{F}=q(\mathbf{E}+\mathbf{v}\times\mathbf{B})$, and the fields obey **Maxwell's equations** (SI units; ρ is the charge density, $\mathbf{J}$ the current density, $c^2=1/\varepsilon_0\mu_0$).

$$\nabla\cdot\mathbf{E}=\frac{\rho}{\varepsilon_0},\qquad \nabla\times\mathbf{B}-\frac{1}{c^2}\frac{\partial\mathbf{E}}{\partial t}=\mu_0\mathbf{J}\ \text{(M1)}$$

$$\nabla\cdot\mathbf{B}=0,\qquad \nabla\times\mathbf{E}+\frac{\partial\mathbf{B}}{\partial t}=\mathbf{0}\ \text{(M2)}$$

The two Maxwell equations with sources (**Gauss's law** and the **Ampère–Maxwell law**) say how charges and currents make the fields; the two source-free Maxwell equations (the absence of magnetic charge and **Faraday's law of induction**) are constraints on the form of the fields. **Example.** Where there are no sources, the four Maxwell equations give $\partial_t^2\mathbf{E}=c^2\nabla^2\mathbf{E}$ (the three-dimensional version of the wave equation of 2-02), and this wave is light. The dispersion $\omega=c|\mathbf{k}|$ has no lower bound.

Since the source-free Maxwell equations contain no sources, the form of $\mathbf{E},\mathbf{B}$ can be narrowed down before solving the equations. That is the next section.

02

Potentials — satisfying the two source-free equations automatically

With the identities $\nabla\cdot(\nabla\times\mathbf{A})=0$ and $\nabla\times\nabla f=\mathbf{0}$, the source-free Maxwell equations can be satisfied automatically. Writing $\mathbf{B}$ as a curl makes the first equation hold; the second becomes $\nabla\times(\mathbf{E}+\partial_t\mathbf{A})=\mathbf{0}$, so we write $\mathbf{E}+\partial_t\mathbf{A}$ as a gradient.

$$\mathbf{B}=\nabla\times\mathbf{A},\qquad \mathbf{E}=-\nabla\phi-\frac{\partial\mathbf{A}}{\partial t}\ \text{(P)}$$

We call $\mathbf{A}$ the **vector potential** and ϕ the **scalar potential**. On the whole of $\mathbb{R}^3$, if $\nabla\cdot\mathbf{B}=0$ then $\mathbf{B}$ can always be written as a curl Known (1-06).

Example 1. For a uniform $\mathbf{B} = B_0 \mathbf{e}_z$, both $\mathbf{A}_1 = \frac{B_0}{2}(-y, x, 0)$ and $\mathbf{A}_2 = B_0(0, x, 0)$ can be used. **Example 2.** A long solenoid of radius a, with a uniform $B_0\mathbf{e}_z$ inside, no magnetic field outside, and flux $\Phi = \pi a^2 B_0$. With the distance from the axis r and the unit vector $\mathbf{e}_\theta$ pointing around the axis,

$$\mathbf{A} = \frac{B_0 r}{2}\,\mathbf{e}_\theta \quad (r < a), \qquad \mathbf{A} = \frac{\Phi}{2\pi r}\,\mathbf{e}_\theta \quad (r > a)$$

Outside, $\nabla \times \mathbf{A} = \frac{1}{r}\partial_r(rA_\theta)\mathbf{e}_z = \mathbf{0}$, and yet $\mathbf{A} \neq \mathbf{0}$. The line integral once round a circle of radius r is Φ, which by Stokes's theorem is the flux passing through it. Even for a path that runs only where the curl is 0, the value of one circuit is not 0. This is because the path encloses a hole (the solenoid), and section 06 is where quantum mechanics reads off this value.

03

Gauge transformations — the freedom in writing that gives the same fields

The potential representation does not determine $(\phi, \mathbf{A})$ uniquely from $\mathbf{E}, \mathbf{B}$. With a smooth function $\chi(t, \mathbf{x})$,

$$\mathbf{A} \mapsto \mathbf{A} + \nabla\chi, \qquad \phi \mapsto \phi - \frac{\partial \chi}{\partial t} \quad \text{(G)}$$

this replacement leaves $\mathbf{E}, \mathbf{B}$ unchanged ($\nabla \times \nabla\chi = \mathbf{0}$, $\nabla\partial_t\chi = \partial_t\nabla\chi$). We call this a **gauge transformation**, and χ the **gauge function**. **Example.** The two of Example 1 are carried into each other by $\chi = B_0 xy/2$ ($\mathbf{A}_2 - \mathbf{A}_1 = \nabla\chi$).

Reducing the freedom by a condition (such as $\nabla \cdot \mathbf{A} = 0$) is called **gauge fixing** (the quantum version is in 4-08).

Seen on a lattice

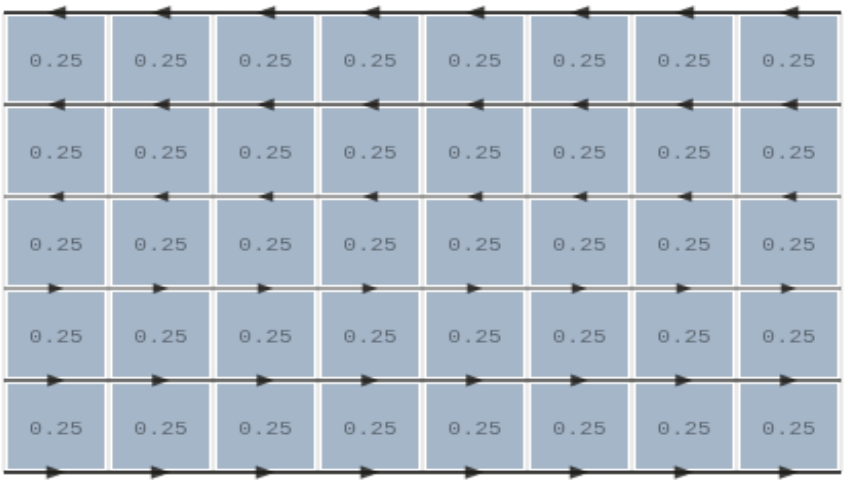
Cut the plane into a square lattice and hold $\mathbf{A}$ as a number $A_{ij} = \int_i^j \mathbf{A} \cdot d\mathbf{l}$ on each edge ($A_{ji} = -A_{ij}$). Label the corners of a square face p 1 to 4 anticlockwise; by Stokes's theorem the flux B_p through the face is the sum over the four edges, and the gauge function becomes the values χ_i at the lattice points.

$$B_p = A_{12} + A_{23} + A_{34} + A_{41}, \qquad A_{ij} \mapsto A_{ij} + \chi_j - \chi_i$$

The χ of each corner appears once in $+$ and once in $-$ and cancels, so B_p does not change. **Example.** The sum of the edge values $(0.3,\ 0.5,\ -0.1,\ -0.2)$ is 0.5. Transforming by $\chi = (1, 0, 2, -1)$ gives $(-0.7,\ 2.5,\ -3.1,\ 1.8)$, but the sum stays 0.5.

Putting group elements on the edges gives the starting point of Wilson's lattice gauge theory (5-01).

Figure 1 Changing the size s of the gauge function, or pressing Play, changes the edge values (line thickness and arrows), but not the face values (shading and numbers)



s = 0.00, max. change on edges 0.000, max. change on faces 0.0e+0

Computed The numbers in the figure are computed on the spot from the sums over four edges and the addition of $\chi_j - \chi_i$.

"Flux through a single face only" is the lattice version of the solenoid of section 06. The row of non-zero edges changes shape under gauge transformations but does not disappear (it corresponds to the cut where the angle θ of section 06 jumps). Under "No magnetic field" only the edges fluctuate; this is a **pure gauge**.

04

From the action — constraints and the conservation of charge

Maxwell's equations too are the stationarity conditions of an action (2-01, 2-02). The unknowns are $(\phi, \mathbf{A})$, and $\mathbf{E}, \mathbf{B}$ are understood as written by the potential representation.

$$\mathcal{L} = \frac{\varepsilon_0}{2}|\mathbf{E}|^2 - \frac{1}{2\mu_0}|\mathbf{B}|^2 - \rho\,\phi + \mathbf{J}\cdot\mathbf{A}, \qquad S = \int \mathcal{L}\, dt\, d^3x \ \text{(L)}$$

The source-free Maxwell equations hold automatically from the potential representation, and the Maxwell equations with sources come out as the Euler–Lagrange equations Known . $\mathcal{L}$ contains $\nabla\phi$ but not $\partial_t\phi$, so the $\varepsilon_0\nabla\cdot\mathbf{E} = \rho$ obtained by varying ϕ is a condition at equal times containing no second time derivative: a **constraint**. It corresponds to the equation for λ in the toy of 2-01. In this way, of the four components of $(\phi, \mathbf{A})$, one disappears through the formula for gauge transformations and one through this constraint, and the remaining $4 - 1 - 1 = 2$ is the number of polarisations of light Known (the quantum version is in 4-01).

Under the formula for gauge transformations the first two terms do not change, and the last two increase by just $\rho\,\partial_t\chi + \mathbf{J}\cdot\nabla\chi$. Integrating by parts ($\chi$ is 0 far away),

$$\delta S = -\int \chi\left(\frac{\partial\rho}{\partial t} + \nabla\cdot\mathbf{J}\right) dt\, d^3x$$

That the action does not change for any χ is the same as the **conservation of charge** $\partial_t \rho + \nabla\cdot\mathbf{J} = 0$. Seen from the side of the equations, the two Maxwell equations with sources are not independent. Combining the time derivative of Gauss's law with the divergence of the Ampère–Maxwell law, the field terms vanish identically by $\nabla\cdot(\nabla\times\mathbf{B}) = 0$, and only $\partial_t\rho + \nabla\cdot\mathbf{J} = 0$ is left. If charge is conserved, part of the equations follows from the rest. This is the form of the second theorem of 2-01.

Example: a mass term cannot be put in. Following 2-02, adding $-\frac{\varepsilon_0\Omega^2}{2}|\mathbf{A}|^2$ makes the dispersion of transverse waves $\omega^2 = c^2|\mathbf{k}|^2 + \Omega^2$, with a lower bound Ω. But this term changes under the formula for gauge transformations. Under gauge invariance, a mass term cannot be added directly to the action. For the electromagnetic field without interactions no lower bound arises, the mass of the photon is 0 in quantum theory too, and the free electromagnetic field has no mass gap Known .

Whether mass can arise from interactions while keeping gauge invariance is a separate problem. The Meissner effect in a superconductor and the Higgs mechanism (3-07) are examples Known . For Yang–Mills, the question is whether this happens without adding another field (section 09, 4-11).

05

Quantum phase and U(1)

A classical particle feels only $\mathbf{E}, \mathbf{B}$, but in quantum mechanics the potentials enter the formulas. We use three things (2-05). (i) A state is represented by a complex-valued **wave function** $\psi(t, \mathbf{x})$, and $|\psi|^2$ is the probability density of position. (ii) For a constant α, ψ and $e^{i\alpha}\psi$ are the same state. (iii) A particle of mass m and charge q obeys the following Schrödinger equation ($\hbar = h/2\pi$).

$$i\hbar\frac{\partial\psi}{\partial t} = \frac{1}{2m}\left(-i\hbar\nabla - q\mathbf{A}\right)^2\psi + q\phi\,\psi \quad \text{(S)}$$

(S) seems to break under the formula for gauge transformations, but if ψ is also changed as

$$\psi \mapsto e^{iq\chi/\hbar}\,\psi \quad \text{(G')}$$

then (S) holds as it is. The key is the **covariant derivative** $D = \nabla - \frac{iq}{\hbar}\mathbf{A}$: the $\frac{iq}{\hbar}(\nabla\chi)\psi$ coming from the derivative of the product cancels the change in $\mathbf{A}$.

$$\Big(\nabla - \frac{iq}{\hbar}(\mathbf{A} + \nabla\chi)\Big)\big(e^{iq\chi/\hbar}\psi\big) = e^{iq\chi/\hbar}\Big(\nabla - \frac{iq}{\hbar}\mathbf{A}\Big)\psi$$

The time side works in the same way. Fock (1926) noticed this invariance, London (1927) connected it with the theory of scale (Eich) that Weyl had set up earlier, and Weyl (1928–29) raised it as a principle under the name of gauge invariance, it is said (Jackson–Okun 2001).

$e^{iq\chi/\hbar}$ is a complex number of absolute value 1, and the set of all of them forms a group under multiplication. This is **U(1)** (the unit circle; 1-02). **A gauge transformation is the choice of one element of U(1) at each point.** If the element is the same at every point, it is (ii) itself; if it differs from point to point, $\mathbf{A}$ makes things consistent (the argument in the opposite direction is in section 09).

The phase factor along a path

If the reference is chosen point by point, how are phases at distant points compared? For a path C from $\mathbf{x}$ to $\mathbf{y}$,

$$W(C) = \exp\Big(\frac{iq}{\hbar}\int_C \mathbf{A}\cdot d\mathbf{l}\Big), \qquad W(C) \mapsto e^{iq\chi(\mathbf{y})/\hbar}\, W(C)\, e^{-iq\chi(\mathbf{x})/\hbar}$$

Since it cancels the replacements at both ends, $W(C) \in U(1)$ is a tool that carries the phase from $\mathbf{x}$ to $\mathbf{y}$ (**parallel transport**, 1-07). For a closed path it is gauge invariant, and with the enclosed flux Φ_C, $W(C) = e^{iq\Phi C/\hbar}$. On a lattice it is the product of the $e^{iqA_{ij}/\hbar}$ of the edges, and the addition of section 03 becomes multiplication in the group.

06

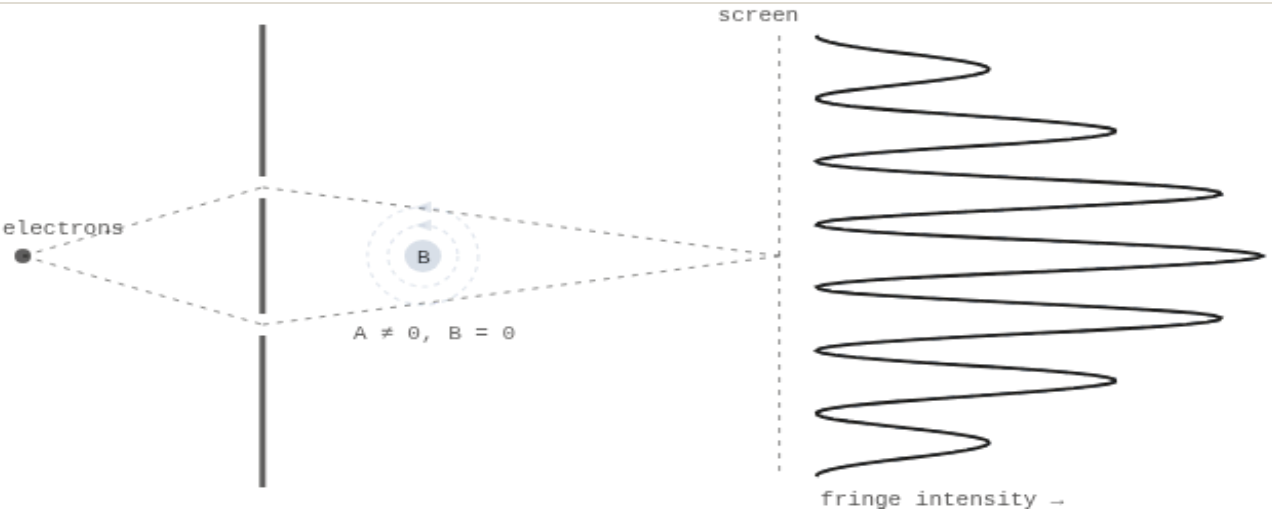
The Aharonov–Bohm effect — the fringes move where there is no magnetic field

Consider a double slit in which electrons pass on either side of the solenoid of Example 2. Where the electrons pass, $\mathbf{B} = \mathbf{0}$, and classically nothing changes. In quantum mechanics the contributions of the path C_1 through slit 1 and the path C_2 through slit 2 are added, and the phase difference makes the fringes. Joining C_1 with the reversed C_2 goes once round the solenoid, so the phase difference due to $\mathbf{A}$ is

$$\Delta = \frac{q}{\hbar}\oint \mathbf{A}\cdot d\mathbf{l} = \frac{q\Phi}{\hbar} = \pm 2\pi\,\frac{\Phi}{\Phi_0}, \qquad \Phi_0 = \frac{h}{|q|}$$

and the fringes shift by Φ/Φ_0 of the fringe spacing. If Φ is an integer multiple of Φ_0, the result cannot be told apart from the original. The prediction is Aharonov–Bohm (1959). Ehrenberg–Siday (1949) had reached the same conclusion earlier, and a clue from 1939 has also been pointed out Known (Hiley 2013).

Figure 2 Changing the flux Φ/Φ_0, or pressing Play, shifts the fringes (solid line on the right) away from the fringes for $\Phi = 0$ (dashed), and they coincide each time Φ/Φ_0 increases by 1. B = 0 on the paths of the electrons



Φ/Φ₀ = 0.00, phase difference = 0.00π, fringe shift = fringe spacing × 0.00

Computed A schematic in which the fringes $I(y) \propto \operatorname{sinc}^2(0.6y)\cos^2(3.5\pi y + \pi\Phi/\Phi_0)$ ($\operatorname{sinc} u = \sin(\pi u)/(\pi u)$) are computed on the spot (the number and width of the fringes are chosen for legibility).

Why it cannot be removed by a gauge transformation. Outside, $\mathbf{A} = \nabla(\Phi\theta/2\pi)$ (θ is the angle around the axis), so it looks removable by $\chi = -\Phi\theta/2\pi$, but θ increases by 2π in one circuit and is not determined as a single function. What needs to be single-valued is the phase $e^{iq\chi/\hbar} = e^{-iq\Phi\theta/2\pi\hbar}$, and that holds only when Φ/Φ_0 is an integer; only then does $\mathbf{A}$ vanish and the effect disappear. The period Φ_0 comes from here.

Computed **Table 1** Numerical integration of $\oint \mathbf{A}\cdot d\mathbf{l}$ around a solenoid of radius 1 with $\Phi = 1$. The right two columns are the $\mathbf{A}$ of Example 2 and the $\mathbf{A}'$ obtained by transforming with $\chi = \sin x\cos 2y + x^2y/3$. The difference from enclosed flux × winding number is within 4×10^{-10}.

Closed path	$\lvert\mathbf{B}\rvert$ on the path	$\oint \mathbf{A}\cdot d\mathbf{l}$	$\oint \mathbf{A}'\cdot d\mathbf{l}$
circle radius 2, centre (0, 0)	0	1.000000	1.000000
square side 5, centre (0.7, −0.4)	0	1.000000	1.000000
circle radius 1, centre (3, 0) (not enclosing)	0	0.000000	0.000000
circle radius 2, twice round	0	2.000000	2.000000
circle radius 0.5 (inside the solenoid)	0.3183	0.250000	0.250000

Computed For electrons $\Phi_0 = h/e = 4.1357 \times 10^{-15}$ Wb; the magnetic field that sends this through a circle of diameter 1 μm is 5.27 mT. Known The first experimental test is said to be Chambers (1960), and Tonomura et al. (1986) measured it with the magnetic field confined in a superconductor.

07

A particle on a ring — the levels are periodic functions of the flux

The shift of the fringes also appears in the levels. Consider a particle of charge $q > 0$ confined to a ring of radius R, with a flux Φ through the centre of the ring. On the ring $\mathbf{B} = \mathbf{0}$, $A_\theta = \Phi/2\pi R$, and ∇ reduces to the $\frac{1}{R}\partial_\theta$ along the ring. Putting a **stationary state** $\psi = e^{-iEt/\hbar} e^{in\theta}$ (a state in which $|\psi|^2$ does not depend on time; since it returns after one circuit, n is an integer) into (S), the left side becomes $E\psi$ and the right side $\frac{1}{2mR^2}(\hbar n - q\Phi/2\pi)^2\psi$, and the energy is

$$E_n = \frac{\hbar^2}{2mR^2}\left(n - \frac{\Phi}{\Phi_0}\right)^2, \qquad n \in \mathbb{Z}$$

For $\Phi \mapsto \Phi + \Phi_0$ the labels are merely renumbered as $n \mapsto n+1$, so **the set of levels is a periodic function with period** Φ_0, and at $\Phi = \Phi_0/2$ the lowest two coincide. There is no magnetic field on the ring, and yet the levels know the flux. Byers–Yang (1961) is said to have discussed from this periodicity the quantisation of flux in a superconducting cylinder (period $h/2e$) Known .

From an examination of the AB effect, Wu–Yang (1975) summarised that the field strength describes electromagnetism too little, the phase $\frac{q}{\hbar}\oint \mathbf{A}\cdot d\mathbf{l}$ describes it too much, and the phase factor $e^{\frac{iq}{\hbar}\oint \mathbf{A}\cdot d\mathbf{l}}$ is what describes it neither more nor less Known . These levels are one example.

08

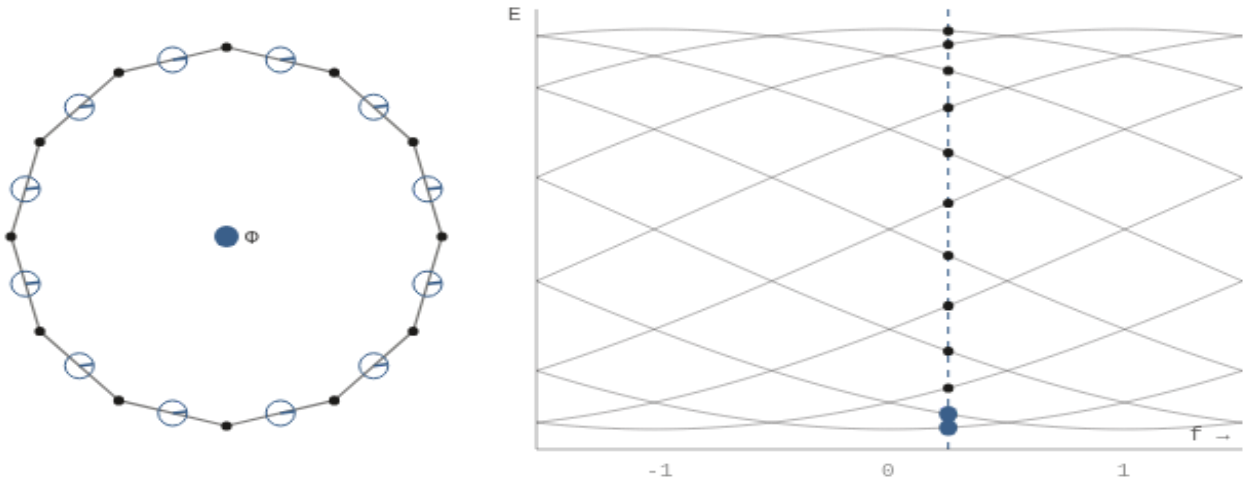
A lattice ring — putting elements of U(1) on the edges

On a ring of N points, consider the Hermitian matrix $H_{j+1,j} = -t\,e^{i\theta_j}$, $H_{j,j+1} = H_{j+1,j}$ (indices modulo N), in which the amplitude $-t$ for moving to a neighbour ($t > 0$) is multiplied by an element of U(1) for each edge. A gauge transformation is $\theta_j \mapsto \theta_j + \chi_{j+1} - \chi_j$, and the only thing that does not change is the product once round, $\prod_j e^{i\theta_j} = e^{2\pi i f}$ ($f = \Phi/\Phi_0$). For a uniform $\theta_j = 2\pi f/N$ the plane waves are the eigenvectors, and

$$E_n = -2t\cos\frac{2\pi(n-f)}{N}, \qquad n = 0, \dots, N-1$$

For $|n-f| \ll N$ this is $E_n \approx -2t + t(2\pi/N)^2(n-f)^2$, the same form as the continuous ring of section 07.

Figure 3 Changing f = Φ/Φ_0 and the number of points N (or pressing a row of Table 2) moves the levels (dots on the right). Pressing "Gauge" changes how the edge phases (the clocks on the left) are distributed, but not the levels



N = 12, f = 0.25, E_0 = -1.98289, E_1 = -1.84776, $E_1 - E_0$ = 0.13513, sum of phases once round ÷ 2π = 0.25 (mod 1)

Computed The levels are computed on the spot from the closed formula ($t = 1$).

Computed **Table 2** The lowest two levels (eigenvalues of the matrix) for $N = 12$, $t = 1$. The rightmost column is the maximum of the differences among three gauges (uniform / gathered on one edge / with a random χ added) and of the difference from the closed formula. Pressing a row moves to Figure 3.

f	E_0	E_1	$E_1 - E_0$	Max. difference
0	-2.000000	-1.732051	0.267949	1.0e-15
0.25	-1.982890	-1.847759	0.135131	6.7e-16
0.5	-1.931852	-1.931852	0.000000	7.8e-16
0.75	-1.982890	-1.847759	0.135131	1.3e-15
1.25	-1.982890	-1.847759	0.135131	8.9e-16

The levels agree among the three gauges within rounding error, and are the same for f, $1-f$, $f+1$. For $N = 192$, $(E+2t)/t(2\pi/N)^2$ approaches the $(n-f)^2$ of the continuous ring, and for the lowest two levels the relative difference was at most 5.0×10^{-5} (E_1 of $f = 0.25$) Computed .

This "gap" and the mass gap

$E_1 - E_0$ is a statement about a finite ring, and it approaches 0 as the ring is made larger. It is a different thing from the mass gap of the problem statement (4-11). What we want you to see is that the levels are determined by the phase factor once round alone. The observables of Yang–Mills on the lattice are also made from the phase factors of closed paths (Wilson loops, 5-01).

09

From U(1) to SU(N) — where this book is going

Yang–Mills theory replaces the element chosen at each point, from U(1) to SU(N) (the $N \times N$ unitary matrices of determinant 1, 1-04) (3-02). Because the result depends on the order of multiplication, a commutator term is added to the field strength, the field itself carries colour charge (the counterpart of electric charge), and the equations become non-linear (3-03). Classical waves have no lower bound, as with light, but in the quantum theory a gap is thought to arise Physics (5-08). The difference from U(1) is in 5-06.

The next bridge is 3-01. This chapter put **A** first and found (G') afterwards. 3-01 goes the other way: it demands that (S) not change even if the phase is turned arbitrarily at each point, and derives **A** as the field that makes things consistent. Making the same demand with SU(N) gives the Yang–Mills field. This is the **gauge principle**, and the parallel transport of section 05 and the edge elements of section 08 are its parts.

What this chapter can say

Content	Status
Sections 01–05 (gauge transformations, constraints, conservation of charge, that a mass term cannot be put directly into the action, (G'), phase factors)	Said Known . The lattice is Figure 1 Computed
Sections 06–08 (the phase difference of the AB effect and the period Φ_0, the periodicity and gauge invariance of the levels on the ring)	Said Known (as a comparison of phases along paths; exact solutions of scattering are not treated). Tables 1 and 2 Computed
That the free electromagnetic field has no mass gap / mechanisms by which mass arises while keeping gauge invariance, and the mass gap of Yang–Mills	The former is known. For the latter, examples are only named, and this chapter shows nothing

Sources and reproduction

Item	Kind	Source / tools
Tables 1 and 2, the numbers in sections 06 and 07, Figures 1–3	Computed	python3, NumPy (midpoint rule for line integrals, eigenvalues of Hermitian matrices). The figures by JavaScript in the browser (the formulas in the text)
Maxwell's equations, the history of gauge invariance (Fock, London, Weyl)	Known	J. D. Jackson, L. B. Okun, Rev. Mod. Phys. 73 (2001) 663, arXiv:hep-ph/0012061 (abstract and section III checked). J. C. Maxwell, Phil. Trans. R. Soc. Lond. 155 (1865) 459, V. Fock, Z. Phys. 39 (1926) 226, F. London, Z. Phys. 42 (1927) 375, H. Weyl, Gruppentheorie und Quantenmechanik (1928), Z. Phys. 56 (1929) 330 (bibliographic data only)
Examples in which mass arises while keeping gauge invariance	Known	P. W. Anderson, Phys. Rev. 130 (1963) 439 (title and abstract checked). For details see 3-07
The prediction, history and experiments of the AB effect	Known	Y. Aharonov, D. Bohm, Phys. Rev. 115 (1959) 485, A. Tonomura et al., Phys. Rev. Lett. 56 (1986) 792 (abstracts checked for both). B. J. Hiley, arXiv:1304.4736 (text checked). W. Ehrenberg, R. E. Siday, Proc. Phys. Soc. B 62 (1949) 8 (bibliographic data only). R. G. Chambers, Phys. Rev. Lett. 5 (1960) 3 (bibliographic data only; title checked; "first" not checked)
Description by phase factors	Known	T. T. Wu, C. N. Yang, Phys. Rev. D 12 (1975) 3845 (conclusion of § II checked)
Periodicity in the flux and superconductivity	Known	N. Byers, C. N. Yang, Phys. Rev. Lett. 7 (1961) 46 (bibliographic data only; the subject checked from the title, content not checked)

Next chapter: 2-04 Special relativity and spacetime. To go on to the gauge principle, 3-01.

Special relativity and spacetime — preserving a quadratic form with one sign different narrows down the form of the action

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That Lorentz transformations are the transformations that preserve the form of the wave equation, and that the sign of the interval divides the relation between two points into three.
That mass is the invariant length of the four-momentum, and that the action of a free particle is the proper time.
That the form of the action of a field is narrowed down, and that the action of pure Yang–Mills in four dimensions has no length scale.

Prerequisite chapters — 2-01 (the action, Euler–Lagrange, Noether) and the product of matrices (1-01). The wave equation and the mass term of 2-02 are also used (the formulas are restated on the spot). The groups of section 04 anticipate 1-02, "why it is called the mass gap" in section 05 anticipates 2-05, the electromagnetic field of section 07 anticipates 2-03, and its Yang–Mills anticipates 1-03 and 1-04. On a first reading you can take only the conclusions and move on.

Said first — Only the classical story in flat spacetime without gravity (quantum theory starts in 2-05). The imaginary time of section 09 is a formal substitution, and its justification is the work of 4-02 and 4-10. Including "the classical action has no scale" of section 08, this chapter shows nothing about the mass gap.

Order of this chapter

1. The wave equation changes form under Galilean transformations — making the speed of light the same for everyone
2. Lorentz transformations — preserving the interval (Figure 1, Table 1)
3. The light cone and the relativity of simultaneity — three cases by sign
4. Four-vectors and the Lorentz group — indices, non-compactness

5. Proper time and four-momentum — mass is an invariant length (why a "mass" gap)
6. The action of a free particle — the straight line maximises proper time (Figure 2, Table 2)
7. Lorentz-invariant actions of fields — scalar, electromagnetic, Yang–Mills
8. Counting dimensions — in four dimensions the coupling is dimensionless (Tables 3 and 4)
9. A bridge to imaginary time — mass becomes a rate of decay

01

The wave equation changes form under Galilean transformations

We write the wave equation built in 2-02 with one dimension of space and wave speed c.

$$\left(\frac{1}{c^2}\partial_t^2 - \partial_x^2\right)u = 0 \quad (1)$$

In Newtonian mechanics the coordinates of an observer moving to the right with speed v are $t' = t,\ x' = x - vt$ (a Galilean transformation). Substituting $\partial_t = \partial_{t'} - v\partial_{x'}$, $\partial_x = \partial_{x'}$, the operator of (1) changes form.

$$\frac{1}{c^2}\partial_{t'}^2 - \frac{2v}{c^2}\partial_{t'}\partial_{x'} - \left(1 - \frac{v^2}{c^2}\right)\partial_{x'}^2$$

Example. Substituting $u = f(x' - st')$ gives $s^2 + 2vs - (c^2 - v^2) = 0$, that is, $s = c - v$ and $-(c + v)$. For $v = c/2$ the waves look like $0.5c$ to the right and $1.5c$ to the left Computed .

For sound this is correct, because the observer at rest relative to the air is special. But the wave equation of light comes from Maxwell's equations (2-03) and needs no medium. Einstein (1905) took as premises that the laws have the same form in every inertial frame and that the speed of light does not depend on the motion of the source Known . We turn the question round. **Which transformations do not change the form of (1)?**

02

Lorentz transformations — the transformations that preserve the interval

For linear transformations, preserving the operator (1) as it is is equivalent to preserving the following **interval** (the two coefficient matrices change in mutually inverse ways; scaling up or down preserves equation (1) but not the interval).

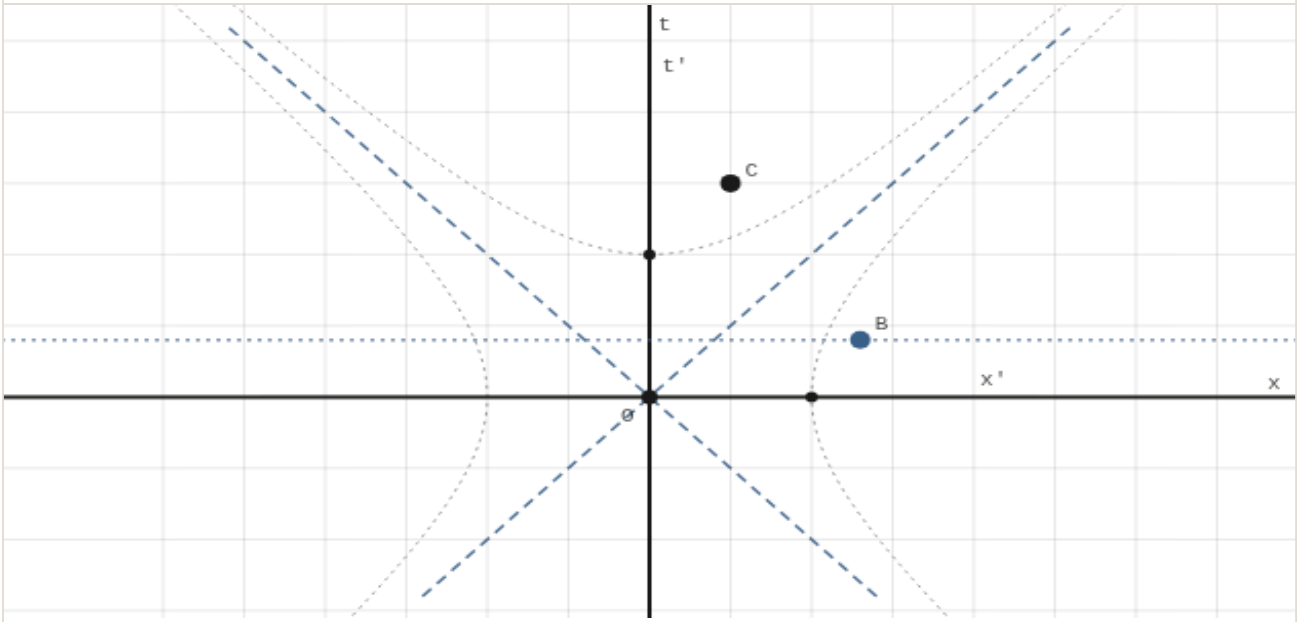
$$s^2 = -c^2t^2 + x^2 \quad (2)$$

Rotations preserve $x^2 + y^2$ and could be written as $\cos, \sin$. The transformations that preserve s^2 can be written, with $\cosh^2\varphi - \sinh^2\varphi = 1$, as $\cosh, \sinh$.

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \begin{pmatrix} \cosh\varphi & -\sinh\varphi \\ -\sinh\varphi & \cosh\varphi \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix} \quad (3)$$

We call φ the **rapidity** ($\tanh\varphi = v/c$), and this transformation a **boost**. Putting $\gamma = \cosh\varphi = 1/\sqrt{1 - v^2/c^2}$ gives $t' = \gamma(t - vx/c^2)$, $x' = \gamma(x - vt)$, and for $c \to \infty$ we return to the Galilean transformation (§ 3 of Einstein 1905 has the same form, except that the speed of light is written V) Known . The difference from rotations is in § 01 and § 04 of the spacetime article. **Example.** For $v = 0.5c$, $\gamma = 1.1547$. The time t of a clock at $x = 0$ is $t' = \gamma t$ for the moving observer, and each tick looks longer.

Figure 1 Changing the rapidity φ tilts the axes t' and x' of the moving observer (solid) towards the light cone (dashed). The hyperbolas $s^2 = \pm 1$ (dotted) do not move. The coloured dotted line is the line of events simultaneous with B for the moving observer



φ = 0.00, v/c = 0.000, γ = 1.0000, B: t' = 0.400 (after O), C: t' = 1.500, s²: B 1.53, C −2 (invariant)

Computed The numbers in the figure are computed on the spot from (3). Pressing a row of Table 1 (or selecting it with Tab and pressing Enter) sets the figure to that speed.

Computed **Table 1** γ, rapidity, and the ratio of kinetic energies ($(\gamma-1)mc^2 \div mv^2/2$). $\gamma = \cosh\varphi$ is cross-checked.

v/c	γ	φ	Energy ratio
0.1	1.005038	0.1003	1.0076
0.5	1.154701	0.5493	1.2376
0.9	2.294157	1.4722	3.1955
0.999	22.366272	3.8002	42.8181

03

The light cone and the relativity of simultaneity — three cases by the sign of the interval

The s^2 of the difference between two events does not change under boosts, and its **sign** is the same for whoever looks.

$$s^2 < 0:\ \text{timelike}, \quad s^2 = 0:\ \text{lightlike}, \quad s^2 > 0:\ \text{spacelike}$$

Two timelike-separated points can be connected by a signal slower than light, and under transformations that do not reverse the direction of time (section 04) their order is the same for whoever looks. Two spacelike-separated points cannot be connected by any signal, and **their order is exchanged depending on the observer**. The view of time and space as a single four-dimensional geometry became widely known through Minkowski's lecture of 1908. The treatment that introduces an imaginary time coordinate $t\sqrt{-1}$ and sees Lorentz transformations as rotations in four dimensions is already in Poincaré (1906) Known .

Example (Figure 1). $c = 1$. The origin O and B $= (t, x) = (0.4, 1.3)$ are spacelike-separated, with $s^2 = 1.53 > 0$. $t'_B = \gamma(0.4 - 1.3v)$ becomes negative at $v > 0.308$ ($\varphi > 0.318$), and B moves from "after" O to "before" it. C $= (1.5, 0.5)$ is timelike-separated, with $s^2 = -2$, and stays $t'_C = \gamma(1.5 - 0.5v) > 0$.

If two points whose order is exchanged were causally connected, there would be a contradiction. In quantum field theory this becomes the axiom "field operators at spacelike-separated points commute" (Jaffe–Witten § 3, 4-09). For the light cone and "the degeneration of distance", see § 03 of the spacetime article.

04

Four-vectors and the Lorentz group — writing with indices makes the invariants visible

We return space to three dimensions and write $x^\mu = (ct, x, y, z)$. When the same index appears once up and once down, it is summed.

$$\eta = \mathrm{diag}(-1,1,1,1), \quad s^2 = \eta_{\mu\nu}\Delta x^\mu \Delta x^\nu \qquad \Lambda^{\mathsf{T}}\eta\Lambda = \eta \tag{4}$$

To lower an index we multiply by η ($x_\mu = (-ct, x, y, z)$). $\partial_\mu = \partial/\partial x^\mu$, $\Box = \eta^{\mu\nu}\partial_\mu\partial_\nu = -\partial_t^2/c^2 + \nabla^2$. The signature $(-,+,+,+)$ is the same as in the spacetime article; some textbooks use the opposite convention.

A **Lorentz transformation** is a 4×4 real matrix Λ satisfying the right-hand formula of (4), and these form a group $\mathrm{O}(1,3)$ (1-02). $\mathrm{O}(1,3)$ splits into four parts according to $\det\Lambda = \pm 1$ and the direction of time (whether the top-left component is ≥ 1 or ≤ -1; by (4) its square is at least 1). The part containing the identity matrix, $\mathrm{SO}^+(1,3)$ ($\det\Lambda = 1$, preserving the direction of time), is generated by 3 rotations and boosts in 3 directions, and the group obtained by adding the 4 translations is the **Poincaré group** Known . The quantum field of the problem statement is placed on a Hilbert space carrying a representation of this group (Jaffe–Witten § 3).

Example. $\mathrm{SO}(4)$ is compact, but $\mathrm{SO}^+(1,3)$ is not, since φ can be taken as large as we like. The "compact simple gauge group" of the problem statement concerns the internal symmetry (1-04); the Lorentz group of spacetime is non-compact from the start.

05

Proper time and four-momentum — mass is a length that does not depend on the observer

We call the time of a clock carried by the particle its **proper time** τ. From the invariance of s^2, $c^2 d\tau^2 = c^2 dt^2 - |d\vec{x}|^2$, that is, $d\tau = dt/\gamma$. Since τ is the same for everyone, the **four-momentum** $p^\mu = m\,dx^\mu/d\tau$ changes by Λ, just as x^μ does.

$$p^\mu = \left(\frac{E}{c}, \vec{p}\right) = (\gamma mc,\ \gamma m\vec{v}) \qquad \eta_{\mu\nu}p^\mu p^\nu = -m^2c^2 \tag{5}$$

The right-hand formula is $E^2 = |\vec{p}|^2c^2 + m^2c^4$. Although E and $\vec{p}$ differ from observer to observer, **mass is the invariant length of the four-momentum**. **Example.** $E = \gamma mc^2$

$= mc^2 + \frac{1}{2}mv^2 + \frac{3}{8}mv^4/c^2 + \cdots$, and the rightmost column of Table 1 is this correction.

Why it is called the "mass" gap

The dispersion relation $\omega^2 = c^2k^2 + \Omega^2$ of 2-02, read as $E = \hbar\omega,\ p = \hbar k$ (2-05), is (5) itself, with $\Omega = mc^2/\hbar$. The problem statement defines the mass gap as "the spectrum of the Hamiltonian H (the set of values the energy can take; defined in 2-05 and 1-10) has no points in $(0, \Delta)$" (Jaffe–Witten § 4). When three things hold together, Lorentz invariance, positive energy ($H \geq 0$ holds for every observer) and a unique vacuum (Jaffe–Witten § 3), this is the same as "the pairs of energy and momentum $(E, \underline{p})$ of states other than the vacuum satisfy $E \geq \sqrt{|\underline{p}|^2c^2 + \Delta^2}$". For if a pair violates it, then if it is spacelike ($E < |\underline{p}|c$) a boost makes the energy negative, contradicting positive energy; if it is timelike, in its rest frame, and if it is lightlike, by a boost that makes it as small as we like, it enters $(0, \Delta)$. That only the vacuum has the pair $(0, \underline{0})$ uses the premise that the vacuum is unique Known . Δ can be read as the mc^2 of the lightest state (4-11).

06

The action of a free particle — the straight path maximises proper time

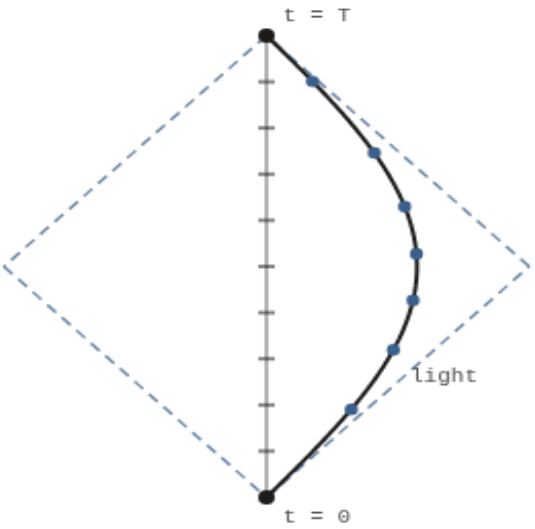
If the action (2-01) is the same for everyone, so is the form of the equation of motion. The simplest invariant of a world line is the proper time, so

$$S = -mc^2 \int d\tau$$
$$= -mc^2 \int \sqrt{1 - |\underline{v}|^2/c^2}\ dt \quad (6)$$

The coefficient was chosen so that the slow-motion expansion $L = -mc^2 + \frac{1}{2}m|\underline{v}|^2 + \cdots$ matches Newton. The equation of motion is $\frac{d}{dt}(\gamma m\underline{v}) = 0$, so the world line is straight, and the Noether conserved quantity of time translation (2-01) is $E = \gamma mc^2$. From the sign, **minimising the action and maximising the proper time are the same thing**, and among timelike world lines with fixed endpoints the straight line maximises the proper time. Seen in the frame of an observer for whom both ends are at the same place (such a frame exists because the ends are timelike-separated; call the time difference there T), $d\tau = \sqrt{1 - |\underline{v}|^2/c^2}\,dt \leq dt$, with equality only when $\underline{v} \equiv 0$. Hence $\tau \leq T$, and the maximum is attained only by the straight line (τ is the same in every frame). We call this the **reverse triangle inequality** Known . A stationary point need not be a minimum (2-01), but here it really is a minimum, and the twin paradox is a restatement of this.

Example. The proper time of $T = 1$, $x(t) = (u/\pi)\sin \pi t$ (top speed u) is $\tau = \frac{2}{\pi}\mathcal{E}(u)$ with $\mathcal{E}(k) = \int_0^{\pi/2} \sqrt{1 - k^2 \sin^2\theta}\, d\theta$ (the complete elliptic integral of the second kind). For $u \to 1$, $2/\pi \approx 0.637$.

Figure 2 Changing the top speed u bends the world line with fixed endpoints, and the proper time (coloured dots: every 0.1 T) decreases. Pressing Play shows the two clocks at the same coordinate time



u = 0.90, proper time of the bent path 0.7459 T (the straight path is 1 T)

Computed The numbers in the figure are computed on the spot by the trapezoidal rule for $d\tau = dt/\gamma$ (400 subdivisions).

Computed **Table 2** τ/T of world lines with fixed endpoints. For the curve, the numerical integral and $\frac{2}{\pi}\mathcal{E}(u)$ agree to 2×10^{-16}. The broken line is the path that goes out and back at speed u (the twins).

u/c	Curve	Approx. 1 − u²/4	Broken line √(1 − u²)
0.3	0.977105	0.977500	0.953939
0.6	0.902780	0.910000	0.800000
0.9	0.745926	0.797500	0.435890
0.99	0.654748	0.754975	0.141067

07

Lorentz-invariant actions of fields — scalar fields, the electromagnetic field, Yang–Mills fields

The action of a field is $S = \frac{1}{c}\int \mathcal{L}\, d^4x$. Since d^4x does not change under $|\det \Lambda| = 1$, the action is invariant if the Lagrangian density $\mathcal{L}$ is a scalar. If we also impose the condition of being **local** (only the field at that point and its first derivatives), the form of the terms is strongly narrowed down.

For a **scalar field**, $\partial_\mu\phi$ is contracted with η.

$$\mathcal{L} = -\tfrac{1}{2}\eta^{\mu\nu}\partial_\mu\phi\,\partial_\nu\phi - \tfrac{1}{2}\mu^2\phi^2$$
$$(\Box - \mu^2)\phi = 0, \quad \mu = \frac{mc}{\hbar} \qquad (7)$$

The right-hand one is the Klein–Gordon equation. **Example.** In one dimension, multiplying the whole by the tension σ gives the string with a mass term of 2-02, with the correspondence $c^2 = \sigma/\rho$, $\Omega = c\mu$.

The electromagnetic field. Gather the potentials $\phi, \boldsymbol{A}$ of 2-03 (this ϕ is different from the field of (7)) as $A_\mu = (-\phi/c, \boldsymbol{A})$, and put $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. The components are $F_{0i} = -E_i/c$, $F_{ij} = \varepsilon_{ijk}B_k$ (ε_{ijk} is the symbol with $\varepsilon_{123} = 1$ that changes sign when indices are exchanged), and it gathers the electric and magnetic fields into one antisymmetric matrix. It does not change under the gauge transformation $A_\mu \to A_\mu + \partial_\mu\chi$ (the formula for gauge transformations in 2-03). Since $F^{0i} = -F_{0i}$, $F^{ij} = F_{ij}$, we have $F_{\mu\nu}F^{\mu\nu} = -2|\boldsymbol{E}|^2/c^2 + 2|\boldsymbol{B}|^2$, and

$$\mathcal{L} = -\tfrac{1}{4}F_{\mu\nu}F^{\mu\nu} = \tfrac{1}{2}\left(|\boldsymbol{E}|^2/c^2 - |\boldsymbol{B}|^2\right) \qquad (8)$$

In SI units the whole is multiplied by $1/\mu_0$, and this becomes the part of (L) of 2-03 without $\rho, \boldsymbol{J}$. The equation of motion $\partial_\mu F^{\mu\nu} = 0$ is half of Maxwell's equations in vacuum, and the other half holds as soon as F is built from A.

Yang–Mills fields. Make $A_\mu = A^a_\mu T^a$ a field with values in a Lie algebra (1-03, 1-04), and add a term with the structure constants f^{abc} to F (3-02).

$$\mathcal{L} = -\frac{1}{4g^2}\sum_a F^a_{\mu\nu}F^{a\,\mu\nu}$$
$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + f^{abc}A^b_\mu A^c_\nu \qquad (9)$$

In matrices, $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - i[A_\mu, A_\nu]$ (T^a Hermitian with $[T^a, T^b] = if^{abc}T^c$, the convention of 1-03; g absorbed into A). The last term is the only difference from Maxwell, and the field becomes its own source. Even so, the classical equations describe "massless waves travelling at the speed of light" (Jaffe–Witten § 1) Known .

If we also impose gauge invariance (3-02) and the group is simple (the invariant quadratic form is unique up to a constant multiple; 1-03), the Lorentz-invariant quadratic expressions in F are only the linear combinations of $\sum_a F^a_{\mu\nu}F^{a\,\mu\nu}$ and $\sum_a \varepsilon^{\mu\nu\rho\sigma}F^a_{\mu\nu}F^a_{\rho\sigma}$ (the four-dimensional version of the symbol $\varepsilon^{0123} = 1$) Known . For the electromagnetic field the second is proportional to $\boldsymbol{E}\cdot\boldsymbol{B}$; it changes sign under spatial reflection and, being a total derivative, does not change the equations of motion (the θ

term of 3-05). Without the condition, the $\sum c_{ab} F^{a}_{\mu\nu} F^{b\,\mu\nu}$ built from an arbitrary symmetric matrix c_{ab} is also invariant.

08

Counting dimensions — ħ = c = 1, and the dimension of the coupling constant

With $c = 1$, length and time are put on the same footing, and with $\hbar = 1$ energy and inverse time, so one unit remains. If mass has dimension 1, length has -1. The only conversion is $\hbar c = 197.3269804\ldots$ MeV fm ($1\,\text{fm} = 10^{-15}\,\text{m}$). The gap Δ corresponds to a length $\hbar c/\Delta$, a rough measure of how far correlations reach. If there is a gap, then for a local quantity O whose vacuum expectation value $\langle O\rangle$ is 0, the correlation $|\langle O(\underline{x})O(\underline{y})\rangle|$ at two points at the same time and far enough apart is at most $\exp(-C|\underline{x}-\underline{y}|)$ for any C satisfying $C < \Delta/(\hbar c)$ (Jaffe–Witten § 5; for $\hbar = c = 1$, $C < \Delta$). If the expectation value is not 0, the correlation approaches $\langle O\rangle^2$.

Computed **Table 3** Converting masses into lengths ($\hbar c$ computed from the defined SI values h, c, e).

Item	mc² (MeV)	ħ/(mc) (fm)
Electron	0.51099895069	386.159267
Proton	938.27208943	0.210309

The action is dimensionless (measured in $\hbar$), so in a spacetime of dimension d, $[\mathcal{L}] = d$, $[\partial] = 1$. Since $(\partial\phi)^2$ and $\lambda\phi^4$ have dimension d, and since F of (9) is the sum of ∂A and AA with $[A] = 1$, $[F^2] = 4$,

$$[\phi] = \frac{d-2}{2}, \quad [\lambda] = [g^2] = 4 - d \tag{10}$$

Computed **Table 4** Mass dimensions ($\hbar = c = 1$, computed with fractions).

d	[φ]	[λ]	[g²]	Reading
2	0	2	2	coupling has dimension mass²
3	1/2	1	1	coupling has dimension mass
4	1	0	0	coupling is dimensionless
5	3/2	−1	−1	coupling has dimension length

In four dimensions, the action (9) contains not a single constant with dimensions. Neither a mass nor a length can be built from the coefficients, and classical Yang–Mills theory has no scale Known . If in the quantum theory there is $\Delta > 0$, the scale must come from outside the classical action; in physics it is thought to enter in the process of renormalisation, and this is called dimensional transmutation (4-07) Physics . The

exposition too gives the fact that in four-dimensional gauge theory the mass gap is not classically visible ('the mass gap is not classically visible') as a reason why a non-classical change of variables may be needed (Jaffe–Witten § 6.6). In two and three dimensions g^2 itself is a scale (4-04, 7-06). For ϕ^4 see 7-08; for other ways four dimensions is special, see the four-dimensions article. A figure that moves the d of Table 4 continuously and sets it beside other quantities that switch at four is the dimension dial of 1-12 § 10.

09

A bridge to imaginary time — putting one sign back, mass becomes a rate of decay

Substituting $t = -i\tau$ (τ real) into (2) gives $-c^2t^2 = c^2\tau^2$, and the interval becomes the square of the Euclidean distance $c^2\tau^2 + |\underline{x}|^2$. This is the **Wick rotation** (the name is said to come from Wick 1954), which formally puts back the difference of sign; it is the same operation as the ict of § 02 of the spacetime article (the authoritative treatment of its form for actions and correlations of fields is 4-02 § 02). Since $\partial_t^2 = -\partial_\tau^2$, $\Box$ becomes the four-dimensional Laplacian, (7) becomes $(-\partial_\tau^2/c^2 - \nabla^2 + \mu^2)\phi = 0$, and oscillation turns into decay. The one-dimensional version, dropping the dependence on space and putting $c = 1$, is (11). $\delta(\tau)$ is the thing concentrated at $\tau = 0$ with $\int f(\tau)\delta(\tau)\,d\tau = f(0)$ (strictly a distribution rather than a function; 1-11).

$$\left(-\frac{d^2}{d\tau^2} + \mu^2\right) G = \delta(\tau)$$
$$G(\tau) = \frac{e^{-\mu|\tau|}}{2\mu} \qquad (11)$$

Example. For $\tau \neq 0$, $G'' = \mu^2 G$; at $\tau = 0$ the slope G' jumps by -1, so G'' contains $-\delta$, and $-G'' + \mu^2 G = \delta$. In three dimensions, $e^{-\mu r}/(4\pi r)$ Known . **On the Euclidean side, mass appears as a rate of decay (an inverse length).** This is the starting point of 4-11, which ties the mass gap to the correlation length. Writing the Euclidean theory as a probability measure (4-02) and the conditions for reconstructing the Minkowski side (Osterwalder–Schrader) are the subject of 4-10. This chapter has only shown the substitution.

What this chapter can say

Content	Status
Lorentz transformations, the three classes of intervals, four-momentum, the action of a free particle and the maximum of proper time, the forms of the actions of three kinds of fields	Said Known (numbers also checked in Tables 1 and 2 Computed)
That under Lorentz invariance, positive energy and a unique vacuum, the definition of the gap becomes the same as $E \geq \sqrt{\lvert p\rvert^2 c^2 + \Delta^2}$	Said Known (section 05; the axiomatic framework is in 4-09 and 4-11)
That the classical action of pure Yang–Mills in four dimensions has no constant with dimensions	Said Known (Table 4 Computed)
That the scale of the gap comes from renormalisation (dimensional transmutation)	At the level of physics Physics (4-07)
Justification of the Wick rotation, curved spacetime, spin	Not treated
The mass gap of Yang–Mills	This chapter shows nothing

Sources and reproduction

Item	Kind	Source / tools
The coefficients of section 01, Tables 1–4, Figures 1–2	Computed	Tables by python3 (NumPy, SciPy). Figures by JavaScript in the browser ((3) and the trapezoidal rule). The static images shown when JavaScript is disabled are SVGs made by python3 from the same formulas
The two premises and Lorentz transformations	Known	A. Einstein, “Zur Elektrodynamik bewegter Körper”, Ann. Phys. 17 (1905) 891–921. Introduction and § 3 checked (the speed of light is written V)
The view as the geometry of spacetime	Known	H. Minkowski, “Raum und Zeit”, Phys. Z. 10 (1909) 104–111 (lecture in Cologne on 21 September 1908). Opening checked
The imaginary time coordinate, seeing Lorentz transformations as rotations in four dimensions	Known	H. Poincaré, “Sur la dynamique de l’électron”, Rend. Circ. Mat. Palermo 21 (1906) 129–176. Bibliographic data only (only the relevant passage checked in the Wikisource transcription)
The definition of the gap, the Poincaré group, positive energy and uniqueness of the vacuum, locality, the cluster property, classical waves	Known	A. Jaffe, E. Witten, “Quantum Yang–Mills theory” (the Clay Mathematics Institute's problem description). § 1, § 3– § 5 and § 6.6 checked
The name of the Wick rotation	Known	G. C. Wick, Phys. Rev. 96 (1954) 1124–1134. Bibliographic data only (attribution not checked)

Item	Kind	Source / tools
ħc, the masses of the electron and the proton	Known	CODATA 2022 (checked at NIST). ħc agrees with the calculation from the defined SI values
The maximum of proper time, the four parts of $\mathrm{O}(1,3)$, the two invariants of F (under gauge invariance and a simple group), the components of the electromagnetic field, counting dimensions	Known	Standard facts (can be checked with the formulas in the text)

Next chapter: 2-05 The bones of quantum mechanics. Imaginary time continues in 4-02, and spacetime itself in One sign is different.

The bones of quantum mechanics — states are vectors, energies are a spectrum, and the gap is the space at the very bottom

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That a state is a unit vector, an observable is a self-adjoint operator, and the values of the energy are a spectrum.
That the gap $E_1 - E_0$ can be read in three ways: as "the price of an excitation", "the speed of oscillation" and "the rate of decay in imaginary time".
Gaps that close and gaps that remain, a gap invisible to perturbation theory, and the agreement with the Poincaré constant.

Prerequisite chapters — 1-10 (Hilbert spaces, spectra, the variational principle, the Poincaré inequality). For the case of matrices, 1-01.

Said first — Only quantum mechanics with finitely many degrees of freedom is treated. The mass gap of quantum field theory is treated in 4-11, and this chapter shows nothing about the mass gap of Yang–Mills.

Order of this chapter

1. States and observables — states are vectors, measured quantities are operators
2. Time evolution — the gap between levels sets the speed of the back-and-forth
3. The spectrum — values that are not eigenvalues also enter (Detour: a well whose valleys go on for ever)
4. The ground state — trial functions give only upper bounds (Table 1)
5. Three readings of the gap — price, oscillation, decay (Figure 1, Table 2)
6. Gaps that close, gaps that remain — comparing as the box is widened (Table 3)
7. The double well — a gap invisible to perturbation theory (Figure 2, Table 4)
8. The gap and the Poincaré inequality — the gap is the inverse of a weighted constant

01

States and observables — vectors and self-adjoint operators

In quantum mechanics a state is a vector ψ of length 1 in a **complex Hilbert space** $\mathcal{H}$ (1-10). A sum $\alpha\psi + \beta\phi$ rescaled to length 1 is also a state (**superposition**), and $z\psi$, multiplied by z with $|z| = 1$, is the same state (the phase of 2-03). The inner product is taken to be conjugate-linear in the left argument.

$$\mathcal{H} = \mathbb{C}^2 : \ |\alpha|^2 + |\beta|^2 = 1; \qquad \mathcal{H} = L^2(\mathbb{R}) : \ \int_{\mathbb{R}} |\psi(x)|^2 \, dx = 1 \tag{1}$$

On the left is a system with two choices (spin, or "the left valley or the right valley"); on the right is a particle on a line, and we read $|\psi(x)|^2$ as the probability density of position (this is said to go back to Born 1926). Measurable quantities correspond to **self-adjoint operators** A (Hermitian matrices in finite dimensions, 1-01); the value of a measurement is one of the eigenvalues a_k, and its probability, using orthonormal eigenvectors e_k, is $|\langle e_k, \psi\rangle|^2$.

$$\langle A \rangle_\psi = \sum_k a_k \, |\langle e_k, \psi\rangle|^2 = \langle \psi, A\psi \rangle \tag{2}$$

Example. Measuring $\sigma_z = \mathrm{diag}(1, -1)$ in $\psi = (1, 1)/\sqrt{2}$ gives ± 1 half and half. For $\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $\sigma_x \psi = \psi$, and ψ is an eigenvector with eigenvalue $+1$, so a measurement certainly gives $+1$ (σ_x, σ_z are the Pauli matrices, 1-04). In $L^2(\mathbb{R})$ the position X is the operator of multiplication by x, the momentum is $P = -i\hbar \, d/dx$, and from $XP - PX = i\hbar$ comes the uncertainty relation $\Delta X \, \Delta P \geq \hbar/2$ Known .

02

Time evolution — stationary states and the back-and-forth of two levels

We call the operator of energy the **Hamiltonian** H. For a particle on a line it is $H\psi = -\frac{\hbar^2}{2m}\psi'' + V\psi$. States obey the Schrödinger equation.

$$i\hbar \frac{d\psi}{dt} = H\psi, \qquad \psi(t) = e^{-itH/\hbar}\psi(0) = \sum_n c_n \, e^{-iE_n t/\hbar} \, \phi_n \tag{3}$$

The right-hand end is the expansion in eigenvectors $H\phi_n = E_n \phi_n$, and each component only has its phase turning. $e^{-itH/\hbar}$ is unitary and preserves total probability (1-01; the converse is Stone's theorem Known). A state with a single

eigenvector is a **stationary state** whose distribution does not change, and change arises only from the phase differences $(E_n - E_m)t/\hbar$.

Example (two levels). A system with the left valley e_L, the right valley e_R and the tunnelling amplitude $g > 0$.

$$H = \begin{pmatrix} \varepsilon & -g \\ -g & \varepsilon \end{pmatrix}, \qquad E_{\mp} = \varepsilon \mp g, \qquad P_R(t) = \sin^2 \frac{gt}{\hbar} \quad (4)$$

Starting from e_L, the probability P_R of being on the right goes back and forth with angular frequency $2g/\hbar = (E_+ - E_-)/\hbar$. **The gap between levels sets the speed of the back-and-forth** (the $x = 0$ of Figure 2 of 1-01). Section 07 gives an example where g is exponentially small.

03

The spectrum — eigenvalues alone are not enough

In infinite dimensions, values without eigenvectors also appear, so we use the **spectrum** $\sigma(H)$ (the set of λ for which $H - \lambda$ has no bounded inverse, 1-10). For a self-adjoint operator, $\lambda \in \sigma(H)$ is the same as "there is a sequence with $\|\psi_k\| = 1$, $\|(H - \lambda)\psi_k\| \to 0$" (Weyl's criterion) Known .

$$\text{box } [0, L]: \ E_n = \frac{\hbar^2 \pi^2 n^2}{2mL^2} \ (n \geq 1); \qquad \text{line}: \ \sigma\Big(\frac{P^2}{2m}\Big) = [0, \infty) \quad (5)$$

The spectrum of a box with walls where $\psi = 0$ consists only of eigenvalues (**discrete spectrum**). On a line the plane wave $e^{ipx/\hbar}$ is not a state, since $\int |\psi|^2 = \infty$. **Example.** For a wave packet $\ell^{-1/2}\chi(x/\ell)e^{ipx/\hbar}$ of width ℓ (χ a smooth bump), $\|(H - p^2/2m)\psi\|$ approaches 0 like $1/\ell$, so every $E \geq 0$ is in the spectrum without being an eigenvalue (**continuous spectrum**). Widening the box makes the spacing on the left tighter, approaching the right (section 06).

Detour: a well whose valleys go on for ever

A subsection you may skip.

For $V = \frac{1}{2}x^2y^2$ ($\hbar = m = 1$) on the plane, valleys with $V = 0$ run on for ever along the coordinate axes, and a classical particle can escape as far as it likes. In quantum mechanics the valley narrows the further in it goes, and the lowest energy $|x|/2$ of the oscillator in the y direction (2-06) acts as a wall. Strictly, this is an estimate that works only deep in the valleys. Computed Solving on the square $[-R, R]^2$, the lowest eigenvalue stops at $0.5671,\ 0.55413,\ 0.55411$ for $R = 3, 5, 9$, while the value for a free particle $\pi^2/4R^2$ goes to 0. For a uniform $A_1 = x\sigma_1,\ A_2$

$= y\sigma 2$ field SU(2), the potential energy $-\operatorname{tr}[A_1, A_2]^2$ of Yang–Mills is proportional to x^2y^2 (3-02 § 07). What the numbers on this machine show goes only as far as the tendency of the lowest eigenvalue to stop as the box is widened. The exposition of the problem (Jaffe–Witten § 6.6) cites Simon (1983, titled "quantum operators with discrete spectrum but classically continuous spectrum") as an example in which bound states arise even when there are flat directions Known .

04

The ground state — estimates from above are easy

We call $E_0 = \inf \sigma(H)$ the **ground-state energy**, and, if it is an eigenvalue, its eigenvector ϕ_0 the **ground state**. The variational principle (1-10) says the following.

$$E_0 = \inf_{\psi \neq 0} \frac{\langle \psi, H\psi \rangle}{\langle \psi, \psi \rangle} \quad (6)$$

Every trial function gives an **upper bound**, but a lower bound requires control of every ψ (this asymmetry appears throughout the book, so it is summarised in the box in 1-10 § 06). Three remarks. (i) The bottom need not be an eigenvalue (the free particle on a line). (ii) It does not fall to the bottom of the potential. For $V = \frac{1}{2}m\omega^2x^2$, $\langle X^2\rangle\langle P^2\rangle \geq \hbar^2/4$ and the inequality of arithmetic and geometric means give $\langle H\rangle \geq \hbar\omega/2$ (2-06). (iii) In one dimension, if V grows far away, the ground state is non-degenerate and can be taken positive, and the n-th eigenfunction has n zeros (Sturm's oscillation theorem) Known .

Example. Putting a Gaussian $e^{-x^2/4s^2}$ into $\hbar = m = 1$, $H = -\frac{1}{2}d^2/dx^2 + x^4$ gives $\langle H\rangle_s = 1/8s^2 + 3s^4$, and its minimum $\frac{3}{16}48^{1/3} = 0.68142$ is an upper bound for $E_0 = 0.667986$. The odd trial function $xe^{-x^2/4s^2}$ is orthogonal to the even ground state, so it gives an upper bound for E_1.

Computed **Table 1** The best upper bounds from Gaussian trial functions, and the eigenvalues of three-point differences.

V	Upper bound on E_0	E_0	Upper bound on E_1	E_1	Difference of upper bound
x^4	0.681420	0.667986	2.423739	2.393644	1.74231

The difference between two upper bounds is neither an upper nor a lower bound on the gap (depending on the trial functions it can be off in either direction). To bound it from below, we need estimates for every state orthogonal to the ground state. Showing the mass gap in quantum field theory is likewise, by definition, bounding from below the energy of the states orthogonal to the vacuum (4-11).

05

Three readings of the gap — price, oscillation, decay

When the ground state is non-degenerate and E_0 is an isolated point of the spectrum, we call $\Delta = \inf(\sigma(H) \setminus \{E_0\}) - E_0 > 0$ the **gap**. The exposition of the problem sets the energy of the vacuum to 0 and defines the mass gap as "the spectrum of H has no points in $(0, \Delta)$" (Jaffe–Witten § 4) Known . There are three readings. (i) **Price**: precisely, for states orthogonal to ϕ_0, $\langle H \rangle \geq E_0 + \Delta$. (ii) **Oscillation**: a superposition of ϕ_0 and ϕ_1 oscillates with angular frequency $\Delta/\hbar$ ((4)). (iii) **Decay**: formally putting $t = -i\tau$ into (3) (2-04), oscillation turns into decay.

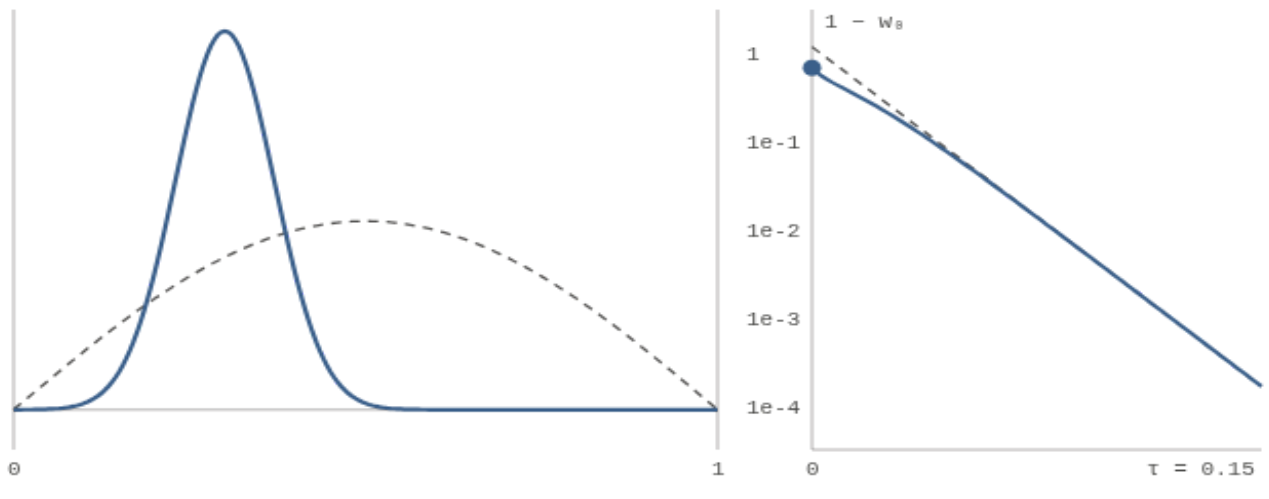
$$e^{-\tau(H-E_0)/\hbar}\psi = c_0\phi_0 + \sum_{n\geq 1} c_n\, e^{-(E_n-E_0)\tau/\hbar}\phi_n \quad (7)$$

For $c_0 \neq 0$ it approaches the ground state, and the rest becomes at most $e^{-\Delta\tau/\hbar}$ times as large. Write the left side of (7) as $\psi(\tau)$, and the **weight** of the ground state as $w_0(\tau) = |\langle\phi_0, \psi(\tau)\rangle|^2/\|\psi(\tau)\|^2$. $1 - w_0$ is the probability of being outside the ground state, and Figure 1 and Table 2 follow it. For the **imaginary-time correlation** of a self-adjoint A ($\langle\phi_0, A\phi_0\rangle = 0$) it takes the following form.

$$0 \leq \langle A\phi_0,\, e^{-\tau(H-E_0)/\hbar}A\phi_0\rangle \leq e^{-\Delta\tau/\hbar}\,\|A\phi_0\|^2 \quad (8)$$

From (8), the rough time for the correlation to vanish in the imaginary-time direction is at most $\hbar/\Delta$. This is a quantity with the dimension of time; in relativity it is multiplied by c to give a length $\hbar c/\Delta$. Decay in the spatial direction (the cluster property) and the correlation length (2-08) are treated in 4-11, and the view as a sum over paths in 2-07.

Figure 1 As imaginary time τ advances, the bump centred at x = 0.3 (colour) approaches the ground state ϕ_0 (dashed). On the right is $1 - w_0$ (common logarithmic scale), which lines up with the prediction keeping only the next level (dashed; slope -2Δ in natural logarithms, $\hbar = 1$)



```
τ = 0.000, w₀ = 0.30949, 1 − w₀ = 6.91e-1, prediction 1.20e+0
```

Computed The numbers in the figure were computed on this machine (box $[0,1]$, $H=-d^2/dx^2$ ($\hbar=1$, $\hbar^2/2m=1$), $\Delta=3\pi^2$; the coefficients were found with python3, and the partial sums of (7) are computed on the spot).

Computed **Table 2** The weight outside the ground state $1-w_0$, and the prediction keeping only the next level $(c_1/c_0)^2 e^{-2\Delta\tau}$.

τ	1 − w₀	Prediction
0	6.905e-1	1.195e+0
0.05	5.831e-2	6.189e-2
0.10	3.194e-3	3.204e-3
0.15	1.659e-4	1.659e-4

The weight is the square of an amplitude, so the exponent is 2Δ. For a state symmetric about the centre, $c_1=0$, and the decrease is set by $E_2-E_0=8\pi^2$. **What the gap guarantees is the slowest decrease, valid for every state.**

06

Gaps that close, gaps that remain — widening the box

The mass gap is asked about in infinitely wide space. A gap that is positive in a finite box can close when the box is widened. The exposition of the problem also says that a **uniform** gap, independent of the volume, could play a role in proving the infinite-volume limit, and that the proof would need new ideas (Jaffe–Witten § 5, § 6.5) Known . With $\hbar=m=1$ and box $[-L/2, L/2]$, we compare three V.

Computed **Table 3** The length of the box L and the gap E_1-E_0 (three-point differences; $V=0$ agrees with the exact value $3\pi^2/2L^2$).

L	V = 0	V = −sech²x	V = x²/2
10	0.14804	0.57518	1.00000
40	0.00925	0.50342	1.00000
160	0.00058	0.50020	1.00000

The gap of the free particle closes like $1/L^2$. $-\operatorname{sech}^2 x$ has exactly one bound state $\operatorname{sech} x/\sqrt{2}$ ($E_0=-1/2$, which can be checked by substitution), and the next level drops to the bottom 0 of the continuous spectrum, so the gap tends to $1/2$ Known . **Above the gap there may be continuous spectrum.** Even in a field theory with a mass gap, above the vacuum is the continuous spectrum of particles, starting from Δ (2-04).

Example. For N independent oscillators, the ground energy $N\hbar\omega/2$ grows with N and diverges as $N\to\infty$. Even so, the gap stays $\hbar\omega$ (2-06). In quantum field theory the

energy of the vacuum is set to 0 and the difference is what is asked about. Jaffe–Witten § 4 states $H\Omega = 0$ from the fact that the vacuum vector Ω is invariant under the Poincaré group (2-04).

07

The double well — a gap invisible to perturbation theory

Put two identical wells side by side and make the hill between them higher ($\hbar = m = 1$).

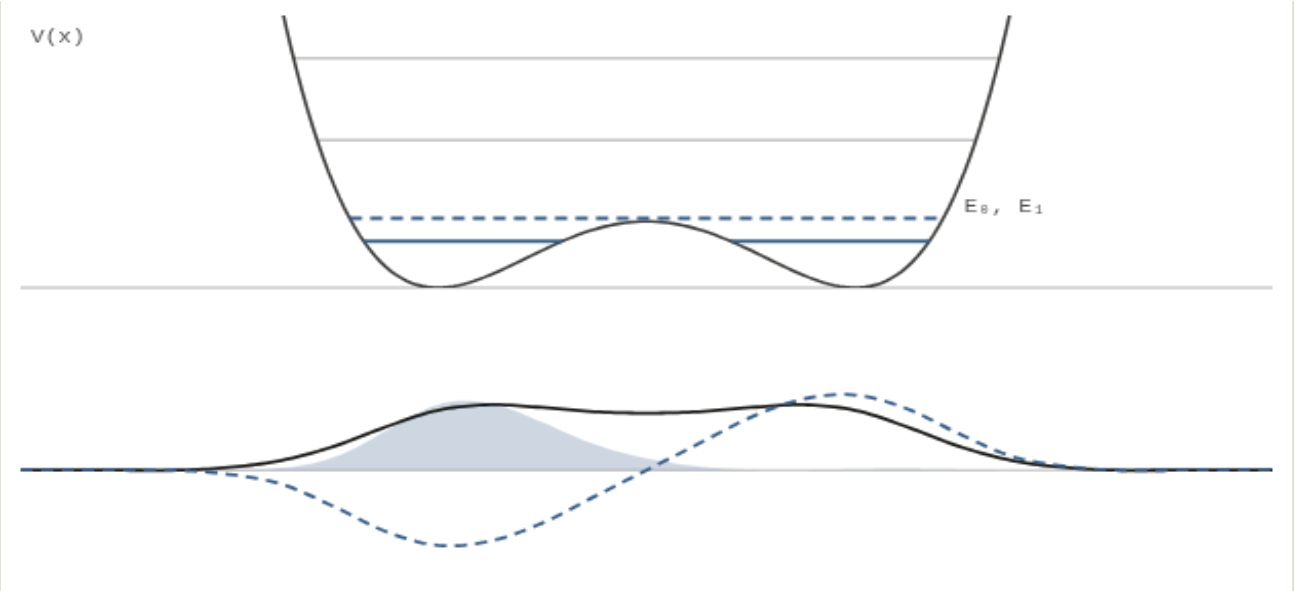
$$V_a(x) = \frac{(x^2 - a^2)^2}{8a^2}, \qquad V_a''(\pm a) = 1, \qquad S_0 = \int_{-a}^{a} \sqrt{2V_a}\, dx = \frac{2a^2}{3} \quad (9)$$

With a single valley the lowest level is about $1/2$, and with two valleys it is doubly degenerate. Tunnelling between the valleys plays the role of the g of (4), the levels split into an even ϕ_0 and an odd ϕ_1, and when the hill is high the splitting is as follows Known .

$$E_1 - E_0 \approx \frac{4a}{\sqrt{\pi}}\, e^{-2a^2/3} = 2\sqrt{\frac{6S_0}{\pi}}\, e^{-S_0} \quad (10)$$

This is equation (2.8) of Zinn-Justin–Jentschura (2004), $\frac{2}{\sqrt{\pi g}} e^{-1/6g}$, with $g = 1/4a^2$ (shifting the potential of that paper by a gives (9); the height of the hill is $a^2/8 = 1/32g$). **Perturbation theory** splits V_a around one of the valleys into an oscillator term and a remainder that is small when g is small, expands the levels in powers of g, and finds the coefficients of each order in turn (4-03). That perturbation series does not depend on parity (evenness or oddness under $x \to -x$) and is the same for the two levels (the explanation of (2.13)–(2.14) in that paper) Known . The $e^{-1/6g}$ of the gap, expanded as a power series in g, has all coefficients 0. **However high the order to which perturbation theory is computed, this gap does not appear.** S_0 is the action of the classical solution (instanton) that rolls from hilltop to hilltop of $-V_a$ in imaginary time. Its counterpart in Yang–Mills is in 3-05, and quantities of the form e^{-1/g^2} appear in 4-07 and 4-12.

Figure 2 Increasing a moves the valleys (±a) apart and raises the hill, and the levels E_0 (coloured solid line) and E_1 (coloured dashed line) draw together and coincide. The lower panel shows ϕ_0 (solid), ϕ_1 (dashed), and the probability density of a state starting from the left valley (shaded). "Play" moves it to the right valley (time shown compressed)




```
a = 2, S₀ = 2.667, E₀ = 0.35024, E₁ = 0.52319, Δ = 0.1729 ((10) gives 0.3136), π/Δ = 18.17,
probability of being on the right at t = 0.000: 0.041
```

Computed The numbers in the figure were computed on this machine (levels and wave functions by three-point differences in python3; the shading is $\frac{1}{2}(\phi_0^2 + \phi_1^2 - 2\phi_0\phi_1 \cos \Delta t)$).

Computed **Table 4** Comparison of the gap of (9) (three-point differences) with (10). The rightmost column is the correction factor $1 - \frac{71}{12}g - \frac{6299}{288}g^2$ from the coefficient (equation (2.19a)) of the expansion (2.13) that the same paper states as a conjecture Physics .

a	Δ (numerical)	Formula (10)	Ratio	Correction factor
2	1.729e-1	3.136e-1	0.551	0.545
3	1.355e-2	1.678e-2	0.807	0.819
4	1.895e-4	2.104e-4	0.901	0.902
5	6.118e-7	6.519e-7	0.9384	0.9386

The ratio approaches 1, and also the correction factor (difference 2×10^{-4} at $a = 5$). The gap at $a = 5$ is 6 orders of magnitude smaller than the level spacing (about 1), and the time π/Δ to move to the right is about 5×10^6.

08

The gap and the Poincaré inequality — weighting by the ground state

In one dimension when $\phi_0 > 0$ (section 04), writing $\psi = f\phi_0$ with a real-valued f and integrating by parts, the $V - E_0$ term vanishes by $H\phi_0 = E_0\phi_0$.

$$\langle \psi, (H - E_0)\psi \rangle = \frac{\hbar^2}{2m} \int |f'|^2 \, d\mu, \qquad d\mu = \phi_0^2 \, dx \tag{11}$$

$\psi \perp \phi_0$ means $\int f\,d\mu = 0$, and then $\|\psi\|^2 = \mathrm{Var}_\mu(f)$. Applying the variational principle to ψ orthogonal to ϕ_0, the following holds with the best constant.

$$\mathrm{Var}_\mu(f) \le C_P \int |f'|^2\,d\mu, \qquad C_P = \frac{\hbar^2}{2m\Delta} \quad (12)$$

The gap is inversely proportional to the Poincaré constant of the ground-state distribution μ Known (the Poincaré inequality is in 1-10 § 07, the weighted form in § 08). Here μ is the distribution of the ground state at a single time. Matching with $\psi = f\phi_0$, $H - E_0$ corresponds to $\frac{\hbar^2}{2m}L$ (L is the operator of 1-10 § 08 for this μ), and the gap of L is $2m\Delta/\hbar^2$. The $C_P \approx \xi^2$ of 1-10 § 09 is a statement about the measure on the whole of the paths in imaginary time (1-08 (10)), and is a different quantity. **Example.** For the oscillator ($\hbar = m = \omega = 1$), $\phi_0 = \pi^{-1/4}e^{-x^2/2}$ (which can be checked by substitution), so μ is the normal distribution with variance $1/2$, and the gap 1 (the right column of Table 3, 2-06) gives $C_P = 1/2$. "The Poincaré constant of a normal distribution equals its variance" has the same content as the gap $\hbar\omega$ of the oscillator (with equality at $f = x$). For the double well $a = 3$, $C_P = 1/2\Delta = 36.9$. Taking f close to "-1 on the left, $+1$ on the right", the variance is about 1, but the slope acts only on the hill between, where μ is almost absent. **A waist in the distribution makes the gap small.** The same form is also used in one-link integration (5-11, this site's papers).

What this chapter can say

What is said	What is not said (and where it is treated)
Known The framework of states, spectra and the variational principle, the levels and gaps of Tables 1–3, the decrease in imaginary time (numbers in Tables 1–3 and Figure 1 Computed)	Stone's theorem, Weyl's criterion and Sturm's oscillation theorem are cited, not proved
Computed The tendency of the lowest eigenvalue of $\frac{1}{2}x^2y^2$ to stop as the box is widened to $R = 3, 5, 9$	A proof that it stops (discrete spectrum is the subject of Simon (1983); its text has not been read)
Known The leading term (10) of the double-well gap, and that it appears at no order of the perturbation series (the tendency of the ratio towards 1 is Table 4 Computed)	The derivation of (10). The expansion (2.13) including the correction coefficients is a conjecture of that paper Physics (instantons are in 3-05, divergent series in 4-12)

What is said	What is not said (and where it is treated)
Known That the gap is inversely proportional to the Poincaré constant of the ground-state distribution (one dimension, $\phi 0 > 0$)	The justification of (11) with infinitely many variables. The relation to $C_P \approx \xi^2$ of the path measure (1-10, 23-01)
—	The mass gap with infinitely many degrees of freedom and the requirements of relativity (4-11). This chapter shows nothing about the mass gap of Yang–Mills

Sources and reproduction

Item	Kind	Source / tools
Tables 1–4, the numbers in the detour, Figures 1–2	Computed	python3 (three-point differences, sparse matrices and numerical integration in NumPy and SciPy). The figures by JavaScript in the browser
The definition of the mass gap, $H\Omega = 0$, the uniform gap, the remark on flat directions	Known	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory" (the official problem statement of the Clay Mathematics Institute; the problem was announced in 2000). In: J. Carlson, A. Jaffe, A. Wiles (eds.), *The Millennium Prize Problems*, CMI / AMS, pp. 129–152 (year of publication of the book not checked). § 4, § 5, § 6.5 and § 6.6 checked
The leading term (10) of the double-well gap, the independence of the perturbation series from parity	Known	J. Zinn-Justin, U. D. Jentschura, Ann. Phys. 313 (2004) 197–267, arXiv:quant-ph/0501136. Equations (2.1)–(2.8) of § 2.1 and the explanation of (2.13)–(2.14) checked
The correction factor of Table 4	Physics	The coefficient (2.19a) of the expansion (2.13) that the same paper states as a conjecture (cited there without proof). On this machine the tendency was checked numerically Computed
The Schrödinger equation	Known	E. Schrödinger, "Quantisierung als Eigenwertproblem", Ann. Phys. 384 (1926) 361–376 (first communication; the time-independent eigenvalue problem), and the fourth communication of the same title, Ann. Phys. 386 (1926) 109–139. The time-dependent form (3) is said to be in the fourth communication (not checked). Bibliographic data only for both
The probabilistic interpretation / Stone's theorem / flat directions and discrete spectrum	Known	M. Born, Z. Phys. 37 (1926) 863–867 (attribution from secondary sources) / M. H. Stone, Ann. of Math. 33 (1932) 643–648 / B. Simon, "Some quantum operators with discrete spectrum but classically continuous spectrum", Ann. Phys. 146 (1983) 209–220 (reference [43] of Jaffe–Witten; the passage treating $x^2 y^2$ not checked). Bibliographic data only for all

Item	Kind	Source / tools
The variational principle, the Poincaré inequality (weighted form), the Poincaré constant of the normal distribution	Known	1-10 § 06, § 07, § 08
Weyl's criterion, Sturm's oscillation theorem, the uncertainty relation, (11)	Known	Standard facts (not in 1-10; no references given)

Next chapter: 2-06 The harmonic oscillator and creation and annihilation. Imaginary time continues in 2-07, and the main treatment of the definition of the gap is in 4-11.

The harmonic oscillator and creation and annihilation — the levels are equally spaced, and the quanta of a field are counted as particles

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That the levels $\hbar\omega(n+\frac{1}{2})$ are determined by the algebra of creation and annihilation alone.
That the gap sets the speed of oscillation and of the decay of correlations, and that with two valleys it is invisible to perturbation theory.
That quantising a field makes particles appear, and that the gap of a free field of mass m is m.

Prerequisite chapters — 2-05 (states, spectra, the gap, the double well), 2-02 (normal modes, the mass term), 2-04 (the $E^2 = p^2 + m^2$ used in section 08, the imaginary time used in section 04), 1-03 (commutators). The ladder of 1-04 appears only for comparison.

Said first — Only single-particle quantum mechanics and fields without interactions are treated. The gap m of section 08 is a value put in by hand, and is a different thing from the mass gap of Yang–Mills (4-11). This chapter shows nothing about the mass gap.

Order of this chapter

1. Why the oscillator — near an equilibrium point everything is quadratic
2. Creation and annihilation — algebra determines the levels
3. The ground state and zero-point energy — a Gaussian (Figure 1)
4. Time evolution — the gap is the speed of oscillation, and in imaginary time the rate at which correlations decrease (Figure 2)
5. Distorting the oscillator — the gap moves (Table 1)
6. Detour: two valleys — seen from the side of the perturbation series (Figure 3, Table 2)

7. Many oscillators and Fock space — counting by occupation numbers
8. Quantising a field — particles appear, and the gap is m (Figure 4, Table 3)
9. The sum of zero-point energies, and a caution about infinitely many — only differences can be measured

---

01

Why the oscillator — near an equilibrium point everything is quadratic

Quantum field theory begins as a collection of harmonic oscillators. Near an equilibrium point x_0 any potential energy can be approximated as $V(x_0) + \frac{1}{2}V''(x_0)(x - x_0)^2$, and the string too was a collection of independent simple oscillations (2-02). The position x and the momentum $p = -i\hbar\, d/dx$ are self-adjoint operators on $L^2(\mathbb{R})$, with $[x, p] = i\hbar$ (2-05).

$$H = \frac{p^2}{2m} + \frac{m\omega^2}{2}x^2 \quad (1)$$

Measuring length in $\ell = \sqrt{\hbar/m\omega}$ and energy in $\hbar\omega$, and putting $\xi = x/\ell$, $\eta = p\ell/\hbar$, we get $H = \frac{\hbar\omega}{2}(\xi^2 + \eta^2)$, $[\xi, \eta] = i$. **Example.** Computed For a spring with 1 g and 1 Hz, $\ell \approx 1.3 \times 10^{-16}$ m and $\hbar\omega \approx 6.6 \times 10^{-34}$ J, a value 13 orders of magnitude smaller than $k_B T \approx 4.1 \times 10^{-21}$ J at room temperature.

02

Creation and annihilation — algebra determines the levels

The eigenvalues of (1) can be found by algebra alone, without solving a differential equation ($a^\dagger$ is the adjoint of a).

$$a = \frac{\xi + i\eta}{\sqrt{2}}, \qquad a^\dagger = \frac{\xi - i\eta}{\sqrt{2}} \quad (2)$$

Since $[a, a^\dagger] = 1$, $a^\dagger a = \frac{1}{2}(\xi^2 + \eta^2 - 1)$,

$$H = \hbar\omega\Big(N + \frac{1}{2}\Big), \qquad N = a^\dagger a \quad (3)$$

From $[N, a] = -a$, $[N, a^\dagger] = a^\dagger$: if $N\psi = \nu\psi$, then $N(a\psi) = (\nu - 1)a\psi$, $N(a^\dagger\psi) = (\nu + 1)a^\dagger\psi$; a lowers by 1 and $a^\dagger$ raises by 1. Since $\|a\psi\|^2 = \langle\psi, N\psi\rangle = \nu\|\psi\|^2 \geq 0$, if ν is not an

integer we keep going down and reach a contradiction; if it is an integer we stop at $a\psi_0 = 0$ with $\nu = 0$. Going up never stops, by $\|a^\dagger\psi\|^2 = (\nu+1)\|\psi\|^2 > 0$.

$$E_n = \hbar\omega\left(n + \frac{1}{2}\right), \qquad n = 0, 1, 2, \ldots \tag{4}$$

Known The levels are **equally spaced**, and the gap is $\hbar\omega$. Normalising, $a^\dagger\psi_n = \sqrt{n+1}\,\psi_{n+1}$, $a\psi_n = \sqrt{n}\,\psi_{n-1}$. This is the same argument as the ladder of 1-04, but that was finite-dimensional, stopping at both ends. No finite matrices satisfy $[a, a^\dagger] = 1$ (the trace, 0 against the dimension, does not match), so **this algebra is realised only in infinite dimensions.**

03

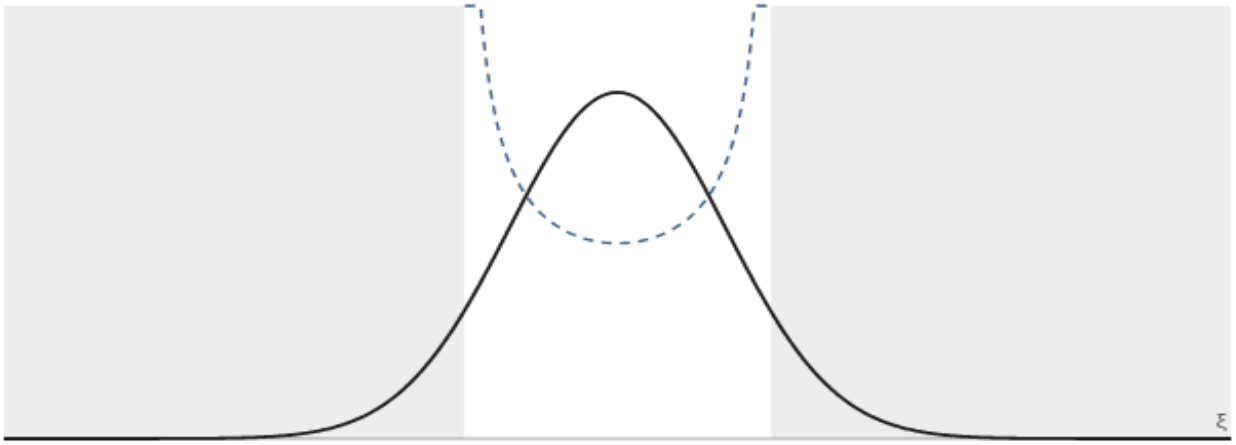
The ground state and zero-point energy — a Gaussian

In coordinates $a\psi_0 = 0$ is the first-order equation $(\xi + \frac{d}{d\xi})\psi_0 = 0$, and its solution is unique up to a constant multiple (no degeneracy).

$$\psi_0(\xi) = \pi^{-1/4} e^{-\xi^2/2}, \quad \psi_n = \frac{(a^\dagger)^n}{\sqrt{n!}}\,\psi_0 \tag{5}$$

The ground state is a **Gaussian**, and $\langle x^2\rangle = \hbar/2m\omega$, $\Delta x\,\Delta p = \hbar/2$ is exactly the lower limit of the uncertainty relation. We call $E_0 = \hbar\omega/2$ the **zero-point energy**; it cannot be made 0 because position and momentum cannot both be brought towards 0 at the same time. Precisely, $\langle H\rangle \geq \omega\sqrt{\langle p^2\rangle\langle x^2\rangle} \geq \omega\,\Delta x\,\Delta p \geq \hbar\omega/2$. ψ_n is the product of a Hermite polynomial and $e^{-\xi^2/2}$.

Figure 1 Changing n changes the probability density $|\psi_n|^2$ (solid) and the fraction of time spent by a classical particle of the same energy (dashed). Grey is the region a classical particle cannot enter



n = 0, E = 0.5 ħω, ⟨ξ²⟩ = 0.5, probability in the forbidden region = 0.1573

Computed The curves are computed on the spot from the recurrence $\psi_{n+1} = \sqrt{\frac{2}{n+1}}\,\xi\psi_n - \sqrt{\frac{n}{n+1}}\,\psi_{n-1}$ (which follows from (2) and the ladder relations) and the classical density $1/(\pi\sqrt{2n+1-\xi^2})$.

Computed The probability of being in the forbidden region, found by numerical integration, is 0.1573, 0.1116, 0.0951, 0.0740, 0.0602, 0.0425 for $n = 0, 1, 2, 5, 10, 30$ ($n = 0$ is $\operatorname{erfc}(1) = \frac{2}{\sqrt{\pi}}\int_1^\infty e^{-t^2}\,dt$). For large n, $|\psi_n|^2$ with its wiggles averaged out approaches the classical density.

04

Time evolution — the gap is the speed of oscillation

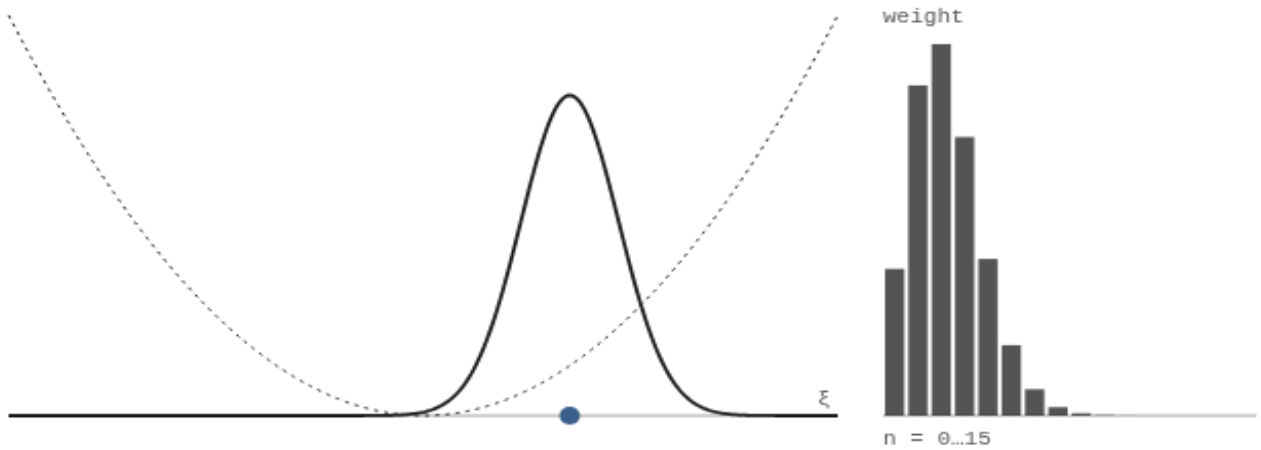
An eigenstate only has its phase turning, and its density does not move. For $\psi = c_0\psi_0 + c_1\psi_1$, with $\Delta = E_1 - E_0$ (the diagonal terms are 0 by parity),

$$\langle x\rangle(t) = 2\,\mathrm{Re}\big(c_0 c_1 e^{-i\Delta t/\hbar}\big)\langle\psi_0, x\psi_1\rangle \quad (6)$$

The gap is the speed at which observable quantities oscillate, and for the oscillator it is the classical ω. The weights of a **coherent state** $\psi_\alpha = e^{-|\alpha|^2/2}\sum_n \frac{\alpha^n}{\sqrt{n!}}\psi_n$ ($a\psi_\alpha = \alpha\psi_\alpha$) follow a Poisson distribution with mean $|\alpha|^2$, and if α is real the density keeps its shape and goes back and forth along the classical orbit.

$$|\psi_\alpha(\xi,t)|^2 = \pi^{-1/2} e^{-(\xi - \sqrt{2}\alpha\cos\omega t)^2} \quad (7)$$

Figure 2 Changing α and pressing "Play", the density of the coherent state (solid) goes back and forth with the same period as the classical particle (dot). The dotted line is the potential energy, and the bars on the right are the weights of the levels n



α = 1.5, mean number of quanta = 2.25, E = 2.75 ħω

Computed Computed on the spot from formula (7) and the Poisson distribution.

In imaginary time, the gap is the rate at which correlations decrease

We compute (8) of 2-05 with the $A = x$ of the oscillator. From (2), $x = \ell\,(a + a^\dagger)/\sqrt{2}$, so $x\psi_0 = \frac{\ell}{\sqrt{2}}\psi_1$, and there are no other components.

$$G(\tau) = \langle x\psi_0,\ e^{-\tau(H-E_0)/\hbar}\,x\psi_0\rangle = \frac{\ell^2}{2}\,e^{-\omega\tau} = \frac{\hbar}{2m\omega}\,e^{-\omega\tau} \quad (8)$$

The rate of decrease is the gap $(E_1 - E_0)/\hbar = \omega$, and (8) of 2-05 is an equality for the oscillator. For a general V it becomes $G(\tau) = \sum_{n\geq 1} |\langle \psi_n, x\psi_0\rangle|^2 e^{-(E_n-E_0)\tau/\hbar}$, and the term $n = 1$ remains at $\tau \to \infty$ (when $\langle \psi_1, x\psi_0\rangle \ /= 0$). **Correlations in imaginary time decrease at the rate of the gap.** Measuring glueball masses on the lattice (5-08) and restating the mass gap as the decay of correlations (4-11) both have this form (for the meaning of $t = -i\tau$ see 2-04; for the path integral, 2-07).

05

Distorting the oscillator — the gap moves

We add a quartic term ($\hbar = m = \omega = 1$).

$$H = \frac{p^2}{2} + \frac{x^2}{2} + \lambda x^4 \quad (9)$$

$x^4 = \frac{1}{4}(a + a^\dagger)^4$ moves n by up to 4, so the ladder no longer closes. At first order of perturbation (replacing by the expectation value), $E_n \approx n + \frac{1}{2} + \frac{3\lambda}{4}(2n^2 + 2n + 1)$, and the gap is $1 + 3\lambda$.

Computed **Table 1** The levels of formula (9) found by finite differences (the change on halving the step is within 3×10^{-6}).

λ	Gap E_1-E_0	First-order estimate 1+3λ	E_2-E_1
0	1.0000	1	1.0000
0.01	1.0284	1.03	1.0552
0.1	1.2104	1.3	1.3691
1	1.9341	4	2.4414
10	3.8166	31	5.0254

The estimate fits at $\lambda = 0.01$ and fails at $\lambda = 1$. **The spacing widens further up**, and the gap stays positive. Known In one dimension, if V is continuous and $V \to \infty$ as $|x| \to \infty$, the spectrum consists only of eigenvalues (discrete spectrum, 1-10), and no eigenvalue is degenerate. Why there is no degeneracy can be said briefly. For eigenfunctions ψ, χ with the same E, the Wronskian $W = \psi\chi' - \psi'\chi$ is constant by the

equation, $W' = 0$, and is 0 far away, so it is identically 0; that is, ψ and χ are proportional. The proof of discreteness is in the textbook in the table of sources (for the ground state see section 04 of 2-05).

06

Detour: two valleys — seen from the side of the perturbation series

A section you may skip.

We look again at the double well of 2-05 from the side of the perturbation series of the oscillator ($\hbar = m = 1$).

$$V(x) = \frac{1}{2}x^2\left(1 - \sqrt{g}\,x\right)^2 = \frac{x^2}{2} - \sqrt{g}\,x^3 + \frac{g}{2}\,x^4 \quad (10)$$

This is (9) of 2-05 with $g = 1/4a^2$, shifted by a ($V(x) = V_a(x - a)$; the valleys are at 0 and $1/\sqrt{g}$), and has the form of equation (2.4) of Zinn-Justin–Jentschura (2004). The bottom of the left valley is the oscillator of section 01, and the remainder $-\sqrt{g}\,x^3 + \frac{g}{2}x^4$ is expanded as a perturbation in powers of g, as in section 05. Starting from the right valley via the reflection $x \mapsto 1/\sqrt{g} - x$ gives the same computation, and the lowest two levels both have the following series as their asymptotic expansion Known (equation (2.6) of that paper).

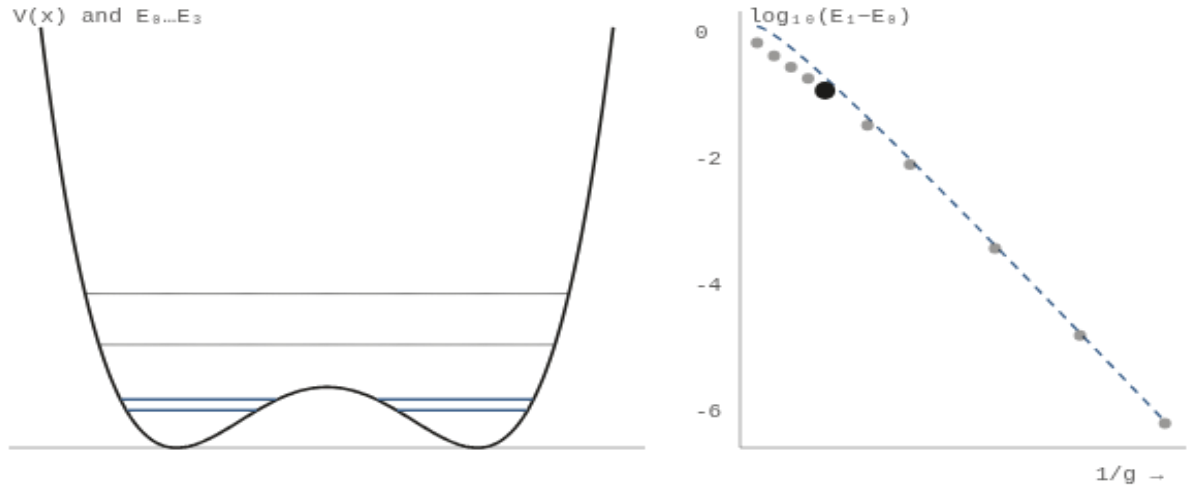
$$E \sim \frac{1}{2} - g - \frac{9}{2}g^2 - \frac{89}{2}g^3 - \frac{5013}{8}g^4 - \cdots$$

The coefficients grow factorially, and **at every order the gap is 0**. The gap lies outside the series; it is the following formula, (10) of 2-05 written in g (equation (2.8) of that paper).

$$E_1 - E_0 \approx \frac{2}{\sqrt{\pi g}}\,e^{-1/(6g)} \quad (11)$$

The $\frac{1}{6g}$ in the exponent is the $S_0 = 2a^2/3$ of 2-05, and is said to be the action of the path that goes from valley to valley by passing under the hill. Precisely, it is the action of the classical solution (instanton, 3-05) that rolls from hilltop to hilltop of $-V$ in imaginary time.

Figure 3 Changing 1/g changes the two valleys (left) and the lowest four levels. On the right is the common logarithm of the gap $E_1 - E_0$ (dots: numerical, dashed: formula (11)). As the valleys move apart the gap shrinks exponentially




```
1/g = 20, height of hill = 0.625, E₀ = 0.38510, E₁ = 0.49649, gap =1.114e-1 (formula (11): 1.800e-1)
```

Computed The levels are values found on this machine; the dashed line is formula (11).

Computed The partial sums only draw towards the **mean** of the two levels, and do not separate them. The partial sums up to five terms draw, for $1/g = 40$, as 0.5, 0.475, 0.472188, 0.471492, 0.471247 towards the mean 0.471054 ($E0 = 0.467282$, $E1 = 0.474826$), and for $1/g = 100$, as 0.5, 0.49, 0.48955, 0.4895055, 0.4894992 towards the mean 0.4894978 ($E0 = 0.4894975$, $E1 = 0.4894981$).

Computed **Table 2** The ratio of the gap to formula (11) on the low-hill side. The rightmost column is the ratio to the formula multiplied by the next correction $(1 - \frac{71}{12}g)$ of the same paper (the coefficient of equation (2.19a)). $1/g = 16$–100 are in Table 4 of 2-05.

1/g	Height of hill	Gap E_1−E_0	Ratio	Ratio with correction
20	0.625	1.114e−1	0.619	0.879
30	0.938	3.153e−2	0.757	0.943
40	1.250	7.544e−3	0.831	0.975

Physics The gap of Yang–Mills too is held to be proportional to the scale Λ, and Λ depends on the coupling constant in the form e^{-c/g^2} ($c > 0$; g is the coupling constant at the scale μ, with a power of g in front), so it is invisible to perturbation theory (4-07, 4-12).

07

Many oscillators and Fock space — counting by occupation numbers

If there are K independent oscillators, $[a_k, a_l^\dagger] = \delta_{kl}$, $[a_k, a_l] = 0$, and

$$H = \sum_{k=1}^{K} \hbar\omega_k \left(a_k^\dagger a_k + \frac{1}{2} \right) \quad (12)$$

The eigenstates are labelled by the tuple of **occupation numbers** $(n_1, \dots, n_K)$, and n_k is "the number of quanta in mode k". All zeros is the **vacuum** $|0\rangle$, and $a_k^\dagger|0\rangle$ is a one-quantum state. Since $a_k^\dagger a_l^\dagger|0\rangle = a_l^\dagger a_k^\dagger|0\rangle$, quanta carry no name tags; only "how many in which mode" has meaning. Quantising the chain of 2-02 gives this form.

The form that does not depend on the number of oscillators is **Fock space**. Let the one-quantum space be $\mathfrak{h}$ (here $\mathbb{C}^K$; for a field, a space of functions). For $\mathfrak{h} = \mathbb{C}^K$, $\mathrm{Sym}^r \mathfrak{h}$ may be read as the space whose orthonormal basis is the tuples of occupation numbers above summing to r, $(n_1, \dots, n_K)$ (precisely, it is the symmetrisation of the tensor product of r copies). If there are infinitely many modes, we take as basis the tuples with finitely many non-zero n_k summing to r.

$$\mathcal{F}(\mathfrak{h}) = \bigoplus_{r=0}^{\infty} \mathrm{Sym}^r \mathfrak{h}, \quad \mathrm{Sym}^0 \mathfrak{h} = \mathbb{C}\,|0\rangle \tag{13}$$

The direct sum is the set of all sequences whose sums of squared norms are finite, and the total number of quanta $\hat{N} = \sum_k a_k^\dagger a_k$ has value r on $\mathrm{Sym}^r \mathfrak{h}$. **Example.** The number of ways to distribute r quanta over K modes is $\binom{r+K-1}{r}$, and $\dim \mathrm{Sym}^2 \mathbb{C}^3 = 6$. Known The method of treating light with creation and annihilation is said to go back to Dirac (1927), and this space to Fock (1932).

08

Quantising a field — particles appear, and the gap is m

On a ring of length L (periodic boundary), we quantise a field with a mass term m (2-02 § 06) ($\hbar = c = 1$). Placing an oscillator for each mode e^{ikx} ($k = 2\pi n/L$, $\omega_k = \sqrt{k^2 + m^2}$; the dispersion relation $\omega^2 = c^2k^2 + \Omega^2$ of 2-02 with $c = 1$, $\Omega = m$),

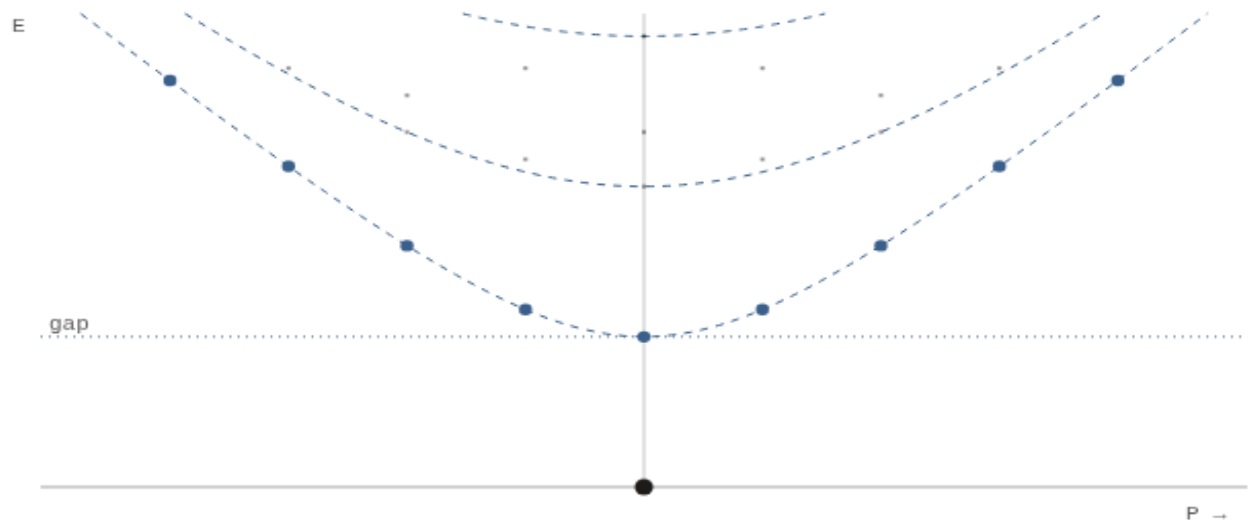
$$\phi(x) = \sum_k \frac{a_k e^{ikx} + a_k^\dagger e^{-ikx}}{\sqrt{2L\omega_k}} \tag{14}$$

$$H = \sum_k \omega_k \left(a_k^\dagger a_k + \frac{1}{2} \right) \tag{15}$$

The total momentum is $P = \sum_k k\, a_k^\dagger a_k$. One quantum of mode k carries energy $\sqrt{k^2 + m^2}$ and momentum k, which is exactly the relation $E^2 = p^2 + m^2$ of a particle of mass m (2-04). **A particle is a quantum of a normal mode of the field.** In the (P, E) plane (Figure 4) a single particle lies on the hyperbola $E = \sqrt{P^2 + m^2}$, and two or more particles fill the side $E \geq \sqrt{P^2 + 4m^2}$. Next after the vacuum is the single particle with $k = 0$, and **even when the box is widened the gap is** m. Known In infinite volume k ranges over all real numbers, the single-particle energies $\sqrt{k^2 + m^2}$ cover $[m,$

$\infty)$, and multi-particle states lie in $[2m, \infty)$, so the spectrum of the Hamiltonian is $\{0\} \cup [m, \infty)$. For $m = 0$ the gap $2\pi/L$ shrinks to 0 (excluding $k = 0$, which is not an oscillator).

Figure 4 Changing L and m changes the momentum P (horizontal) and energy E (vertical) of all states (E $\leq$ 3) of the free field in a box. Large dots are single particles. The dashed lines are $E = \sqrt{P^2 + (jm)^2}$ (j = 1, 2, 3; for m = 0, $E = |P|$), and the dotted line is the gap. When there are more than 10,000 states, the dots for two or more particles are omitted



L = 10.0, m = 1, number of states with E ≤ 3 = 23, gap =1.0000

Computed The dots are tuples of occupation numbers enumerated on the spot (the same procedure as Table 3).

Computed **Table 3** The number of states with $E \leq 3$ and the gap, for the free field in a box of period L.

L	m = 1: gap	Number of states	m = 0: gap	Number of states
5	1.0000	10	1.2566	8
10	1.0000	23	0.6283	38
20	1.0000	69	0.3142	734
40	1.0000	219	0.1571	55535

The gap of a free field, and the gap of the problem statement

The m here is a value put in by hand. The action of Yang–Mills has no mass term (3-02), and self-interaction makes the modes no longer independent. Physics Even so, a gap is thought to arise, and showing it as mathematics is the content of the problem (0-03, 4-11). The way of saying that the uncertainty relation of § 03 is the origin of the gap is corrected in 16-08 § 09. What the theorem gives is only the step that converts a length into an energy, and the length side is this problem.

09

The sum of zero-point energies, and a caution about infinitely many — only differences can be measured

The vacuum energy $\sum_k \omega_k/2$ of (15) diverges. Cutting off at $|k| \le \Lambda$ and putting $\Lambda \gg m$,

$$\sum_{|k|\le\Lambda} \frac{\omega_k}{2} \approx \frac{L}{2\pi}\int_{-\Lambda}^{\Lambda} \frac{\omega_k}{2}\,dk \approx \frac{L\Lambda^2}{4\pi} \quad (16)$$

It is proportional to the volume and grows like Λ^2. Since what can be measured is differences, we discard the constant and put $H = \sum_k \omega_k a_k^\dagger a_k$ (**normal ordering**). The gap is a difference, so it does not change. Physics The **difference** of vacuum energies when the box is changed can remain finite, and the Lüscher term is held to be an example (17-03).

Providing operators satisfying $[a_k, a_l^\dagger] = \delta_{kl}$ on a Hilbert space is called a **representation** of the commutation relations. Known For finitely many oscillators, when the commutation relations are written in exponential form (the Weyl relations; for one oscillator $e^{isx}e^{itp} = e^{-i\hbar st}e^{itp}e^{isx}$), the unitary irreducible representations (those whose only invariant closed subspaces are trivial) are unique up to unitary equivalence (being carried into each other by a unitary map) (Stone–von Neumann). For infinitely many, representations not unitarily equivalent to the Fock representation can be built. For example, the state with one quantum in every mode has an infinite sum of occupation numbers and is not in Fock space, and another representation with it as the vacuum can be considered. Fields with interactions are treated together with Haag's theorem in 4-09.

What this chapter can say

Content	Status
Sections 01–04: the levels and states of the oscillator, the decay of correlations	Said Known
Section 05: the levels and gap of the distorted oscillator	Said as numbers Computed . Discreteness and non-degeneracy are Known
Section 06: the perturbation series common to the two levels and formula (11)	Formulas from the literature Known . Partial sums and the gap checked against numbers Computed . Not derived
Sections 07–08: Fock space, the spectrum {0} ∪ [m, ∞)	Said Known (Table 3 Computed)

Content	Status
Representations of infinitely many oscillators	Introduced only (4-09)
The mass gap of Yang–Mills	This chapter shows nothing

Sources and reproduction

Item	Kind	Source / tools
Tables 1–3, Figures 1–4, the numbers in the text	Computed	python3, NumPy, SciPy (eigenvalues by finite differences, numerical integration, enumeration). The figures by JavaScript in the browser
The method of treating radiation with creation and annihilation	Known	P. A. M. Dirac, Proc. R. Soc. Lond. A 114 (1927) 243–265, doi:10.1098/rspa.1927.0039. Bibliographic data only
Fock space	Known	V. Fock, Z. Phys. 75 (1932) 622–647, doi:10.1007/BF01344458. Bibliographic data only
Two valleys: formulas (10)(11), the series, the correction 71/12	Known	J. Zinn-Justin, U. D. Jentschura, Ann. Phys. 313 (2004) 197–267, doi:10.1016/j.aop.2004.04.004, arXiv:quant-ph/0501136 (equations (2.4)(2.6)(2.8)(2.19a)(2.20) of § 2.1 checked). The introduction and numbers for the same system are in 2-05
That the spectrum is discrete in one dimension when $V \to \infty$	Known	M. Reed, B. Simon, *Methods of Modern Mathematical Physics IV: Analysis of Operators* (Academic Press), chapter XIII. Bibliographic data only (year of publication and the relevant section not checked)
The spectrum $\{0\} \cup [m, \infty)$ of the free field	Known	M. Reed, B. Simon, *Methods of Modern Mathematical Physics II: Fourier Analysis, Self-Adjointness* (Academic Press, 1975), the section on the free relativistic field (said to be § X.7; section number not checked). Bibliographic data only
Stone–von Neumann uniqueness	Known	J. von Neumann, Math. Ann. 104 (1931) 570–578, doi:10.1007/BF01457956. Bibliographic data only. Infinitely many oscillators and Haag's theorem are treated in 4-09

Next chapter: 2-07 The idea of the path integral.

The idea of the path integral — add up every path, and making time imaginary turns it into probability

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That an amplitude can be written as the limit of discretised multiple integrals, "the sum of $e^{iS/\hbar}$ over all paths".
That making time imaginary makes the weight $e^{-S_E/\hbar}$ positive, giving a probability measure, the Wiener measure.
That the partition function becomes an integral over paths on a ring, and quantum mechanics takes the same form as one-dimensional classical statistical mechanics.

Prerequisite chapters — 2-05 (decay in imaginary time), 1-08 (probability, the Gaussian distribution). For the oscillator, 2-06.

Said first — We treat a single particle on a line. The real-time path integral is not a measure, and many of the formulas are read as limits of discretised integrals. We do not go into the construction of path integrals for fields (4-02, Part 7). This chapter shows nothing about the mass gap of Yang–Mills.

Order of this chapter

1. The idea of a sum over paths — add the amplitudes, then square
2. Gaussian integrals — the value is set by the determinant, the correlations by the inverse matrix
3. Discretising time — building discretised integrals from the propagator and the product formula
4. Oscillating integrals and the classical path — cancellation selects the stationary point (Figure 1)
5. Imaginary time — oscillation turns into a weight, and the Wiener measure appears
6. The partition function is paths on a ring — quantum mechanics takes the same form as a one-dimensional classical chain (Figure 2)

7. Computing the oscillator by discretising — the correlation length grows like 1/a (Figure 3)
8. Finding the levels of x^4 with the transfer matrix — even without Gaussians we can compute with products of matrices
9. To fields — the fluctuation at a point diverges for d $\geq$ 2 (tables and figures in 1-09)

01

The idea of a sum over paths — add the amplitudes, then square

The path integral rewrites quantum mechanics as the sum of contributions from **every path** the particle could take. Feynman (1948) wrote the amplitude (the **propagator**) for a particle leaving a at time 0 to be found at b at time t in the following form.

$$K(b,t;a,0) = \langle b|e^{-itH/\hbar}|a\rangle = \text{“}\sum_{\text{paths } x(\cdot)}\text{”}\, e^{iS[x]/\hbar}, \qquad S[x] = \int_0^t \Big(\frac{m}{2}\dot{x}^2 - V(x)\Big)\,ds \quad (1)$$

S is the action of 2-01, and the paths are all those satisfying $x(0) = a$, $x(t) = b$. Every path brings a complex number of absolute value 1, differing only in phase, and the probability is the square of the absolute value of the sum Known . The idea is said to go back to Dirac (1933). As a metaphor: just as light passes through both slits and interferes, the particle passes along every path and interferes. Precisely, the "sum" in (1) is not defined as it stands, and is read as the limit of the discretised integrals of section 03.

02

Gaussian integrals — the foundation of every computation

Many computable path integrals reduce to Gaussian integrals. Let A be a positive-definite symmetric matrix; diagonalising it and using the one-variable $\int e^{-ax^2/2+Jx}\,dx = \sqrt{2\pi/a}\,e^{J^2/2a}$ for each eigenvalue,

$$\int_{\mathbb{R}^n} e^{-\frac{1}{2}x\cdot Ax + J\cdot x}\,d^n x = \frac{(2\pi)^{n/2}}{\sqrt{\det A}}\, e^{\frac{1}{2}J\cdot A^{-1}J} \quad (2)$$

Normalised, this is the Gaussian distribution with mean 0 and covariance matrix A^{-1} (1-08). Differentiating with respect to J and putting $J = 0$,

$$\langle x_i x_j\rangle = (A^{-1})_{ij}, \qquad \langle x1x2x3x4\rangle = \langle x1x2\rangle\langle x3x4\rangle + \langle x1x3\rangle\langle x2x4\rangle + \langle x1x4\rangle\langle x2x3\rangle \quad (3)$$

The expectation of a product of $2k$ factors is the sum over the $(2k-1)!!$ ways of pairing them off (Isserlis 1918) Known . In physics the same combinatorics appears in operator form in Wick (1950), and goes by the name of Wick's theorem. Drawing these pairings as lines gives Feynman diagrams (4-03). **Example.** For $A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ the integral is $2\pi/\sqrt{3}$, $\langle x_1^2 \rangle = \frac{2}{3}$, $\langle x_1 x_2 \rangle = \frac{1}{3}$, $\langle x_1^2 x_2^2 \rangle = \frac{4}{9} + 2 \cdot \frac{1}{9} = \frac{2}{3}$ Computed . **The value is set by** $\det A$**, and the correlations by** A^{-1}**.**

03

Discretising time — the propagator and the product formula

There are three ingredients. **Composition**: from $e^{-i(t_1+t_2)H/\hbar} = e^{-it_2H/\hbar} e^{-it_1H/\hbar}$ comes the formula $K(b, t_1+t_2; a, 0) = \int K(b, t_1+t_2; c, t_1) K(c, t_1; a, 0)\, dc$ that sums over all intermediate places c. **The short-time kernel**: for a free particle, exactly

$$K_0(b, t; a, 0) = \sqrt{\frac{m}{2\pi i\hbar t}}\, \exp\Big(\frac{i}{\hbar} \frac{m(b-a)^2}{2t} \Big) \quad (4)$$

and the exponent is the action of the uniform-velocity path. **The product formula**: split the Hamiltonian as $H = T + V$. $T = P^2/2m$ is the operator of kinetic energy ($P = -i\hbar\, d/dx$ is the momentum of 2-05), and V is the operator of multiplication by the potential energy $V(x)$. T and V do not commute, but if, for example, $T+V$ is self-adjoint on the intersection of the domains of T and V, then $e^{-it(T+V)/\hbar} = \lim_N \big(e^{-i\varepsilon T/\hbar} e^{-i\varepsilon V/\hbar} \big)^N$, $\varepsilon = t/N$ holds Known . It is a special case of Trotter's (1959) product theorem for contraction semigroups, and this form is in chapter VIII of Reed–Simon I. Putting these together, with $x_0 = a$, $x_N = b$,

$$K(b, t; a, 0) = \lim_{N\to\infty} \Big(\frac{m}{2\pi i\hbar\varepsilon} \Big)^{N/2} \int dx_1 \cdots dx_{N-1}\, \exp\Big(\frac{i}{\hbar} \sum_{j=0}^{N-1} \varepsilon \Big[\frac{m}{2} \Big(\frac{x_{j+1} - x_j}{\varepsilon} \Big)^2 - V(x_j$$

The sum in the exponent is a Riemann sum for the action of a broken-line path, that is, the discretised action of 2-01. (1) is shorthand for this limit, also written $\int \mathcal{D}x\, e^{iS/\hbar}$.

Caution — $\mathcal{D}x$ is not a measure

Cameron (1960) is said to have shown that the real-time $e^{iS/\hbar}\mathcal{D}x$ cannot be constructed as a measure in the form of the Wiener measure (section 05) with complex variance (such a measure would have infinite total variation even on bounded sets of the path space) Known . The total variation is the supremum, over fine partitions of a set, of the sum of the absolute values of the measures of the

parts; for a probability measure it is 1. Measures in infinite dimensions in general are treated in 1-09.

04

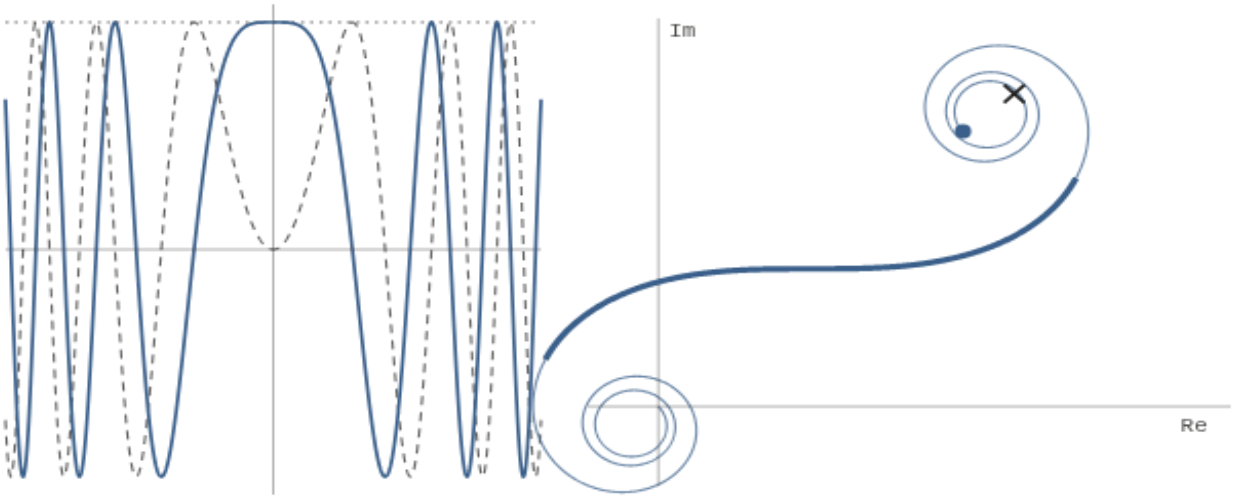
Oscillating integrals and the classical path — cancellation and the stationary point

The integrand of (5) has absolute value 1 and oscillates, and does not converge absolutely. Figure 1 shows the model case $\int e^{ix^2/2}dx = \sqrt{2\pi}\,e^{i\pi/4}$ (the Fresnel integral). On the right is the trace of the partial integral (the integral with a moving upper limit) $F(x) = \int_{-6}^{x} w\,dx'$; for $\theta = 0$ both ends spiral and cancel, and the net progress happens only near the **stationary point** $x = 0$ where the slope of the phase is 0. For $\hbar \to 0$ (the sign is the sign of $S''(x_c)$) Known

$$\int g(x)\,e^{iS(x)/\hbar}\,dx \approx g(x_c)\sqrt{\frac{2\pi\hbar}{|S''(x_c)|}}\;e^{iS(x_c)/\hbar \pm i\pi/4}, \qquad S'(x_c) = 0 \tag{6}$$

In path space too, the contribution comes from near the **classical path** with $\delta S = 0$ (2-01). If the action is quadratic, (6) is exact, and (4) has that form. **Example.** A 1 g ball moving at 1 cm/s for 1 second gives $S/\hbar \approx 4.7 \times 10^{26}$ Computed . For displaced paths the phase turns round many times, and they cancel.

Figure 1 Changing the angle θ changes $w(x) = e^{ie^{i\theta}x^2/2}$ (left; solid is the real part, dashed the imaginary part, dotted the absolute value) and the trace of the partial integral $F(x)$ (right). θ = 0 is real time and 90° is imaginary time (the $t = e^{-i\theta}\tau$ that joins the $t = -i\tau$ of 2-04 by an angle; see also the spacetime article). The thick line is |x| < 1.5, and × is the exact value



θ = 0°, F(6) = 1.517 + 1.560i, exact value = 1.772 + 1.772i, difference 3.3e-1

Computed The numbers in the figure were computed on this machine (the trapezoidal rule on the spot). Table 1 gives values from the complex error function in python3.

Computed **Table 1** The absolute value of the difference between $\int_{-X}^{X} w\,dx$ and the exact value $\sqrt{2\pi}\,e^{i(\pi/2-\theta)/2}$.

X	θ = 0	θ = 22.5°	θ = 90°
2	9.3e-1	4.1e-1	1.1e-1
8	2.5e-1	1.2e-6	< 1e-14
32	6.25e-2	< 1e-14	< 1e-14

In real time the error decreases only like $2/X$; rotating even slightly makes it vanish at the rate $e^{-\sin\theta\,X^2/2}$.

05

Imaginary time — oscillation turns into a weight

Replacing the step ε in (5) by $-i\delta$ ($\delta>0$), the exponent becomes $-\frac{\delta}{\hbar}\left[\frac{m}{2}(\Delta x/\delta)^2+V\right]$, and the sign of the potential energy changes. With $\tau=N\delta$,

$$\langle b|e^{-\tau H/\hbar}|a\rangle=\lim_{N\to\infty}\left(\frac{m}{2\pi\hbar\delta}\right)^{N/2}\int dx1\cdots dxN-1\;e^{-S_E/\hbar},\qquad S_E=\sum_{j=0}^{N-1}\delta\left[\frac{m}{2}\left(\frac{xj+1}{\delta}\right.\right.$$

We call S_E the **Euclidean action**. The weight is positive, and if V is bounded below the integral converges absolutely. The $V=0$ part is a random walk whose increments independently follow normal distributions with variance $\hbar\delta/m$; for $N\to\infty$ it becomes a probability measure on continuous paths, the **Wiener measure** (the distribution of Brownian motion), and V enters as a weight.

$$\left(e^{-\tau H/\hbar}\psi\right)(x)=\mathbb{E}_x\left[\exp\left(-\frac{1}{\hbar}\int_0^\tau V(X_s)\,ds\right)\psi(X_\tau)\right]\quad(8)$$

X_s is Brownian motion leaving x and spreading with variance $\hbar s/m$. (8) is called the Feynman–Kac formula (Kac 1949). It is a theorem when, for example, V is bounded and continuous, and it holds for quite general V bounded below Known . **In real time there is no measure; in imaginary time there is.** This is one reason to construct field theories on the Euclidean side (the rotation for fields is in 4-02 § 02 and 4-10).
Example. The stationary point $\delta S_E=0$ is motion in $-V$, and the S_0 of the double well of 2-05 was the action of that path (the instanton, 3-05).

The roughness of paths and the measure

The paths themselves are nowhere differentiable with probability 1 Known . Since $\Delta x\sim\sqrt{\delta}$, the expectation $\langle S_{\text{kin}}\rangle$ of the kinetic term $S_{\text{kin}}=\sum(\Delta x)^2/2\delta$ ($\hbar=m=1$) of S_E grows in proportion to the number of steps Computed (for the ring of section 07, 6.97,

126.96, 2046.96 for $\beta = 4$, $N = 16, 256, 4096$; almost $N/2$). The potential-energy part $\sum \delta V(x_j)$ on its own approaches $\int_0^\tau V(X_s)\,ds$, but the weight of the kinetic term $e^{-S_{\text{kin}}/\hbar}$ and $\mathcal{D}x$ have no limits separately; only the product survives, as the Wiener measure (both in 1-09).

06

The partition function is paths on a ring — the agreement with statistical mechanics

In thermal equilibrium a state n appears with weight $e^{-\beta E_n}$ ($\beta = 1/(k_B \times \text{temperature})$, k_B is Boltzmann's constant), and their sum is the **partition function** (2-08). Taking the trace means putting $a = b$ and integrating over a, so by (7) it becomes an integral over **closed paths on a ring** of length $\beta\hbar$.

$$Z(\beta) = \operatorname{Tr} e^{-\beta H} = \sum_n e^{-\beta E_n} = \int_{x(0)=x(\beta\hbar)} \mathcal{D}x\; e^{-S_E[x]/\hbar} \tag{9}$$

Read in discretised form, it is the partition function of a **classical chain** in which $x_0, \ldots, x_{N-1}$ arranged on a ring are joined by springs $\frac{m}{2\delta}(x_{j+1} - x_j)^2$ and each point carries V. Quantum mechanics takes the same form as one-dimensional classical statistical mechanics, with $\hbar$ playing the role of temperature and $\beta\hbar$ that of the length of the chain. The exposition of the problem attributes to Symanzik the view of reading the Euclidean expectations of a field theory as a classical statistical-mechanical ensemble with density proportional to e^{-S} (Jaffe–Witten § 3) Known . The chain of this section is its one-dimensional version.

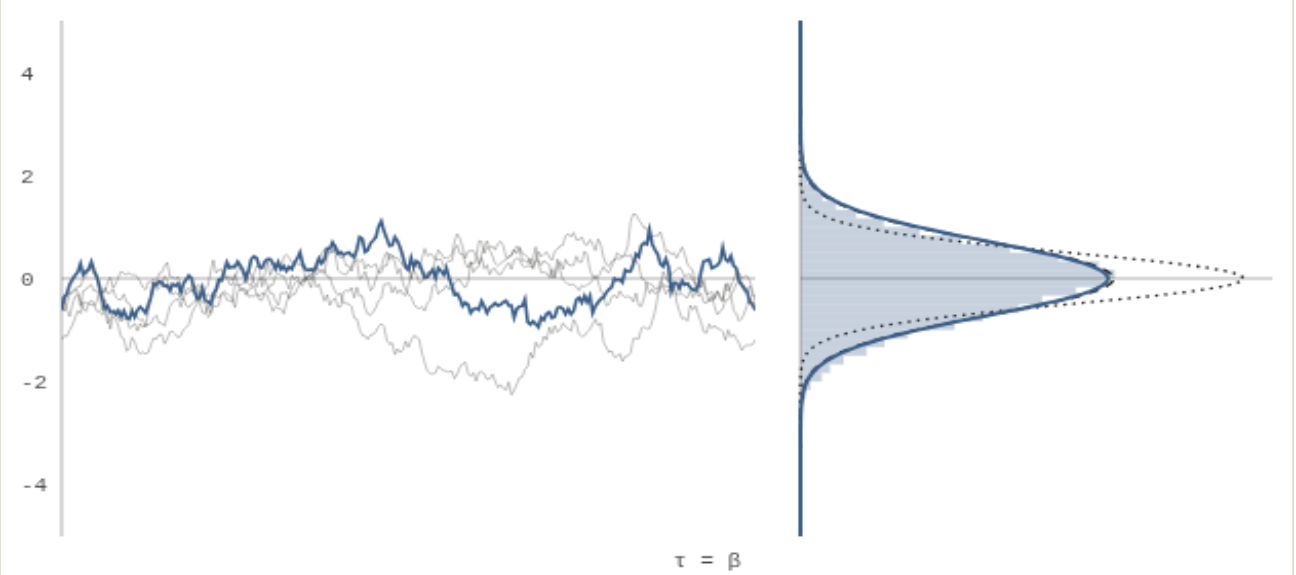
Example: the oscillator ring

For the oscillator, $Z = 1/(2\sinh\frac{\beta\hbar\omega}{2})$; the distribution of one point $x(0)$ of the ring is the distribution of position in thermal equilibrium, with variance

$$\langle x^2\rangle_\beta = \frac{\hbar}{2m\omega}\coth\frac{\beta\hbar\omega}{2} \longrightarrow \begin{cases} \hbar/2m\omega & (\beta\to\infty) \\ 1/m\omega^2\beta & (\beta\to 0) \end{cases} \tag{10}$$

and at low temperature it approaches the $|\psi_0|^2$ of the ground state (2-06), at high temperature the classical Boltzmann distribution.

Figure 2 Changing $\beta\hbar\omega$ (the length of the ring) changes the paths on the ring (left; colour is the latest) and the distribution of positions visited by the paths (shading on the right). The solid line is (10), the dashed line the ground state, the dotted line the classical case. "Play" adds paths ($\hbar = m = \omega = 1$)



βħω = 4, paths: 120, sample ⟨x²⟩ = 0.557, discretised system 0.5186, (10) 0.5187, classical 0.2500

Computed The numbers in the figure were computed on this machine. The paths are exact samples of (7) with the ring cut into 256 pieces (the real Fourier components are independent normal variables, drawn with random numbers from a fixed seed). Table 2 gives python3 values from the same procedure.

Computed **Table 2** $\langle x^2 \rangle$ at one point of the ring (unit $\hbar/m\omega$; the ground state gives 0.5).

βħω	Sample	Exact value for the discretised system	Quantum (10)	Classical
0.25	4.05 ± 0.09	4.0208	4.0208	4.0000
4	0.517 ± 0.006	0.5186	0.5187	0.2500
16	0.504 ± 0.003	0.4998	0.5000	0.0625

At high temperature, on a short ring, the path shrinks almost to a point (a classical particle); at low temperature, **quantum fluctuations are seen as fluctuations of a path stretched in the imaginary-time direction**.

07

Computing the oscillator by discretising — determinant, correlations, correlation length

$\hbar = m = \omega = 1$: cut a ring of length β into N pieces. From here on $a = \beta/N$ is the step width (the lattice spacing), a different symbol from the starting point a of sections 01–06. The A of $S_E = \frac{1}{2} x \cdot Ax$ is $(Ax)_j = \frac{1}{a}(2x_j - x_{j+1} - x_{j-1}) + a x_j$ (indices counted modulo N), a **circulant matrix** connected only to neighbours. Its eigenvectors are $v_j = e^{2\pi i j k/N}$ (discrete Fourier, $k = 0, \dots, N-1$), and substituting $v_{j\pm1} = e^{\pm 2\pi i k/N} v_j$ gives the eigenvalues $\frac{1}{a}(2 - 2\cos\frac{2\pi k}{N}) + a$. From (2) and the normalisation $(2\pi a)^{-N/2}$,

$$Z_N = \prod_{k=0}^{N-1} \left(2 - 2\cos\frac{2\pi k}{N} + a^2\right)^{-1/2} = \frac{1}{2\sinh(N\kappa/2)}, \qquad \cosh\kappa = 1 + \frac{a^2}{2} \quad (11)$$

Compared with the continuous $1/(2\sinh\frac{\beta}{2})$, the discretised system is the same as an oscillator with frequency $\Omega_a = \kappa/a = 1 - a^2/24 + \cdots$. The correlations are set by A^{-1}, and for $N \to \infty$,

$$\langle x_0 x_j \rangle = (A^{-1})_{0j} = \frac{a\, e^{-\kappa|j|}}{2\sinh\kappa} \;\longrightarrow\; \frac{e^{-\tau}}{2} \qquad (\tau = ja,\ a \to 0) \tag{12}$$

As in (8) of 2-06, correlations decrease at the rate of the gap. The **correlation length** counted in lattice units is $\xi = 1/\kappa = 1/(a\Omega_a)$.

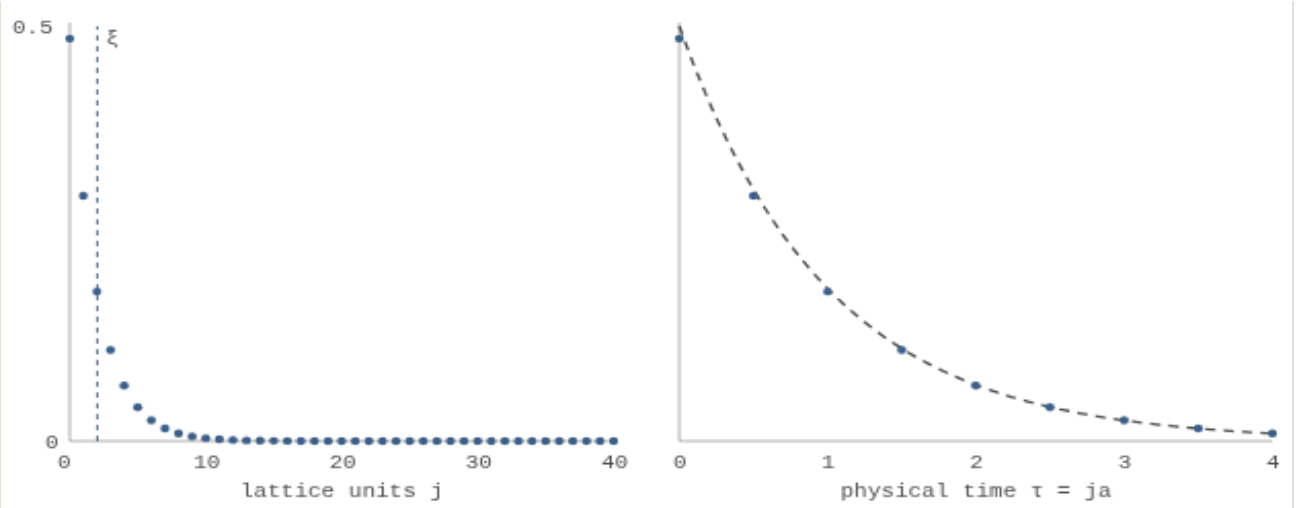
The correlation length and the continuum limit

Computed **Table 3** The ring with $\beta = 4$ (Z_N: the numerical determinant and (11) agree to 10 digits).

N	a	Z_N / Z − 1	Ω_a	ξ
4	1	8.1e-2	0.96242	1.04
16	0.25	5.4e-3	0.99741	4.01
128	0.03125	8.4e-5	0.99996	32.00

The discrepancy vanishes like a^2, and the correlation length grows like $1/a$ (Figure 3). **Approaching the continuum means bringing the lattice chain towards a state in which "the correlation length is many times the lattice spacing", in the language of statistical mechanics towards a critical point.** We hold the physical gap $1/(\xi a)$ fixed and let $\xi \to \infty$. In lattice gauge theory too, the continuum limit is thought to be taken in this form Physics . In four dimensions, the construction published in 2026-11 made the scaling limit by choosing the bare coupling to follow the running of asymptotic freedom (the committee's assessment is that it is proved in Lean; 6-00, 8-02). The form in which the correlation length grows while the physical gap is held fixed, that is, that the gap stays positive in the limit, has not been shown (6-04, 5-05, 5-09).

Figure 3 Changing the step a changes how the correlation (12) $a\,e^{-\kappa j}/(2\sinh\kappa)$ ($\beta \to \infty$) looks. On the left it is counted in lattice units j (the vertical dashed line is the correlation length $\xi = 1/\kappa$); on the right it is measured in physical time $\tau = ja$ (the dashed line is the continuous $e^{-\tau}/2$). "Play" moves a back and forth between 1 and 1/16



Computed The numbers in the figure were computed on this machine (the closed formula (12) computed on the spot; κ, ξ, Ω_a at $a = 1, \frac{1}{2}, \ldots, \frac{1}{16}$ agree with the python3 values to the digits shown). On the left, the smaller a, the further the dots continue; on the right, the dots settle onto the same dashed line.

08

A non-Gaussian example — finding the levels of x⁴ with the transfer matrix

Even without Gaussians, discretised integrals can be computed as products of matrices. Make one step of (7) ($\hbar = m = 1$, step a) into the integral kernel of $e^{-aV/2}e^{-aP^2/2}e^{-aV/2}$

$$T_a(x,y) = \frac{1}{\sqrt{2\pi a}} \exp\Big(-\frac{(x-y)^2}{2a} - \frac{a}{2}\big(V(x)+V(y)\big)\Big) \quad (13)$$

then $Z_N = \operatorname{Tr} T_a^N$; this is the **transfer matrix** of statistical mechanics. Writing the eigenvalues as $e^{-aE_n(a)}$, $E_n(a) \to E_n$ as $a \to 0$. The kernel is positive at every point ($T_a(x,y) > 0$). Moreover $V = x^4$ grows at infinity, so $\iint T_a(x,y)^2\,dx\,dy \le \frac{1}{2\pi a}\big(\int e^{-aV}\,dx\big)^2 < \infty$, and T_a is **compact** (an operator that can be approximated as closely as we like by finite-dimensional matrices, and whose eigenvalues accumulate nowhere but 0). With these two together, the integral-operator version of the Perron–Frobenius theorem (said to begin with Jentzsch 1912; the form for L^2 is in Reed–Simon IV) shows that the largest eigenvalue is simple and its eigenfunction can be taken positive Known . Hence the ground state is non-degenerate (2-05, 5-03).

Computed **Table 4** The levels of the transfer matrix for $H = -\frac{1}{2}\frac{d^2}{dx^2} + x^4$ (lattice width for $x \le 0.02$, $|x| \le 3.2$) and their differences from the values of Table 1 of 2-05 ($E_0 = 0.667986$, $E_1 - E_0 = 1.725658$). Halving the lattice width changes nothing to 9 digits.

a	$E_0(a)$	$E_1(a) - E_0(a)$	Difference in E_0	Difference in gap
0.1	0.666246	1.719680	−1.7e-3	−6.0e-3

a	$E_0(a)$	$E_1(a) - E_0(a)$	Difference in E_0	Difference in gap
0.05	0.667546	1.724144	−4.4e-4	−1.5e-3
0.025	0.667876	1.725278	−1.1e-4	−3.8e-4

The difference is roughly 1/4 each time the step is halved. **The discretised path integral and operator quantum mechanics give the same levels and gap.**

09

To fields — as the step is refined, the fluctuations diverge

We extend the discretisation of section 07 directly to fields. In field theory the path $x(\tau)$ is replaced by a field $\phi(x)$ (d is the dimension of spacetime including imaginary time), and $S_E = \int (\frac{1}{2}|\nabla\phi|^2 + \frac{1}{2}m^2\phi^2 + \cdots)d^dx$ (m is here the mass of the field; 4-02). Discretising on a lattice of spacing a, a field without interactions is again the Gaussian integral of (2), and the fluctuation at a point is a diagonal entry of A^{-1}.

$$\langle \phi(0)^2 \rangle = G_a(0) = \int_{[-\pi/a,\pi/a]^d} \frac{d^dp}{(2\pi)^d} \frac{1}{\sum_{\mu=1}^{d} \frac{4}{a^2}\sin^2\frac{p_\mu a}{2} + m^2} \quad (14)$$

$d = 1$ is the chain of section 07 itself; (14) equals the value $a/(2\sinh\kappa)$ of (12) at $j = 0$, and converges to $1/2m$. For $d \geq 2$ it diverges. The divergence of the variance at a point is also treated in 1-09, from the side of Gaussian measures in infinite dimensions. Here we look, as a continuation of discretising the path integral, at what happens each time the step is halved.

Computed Each time the step is halved, $d = 1$ converges to $1/2$, $d = 2$ grows by about $\ln 2/(2\pi) \approx 0.110$ each time ($G_a(0) \approx \frac{1}{2\pi}\log\frac{1}{a} + 0.276$), $d = 3$ doubles ($\approx 0.2527/a$), and $d = 4$ quadruples ($\approx 0.1549/a^2$) ($m = 1$; (14) is turned into a one-dimensional integral of the Bessel function I_0 and integrated numerically). The table of values for each step and the figure for choosing the dimension are in 1-09 Table 3, Figure 3, which this book takes as authoritative. A figure setting it beside other quantities that switch at four is the dimension dial of 1-12 § 10.

Fields in the continuum limit and Yang–Mills

A field in the continuum limit is a **distribution** without values at points (1-11). Products such as $\phi(x)^2$ need subtractions (renormalisation) (4-04, 7-03).

In four-dimensional Yang–Mills this divergence, the gauge freedom (4-08) and the interaction are handled at once. The exposition of the problem says that for a compact group and a finite (periodic) spacetime, Wilson's lattice makes the functional integral a finite-dimensional integral, and that one must then show that the limit lattice spacing

$\to 0$, volume $\to \infty$ exists for the expectations of gauge-invariant observables (Jaffe–Witten § 6.5) Known . The construction published in 2026-11 takes the limit lattice spacing $\to 0$ with the volume held fixed, and then takes the limit along a subsequence of volumes (8-05). For what is special about four dimensions, see the four-dimensions article.

What this chapter can say

Content	Status
The formulas (2)–(12)	Said Known ((5) and (8) under conditions; numbers also checked in Tables 2 and 3 and Figures 2 and 3 Computed). That $\mathcal{D}x$ is not a measure is a report of the literature
The convergence in Table 4, the constants in the growth of the divergence in § 09	The range checked numerically Computed
That the continuum limit of lattice gauge theory is also taken in the form "correlation length $\to \infty$" (section 07)	A view at the level of physics Physics . The scaling limit in four dimensions was made by the construction published in 2026-11 (at the level of the committee's assessment; 6-00). The form in which the gap stays positive has not been shown (6-04)
Path integrals for interacting fields and gauge fields	Not treated (4-02, Part 7). Nothing is shown about the mass gap of Yang–Mills

Sources and reproduction

Item	Kind	Source / tools
Tables 1–4, the numbers in the examples, Figures 1–3	Computed	python3 (NumPy, SciPy). The figures by JavaScript in the browser (the trapezoidal rule, random numbers with a fixed seed, the closed formula (12))
The constants of growth in § 09 (0.2527, 0.1549, 0.276)	Computed	Computed with python3 (SciPy's I_0 and numerical integration): the values at the origin of the massless lattice ($d = 3$: 0.252727, $d = 4$: 0.154933), and the value of the difference $G_a(0) - \frac{1}{2\pi} \log \frac{1}{a}$ of $d = 2$ up to $a = 2^{-10}$ (0.27579). Agrees with Table 3 of 1-09
(1)	Known	R. P. Feynman, “Space-time approach to non-relativistic quantum mechanics”, Rev. Mod. Phys. 20 (1948) 367–387 (bibliographic data checked with Crossref; text bibliographic data only) / P. A. M. Dirac, Physik. Z. Sowjetunion 3 (1933) 64 (bibliographic data only; that it goes back to Dirac's suggestion is from the Scholarpedia article below)

Item	Kind	Source / tools
The product formula	Known	H. F. Trotter, Proc. Amer. Math. Soc. 10 (1959) 545–551 (bibliographic data only) / M. Reed, B. Simon, Methods of Modern Mathematical Physics I (Academic Press 1972), chapter VIII (bibliographic data only; said to be Theorem VIII.31, number not checked)
That $\mathcal{D}x$ is not a measure, the conditions for (8)	Known	R. H. Cameron, "A family of integrals serving to connect the Wiener and Feynman integrals", J. Math. and Phys. 39 (1960) 126–140 (bibliographic data only). Content checked in S. Albeverio, S. Mazzucchi, "Path integral: mathematical aspects", Scholarpedia 6(1):8832 (2011)
(8), (3)	Known	M. Kac, Trans. Amer. Math. Soc. 65 (1949) 1–13 / L. Isserlis, Biometrika 12 (1918) 134–139 / G. C. Wick, Phys. Rev. 80 (1950) 268–272. Bibliographic data only for all
Symanzik's view, lattices and limits	Known	A. Jaffe, E. Witten, "Quantum Yang–Mills theory" (the Clay problem description), § 3 and § 6.5 checked / K. Symanzik, "Euclidean quantum field theory", in Local Quantum Theory (R. Jost ed., Academic Press 1969) 152–226 (bibliographic data only; reference [46] of Jaffe–Witten)
The integral-operator version of Perron–Frobenius	Known	R. Jentzsch, "Über Integralgleichungen mit positivem Kern", J. reine angew. Math. 141 (1912) 235–244 (bibliographic data checked with Crossref; text bibliographic data only) / M. Reed, B. Simon, Methods of Modern Mathematical Physics IV (Academic Press 1978) (bibliographic data only; said to be § XIII.12, not checked)
Stationary phase	Known	R. P. Feynman, A. R. Hibbs, Quantum Mechanics and Path Integrals (1965). Bibliographic data only
That paths are not differentiable	Known	1-09 (citing Durrett's theorem)

Next chapter: 2-08 Statistical mechanics and phase transitions.

Statistical mechanics and phase transitions — probability weights, the length over which correlations reach, and the point where that length diverges

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — Reading thermal equilibrium as a probability distribution $e^{-\beta H}/Z$, and obtaining means and fluctuations by differentiating $\log Z$.
The correlation length, and that in the Gaussian chain the correlation length grows like $1/m$.
That phase transitions occur only in infinite systems, the universality of critical exponents, and that the continuum limit is an approach to a critical point.

Prerequisite chapters — 1-08 (probability, variance, Gaussian measures). Inverse matrices are in 1-01. The transfer matrix of section 03 also appears in 2-07 (if you have read it; this chapter defines it on the spot too).

Said first — We treat classical statistical mechanics on lattices. The three-dimensional critical exponents are numbers, not theorems, and the correspondence with field theory, except for the Gaussian chain, is a reinterpretation at the level of physics. This chapter shows nothing about the mass gap of Yang–Mills.

Order of this chapter

1. The Boltzmann distribution — temperature is a way of weighting probabilities
2. The partition function is a generating function — differentiating gives means and fluctuations
3. Correlations and the correlation length — exactly for the Gaussian chain; the inverse of the correlation length is the gap of the transfer matrix
4. Phase transitions — kinks that occur only in infinite systems
5. The mean-field model — watching the tug of war (Figure 1, Table 1). Detour: zeros of the partition function (Figure 2)

6. Critical points — critical exponents (Table 2)
7. Universality — exponents are universal, coefficients are not (Figure 3, Table 3)
8. The continuum limit — the critical point is where the lattice disappears (Figure 4, Table 4)

01

The Boltzmann distribution — temperature is a way of weighting probabilities

We call each single state of a system of many spins or particles a **configuration** σ, and its energy $H(\sigma)$. A system in thermal equilibrium is described by a probability distribution on configurations; the lower the energy of a configuration, the heavier it is, and the weight is exponential.

$$\mu_\beta(\sigma) = \frac{e^{-\beta H(\sigma)}}{Z_\beta}, \qquad Z_\beta = \sum_\sigma e^{-\beta H(\sigma)}, \qquad \beta = \frac{1}{kBT} \quad (1)$$

We call μ_β the **Boltzmann distribution** and Z_β the **partition function** (T is the temperature; below, Boltzmann's constant $kB = 1$, and β is the inverse temperature). For $\beta \to 0$ it becomes the uniform distribution, and for $\beta \to \infty$ it concentrates on the configurations of minimal energy. $\langle\cdot\rangle$ is the expectation under μ_β (1-08).

One reason for the exponential is the following variational principle Known . With the entropy $S(P) = -\sum_\sigma P \log P$,

$$-\frac{1}{\beta}\log Z_\beta = \min_P \left(\mathbb{E}_P[H] - \frac{1}{\beta} S(P) \right) \quad (2)$$

holds, and the minimum is attained only by $P = \mu_\beta$ (the difference between the two sides is the relative entropy $\frac{1}{\beta}\sum_\sigma P \log(P/\mu_\beta) \geq 0$). The left side is the **free energy**. It is a tug of war between the force that lowers the energy and the force that increases the number of configurations; at low temperature the former wins, at high temperature the latter. Precisely, in a finite system the minimum of (2) is always the single μ_β. If we consider the minimisation of (2) per site in an infinite system, at low temperature the states attaining the minimum can be several rather than one. This "splitting from one into several" is one precise form of a phase transition (the coexistence of phases) (we see it in one-variable form in section 05).

Example. If N non-interacting spins $\sigma_x = \pm 1$ are in an external field h ($H = -h\sum_x \sigma_x$), then $Z = (2\cosh \beta h)^N$, the spins are independent, and $\langle \sigma_x \rangle = \tanh \beta h$.

02

The partition function is a generating function — differentiating gives means and fluctuations

Differentiating $\log Z$ gives means and variances. In the language of probability, $\log Z\beta$ is the cumulant generating function of $-H$ (plus a constant). The cumulant generating function of a random variable X is $t \mapsto \log \mathbb{E}[e^{tX}]$; differentiating once in $t = 0$ gives the mean, twice the variance.

$$\langle H \rangle = -\frac{\partial}{\partial \beta} \log Z\beta, \qquad \operatorname{Var}(H) = \frac{\partial^2}{\partial \beta^2} \log Z\beta \geq 0 \quad (3)$$

Putting in the external field as $-hM$ with the total magnetisation $M = \sum_x \sigma_x$ gives the magnetisation and the **susceptibility** χ (the response to the external field).

$$\langle M \rangle = \frac{1}{\beta} \frac{\partial}{\partial h} \log Z, \qquad \chi = \frac{1}{N} \frac{\partial \langle M \rangle}{\partial h} = \frac{\beta}{N} \operatorname{Var}(M) \quad (4)$$

The right-hand equality is the fluctuation–response relation: the response to a push is determined by the fluctuations without the push. Since variances are non-negative, $\log Z$ is convex both in β and in h. The limit f of $f_N = -\frac{1}{\beta N} \log Z_N$ per site as $N \to \infty$ is called the **thermodynamic limit**. For short-range interactions the limit is known to exist, and convexity survives Known (for lattice spin systems, the standard argument from the fact that when the system is divided into boxes, $\log Z$ differs from the sum over boxes only by the boundary contributions [subadditivity]; Ruelle 1969 and others).

Example. For independent spins, $\chi = \beta(1 - \tanh^2 \beta h) = \beta \operatorname{Var}(\sigma_x)$. For $h = 0$, $\chi = \beta$, finite at every temperature. Without interactions nothing happens.

03

Correlations and the correlation length — how much two distant points move together

With interactions, distant points move together. The degree of this is the **connected two-point correlation** (covariance).

$$G(x, y) = \langle \sigma_x \sigma_y \rangle - \langle \sigma_x \rangle \langle \sigma_y \rangle \quad (5)$$

In a translation-invariant infinite system, $G(x, y) = G(x - y)$, and (4) is $\chi = \beta \sum_x G(x)$. **For the susceptibility to diverge, the correlations must decrease only slowly.** When correlations decrease exponentially, we call that length the **correlation length** ξ.

$$\frac{1}{\xi} = -\lim_{|x|\to\infty} \frac{\log G(x)}{|x|} \tag{6}$$

The logarithm is taken to allow polynomial prefactors. The correlation length of lattice gauge theory is defined in the same form, except that instead of spins it uses correlations of the variables of small square loops (plaquettes) (5-01, Chatterjee 2018 § 5, 5-09) Known . There are two examples.

The Ising model: at each point of $\mathbb{Z}^d$, $\sigma_x = \pm 1$, $H = -J\sum_{\langle xy\rangle}\sigma_x\sigma_y - h\sum_x \sigma_x$ ($\langle xy\rangle$ is a neighbouring pair, and $J>0$ is a ferromagnetic coupling that rewards alignment; a worked example is in 2-09).

The Gaussian chain: put a real number ϕ_x at each point of $\mathbb{Z}$, and with $\beta = 1$, $m > 0$ (the **mass**), take

$$H(\phi) = \frac{1}{2}\sum_x \Big[(\phi_{x+1}-\phi_x)^2 + m^2\phi_x^2\Big] \tag{7}$$

Restricting to an interval of length L with $\phi = 0$ outside, and taking ϕ as a column vector, we can write $H = \tfrac{1}{2}\phi^{\mathsf T} K \phi$ (K is the $L\times L$ matrix with m^2+2 on the diagonal and -1 next to it); e^{-H} is a finite-dimensional Gaussian measure with covariance K^{-1} (1-08). As $L\to\infty$, the G at the centre of the interval converges to (8) below. (8) is obtained by putting $G = Ae^{-\kappa|x|}$ for $x \neq 0$ into the equation $\sum_y K_{xy}G(y) = \delta_{x0}$ of the infinite chain to get $m^2+2 = e^{\kappa}+e^{-\kappa}$, and fixing A by the equation at $x=0$.

$$G(x) = \frac{e^{-\kappa|x|}}{2\sinh\kappa}, \qquad \cosh\kappa = 1 + \frac{m^2}{2}, \qquad \xi = \frac{1}{\kappa} \approx \frac{1}{m} \tag{8}$$

The smaller the mass, the longer the correlation length, and it diverges at the rate $1/m$ as $m\to 0$.

Computed For $m = 1,\ 0.2,\ 0.05$, $\xi = 1.03904,\ 5.00831,\ 20.00208$ ($\xi m \to 1$). The $G(20)$ obtained by numerically inverting K at the centre of the interval (1601 to 4001 points) agreed with (8) to 13 digits.

This chain is the imaginary-time harmonic oscillator of 2-07 placed on a lattice in time. Distributing $m^2\phi_x^2$ half and half to the couplings on either side, e^{-H} becomes the product of weights of neighbouring pairs

$$T(\phi,\phi') = \exp\Big(-\frac{(\phi-\phi')^2}{2} - \frac{m^2(\phi^2+\phi'^2)}{4}\Big)$$

We call the integral operator $(Tf)(\phi) = \int T(\phi,\phi')f(\phi')\,d\phi'$ with this T as kernel the **transfer matrix** (the $\sum$ of a matrix product replaced by an integral). For a periodically

joined chain of L points, $Z = \int T(\phi_1,\phi_2)T(\phi_2,\phi_3)\cdots T(\phi_L,\phi_1)\,d\phi_1\cdots d\phi_L$, that is, the trace of T^L. The ratio of the two largest eigenvalues of T is $e^{-\kappa}$, and we call this κ the **gap** of the transfer matrix. This is the simplest example in which **the inverse of the correlation length is the gap** Known (the general form is in 4-11). Computed The ratio of eigenvalues of the matrix discretising T with 1500 points agreed with $e^{-\kappa}$ to 10 digits at $m = 1,\ 0.5,\ 0.2$.

04

Phase transitions — kinks that occur only in infinite systems

We call it a **phase transition** when a slight change in temperature or the like changes the behaviour completely. In mathematical language, the $f(\beta, h)$ of the thermodynamic limit is not analytic at that point (cannot be represented by a Taylor expansion around the point). **It does not occur in finite systems.** Precisely, Z_N is a finite sum of exponentials $e^{-\beta H(\sigma)}$ of β, h, positive for real β, h. So $\log Z_N$ can be differentiated any number of times and its Taylor expansion converges (this is called being **real-analytic**).

For the Ising model the following is known Known . For $d = 1$ there is no phase transition (Ising 1925, 2-09). For $d = 2$, $h = 0$,

$$\sinh(2\beta_c J) = 1, \qquad \beta_c J = \tfrac{1}{2}\log\bigl(1+\sqrt{2}\bigr) = 0.440687\ldots \quad (9)$$

a phase transition occurs. Assuming there is just one transition, Kramers–Wannier 1941 located it by duality, and Onsager 1944 found the free energy exactly and showed the $-\log|T - T_c|$ divergence of the specific heat. The argument showing that the alignment survives at low temperature is said to begin with Peierls 1936 (put into shape as a proof in later years). What states this precisely is **spontaneous magnetisation**.

$$\bar{m}_0(\beta) = \lim_{h\downarrow 0}\ \lim_{N\to\infty} \langle\sigma_0\rangle_{N,h} \quad (10)$$

Example. The order of limits matters. Taking $h \to 0$ first, the H of the finite system is invariant under the global flip $\sigma \to -\sigma$, so $\langle\sigma_0\rangle = 0$. Going to infinity first and then switching off the external field, in two dimensions $\bar{m}_0 > 0$ remains for $\beta > \beta_c$ (going to 0 like $(\beta - \beta_c)^{1/8}$ as $\beta \downarrow \beta_c$; Yang 1952). **That the answer changes with the order of limits** is a recurring issue in Yang–Mills too (5-09, 17-10).

The view of phase transitions through the zeros of the partition function is treated in the detour after section 05.

05

The mean-field model — watching the tug of war

The **Curie–Weiss model** $H_N = -\frac{J}{2N}M^2 - hM$ ($M = \sum_{x=1}^{N}\sigma_x$, $J>0$), in which every spin is coupled with the same strength, is the smallest example in which a phase transition can be followed by hand. Grouping by the number of up spins k gives $Z_N = \sum_k \binom{N}{k} e^{-\beta H_N}$. Let $\bar m = M/N$ be the magnetisation per spin (with a bar, to distinguish it from the mass m of the Gaussian chain). Using $\binom{N}{N(1+\bar m)/2} = e^{Ns(\bar m)+o(N)}$, $s(\bar m) = -\frac{1+\bar m}{2}\log\frac{1+\bar m}{2} - \frac{1-\bar m}{2}\log\frac{1-\bar m}{2}$ from Stirling's formula, the sum is determined by its largest term (Laplace's method).

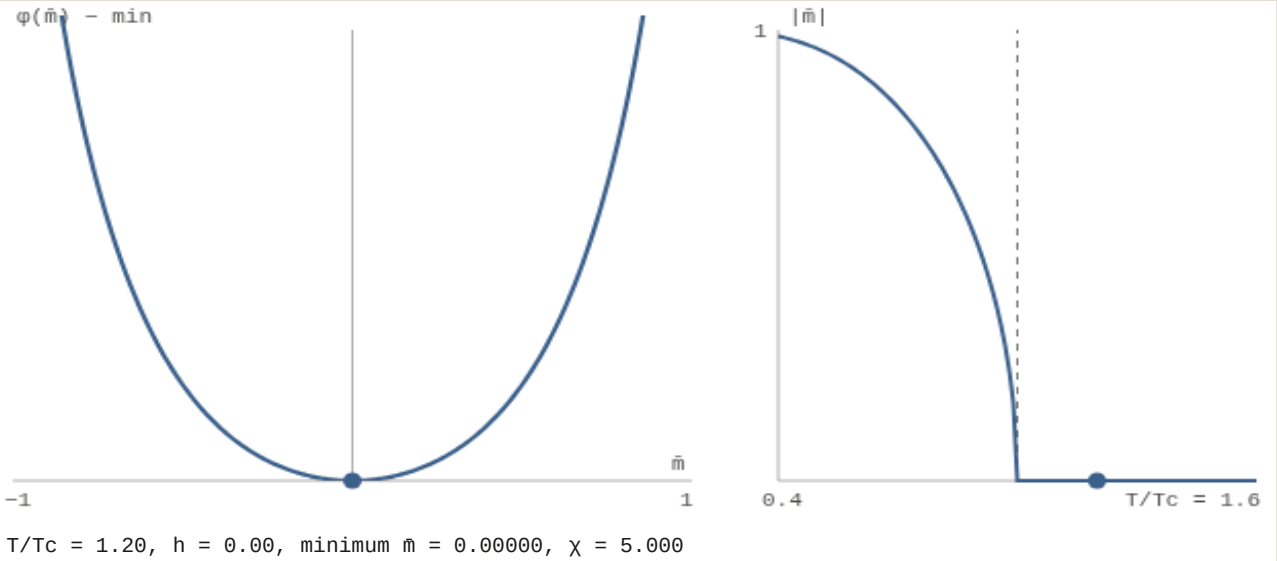
$$f = \min_{-1\le \bar m\le 1} \varphi(\bar m), \qquad \varphi(\bar m) = -\frac{J}{2}\bar m^2 - h\bar m - \frac{1}{\beta}s(\bar m) \tag{11}$$

The tug of war of (2) has shrunk to one variable (the first term is the energy, the last the entropy). $\varphi'(\bar m) = 0$ is

$$\bar m = \tanh\big(\beta(J\bar m + h)\big) \tag{12}$$

For $h=0$: if $\beta J \le 1$, the only solution is $\bar m = 0$; for $\beta J > 1$, $\pm\bar m^{*} \neq 0$ are the minima (critical temperature $T_c = J$). From $\tanh u \approx u - u^3/3$, $\bar m^{*} \approx \sqrt{3(\beta J - 1)}$, and differentiating (12) in h gives $\chi = \beta/(1-\beta J)$ for $T > T_c$ (the Curie–Weiss law).

Figure 1 Moving T/T_c and the external field h, the valley of $\varphi(\bar m)$ on the left (minimum shifted to 0) splits in two at T_c. On the right are the $\bar m^{*}$ of $h=0$ (line) and the current minimum. "Play" cools it



T/Tc = 1.20, h = 0.00, minimum m̄ = 0.00000, χ = 5.000

Computed The figure computes (11)(12) ($J=1$) on the spot. The $\bar m^{*}$ at $h=0$ is $0.52543,\ 0.71041,\ 0.90733$ at $T/T_c = 0.9,\ 0.8,\ 0.6$ (agreeing with python3).

Computed **Table 1** $\sqrt{\langle \bar{m}^2 \rangle}$ ($h = 0$; for finite N the sum over $\binom{N}{k}$ is computed exactly) and the $\bar{m}^*$ of $N = \infty$.

N	βJ = 0.8	βJ = 1	βJ = 1.25
1000	0.07003	0.19172	0.70824
10^5	0.00707	0.06083	0.71039
∞	0	0	0.71041

$\beta J = 0.8$ goes to 0 like $N^{-1/2}$ ($\sqrt{N}$ times it is 2.2358, against the prediction 2.2361 from (4) and the Curie–Weiss law), $\beta J = 1.25$ goes to $\bar{m}^*$, and the critical $\beta J = 1$ goes to 0 like $N^{-1/4}$ (about 1.08 when multiplied by $N^{1/4}$). **For finite N each is a smooth function of β, and the kink first appears as $N \to \infty$.**

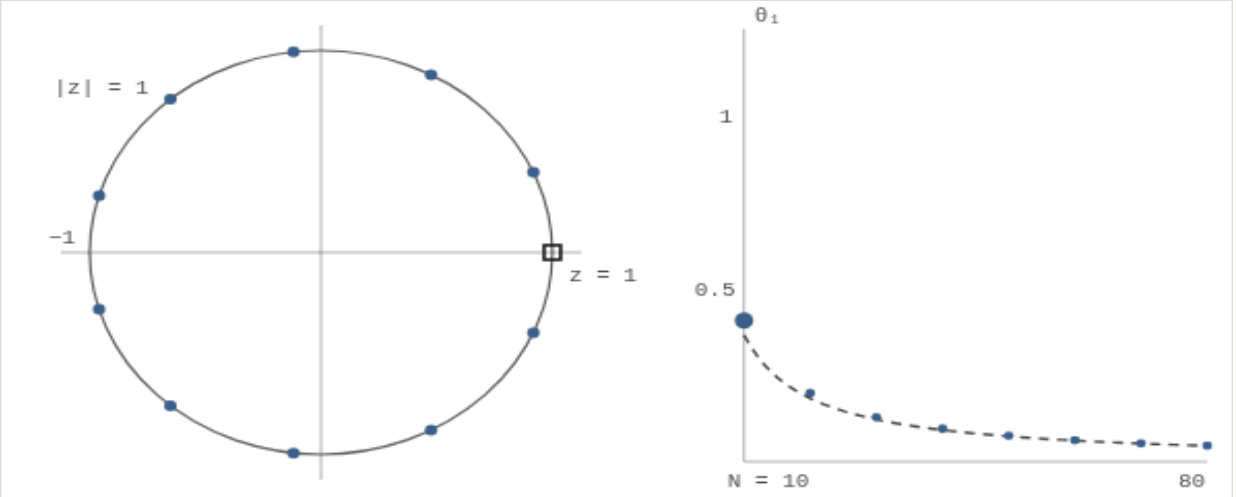
Detour: zeros of the partition function

A subsection you may skip.

Putting $z = e^{-2\beta h}$, Z_N is a polynomial in z with positive coefficients (multiplied by $e^{\beta h N}$), and for a ferromagnet all its zeros lie on the unit circle $|z| = 1$ (Lee–Yang 1952) Known . Real h corresponds to the positive real axis, and $h = 0$ to $z = 1$. Since the coefficients are positive, for finite N no zero lies on $z = 1$, and $\log Z_N$ is smooth near $h = 0$. If the zeros pinch $z = 1$ as $N \to \infty$, f can kink at $h = 0$.

Computed In the model of section 05, the angle θ_1 of the zero nearest to $z = 1$, from $N = 10$ to 80, was $N\theta_1 = 4.08 \to 3.72$ for $\beta J = 1.5$ (closing in on 0; the prediction $\pi/\bar{m}^* = 3.66$ when Z_N is approximated by a constant multiple of $\cosh(N\beta h \bar{m}^*)$ at low temperature), and $\theta_1 = 1.157 \to 0.722$ for $\beta J = 0.5$ (consistent with the prediction 0.571 coming from the point where, with h made complex, the two solutions of (12) merge into one).

Figure 2 Moving N increases the zeros on the unit circle (dots; upper and lower are complex conjugates). At $\beta J = 1.5$ (low temperature) they close in on $z = 1$ ($h = 0$, square); at $\beta J = 0.5$ (high temperature) an arc with no zeros remains around $z = 1$ (the hatching is the prediction $|\theta| < 0.571$ for $N \to \infty$). On the right (below, on narrow screens) are θ_1 and the prediction (dashed). "Play" takes N from 10 to 80



```
βJ = 1.5, N = 10, θ₁ = 0.40767, Nθ₁ = 4.077 (predicted limit of Nθ₁: π/m̄* = 3.659)
```

Computed The zeros in the figure are angles found with python3 and embedded. We followed the changes of sign of $z^{-N/2}Z_N$, which is real on $z = e^{i\theta}$, with 80-digit decimal arithmetic, and checked that there are exactly $N/2$ of them in $0 < \theta < \pi$ (θ_1 agrees with the values above to 6 digits).

06

Critical points — the correlation length diverges and powers appear

We call the point of a transition in which the spontaneous magnetisation rises continuously from 0 a **critical point**. Approaching it, the correlation length diverges, and many quantities behave as powers. Their exponents are the **critical exponents**. With $t = (T - T_c)/T_c$,

$$\xi \sim |t|^{-\nu}, \quad \chi \sim |t|^{-\gamma}, \quad \bar{m}0 \sim (-t)^{\beta}, \quad G(x)|_{T=T_c} \sim \frac{1}{|x|^{d-2+\eta}} \tag{13}$$

we write ($\sim$ means the ratio approaches a constant. The exponent β is a different thing from the inverse temperature. η is written η_s in later chapters when it must be distinguished from the dynamical exponent: 18-03 § 07. The relaxation time and the dynamical exponent z are in 2-10). From the scaling hypothesis the following relations among the exponents follow (the renormalisation group explains them; equations (1.38)–(1.40) of Pelissetto–Vicari 2002) Physics , and the right two are held to hold for $d < 4$.

$$\gamma = \nu(2 - \eta), \qquad \beta = \frac{\nu}{2}(d - 2 + \eta), \qquad \alpha = 2 - d\nu \tag{14}$$

α is the exponent of the specific heat. Putting in the $\nu = 1,\ \eta = 1/4$ of two-dimensional Ising gives back the exact values $\gamma = 7/4,\ \beta = 1/8,\ \alpha = 0$ Computed .

Computed **Table 2** Critical exponents. Mean-field and two-dimensional values are from Pelissetto–Vicari (Table 13 and elsewhere), the three-dimensional ν, η are numerical values from the conformal bootstrap (Kos et al. 2016), and * is computed from (14) (errors move it by about ± 1 in the last digit).

Exponent	Mean field (d > 4)	Two-dimensional Ising	Three-dimensional Ising Physics
ν (correlation length)	1/2	1	0.62997
η (correlation at the critical point)	0	1/4	0.036297

Exponent	Mean field (d > 4)	Two-dimensional Ising	Three-dimensional Ising Physics
γ (susceptibility)	1	7/4	1.23708
β (spontaneous magnetisation)	1/2	1/8	0.32642
α (specific heat)	0	0 (logarithmic)	0.11009

Example. In the Gaussian chain, the squared mass m^2 (section 03) plays the role of t, and $\nu = 1/2$ follows from $\xi \approx (m^2)^{-1/2}$. Together with the $\beta = 1/2,\ \gamma = 1$ of section 05, this is the same as the mean-field column. η measures the departure from Gaussian behaviour (Pelissetto–Vicari), and for three-dimensional Ising it is small but not 0.

07

Universality — forgetting the details

The exponents of Table 2 are thought not to depend on the shape of the lattice or the details of the couplings, but only on the dimension, the symmetry and the range of the interaction. A collection of systems with the same exponents is a **universality class** Physics . The gas–liquid critical point of simple fluids and uniaxial antiferromagnets are held to belong to the three-dimensional Ising class (Pelissetto–Vicari § 3.1), so that liquids and magnets have the same exponents.

Near a critical point ξ becomes much longer than the lattice spacing. Kadanoff 1966 regrouped spins into blocks larger than the lattice spacing and smaller than the correlation length and looked again, and Wilson 1971 turned the repetition of this operation into the equations of the renormalisation group. Details fade, and systems that flow into the same fixed point have the same exponents. Precisely, universality is a principle of physics confirmed in many examples, not a general theorem (7-05, 4-04).

As the smallest example, we add to (7) a coupling $c \geq 0$ to the next-but-one neighbour.

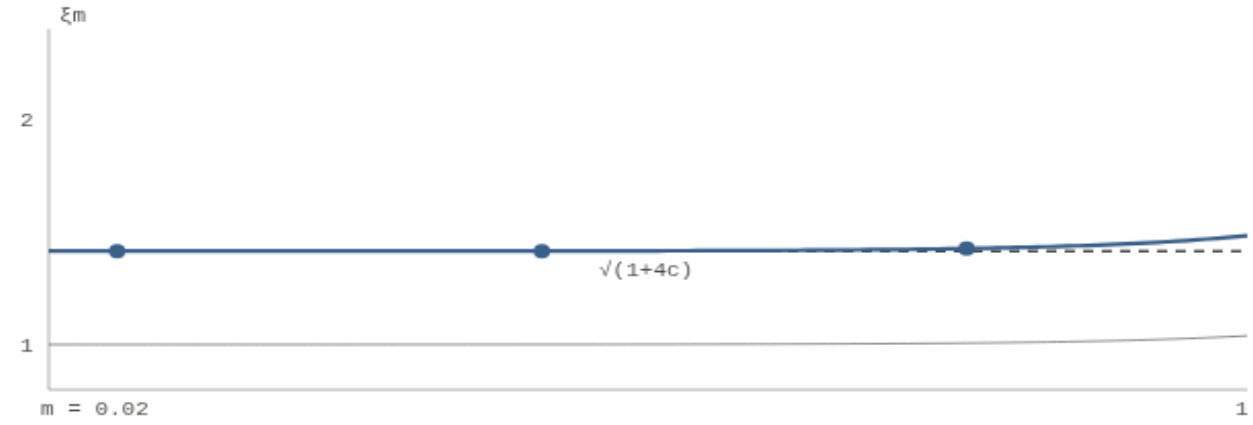
$$H_c = \frac{1}{2} \sum_x \left[(\phi_{x+1} - \phi_x)^2 + c\,(\phi_{x+2} - \phi_x)^2 + m^2 \phi_x^2 \right] \tag{15}$$

By the same computation as section 03, G becomes a sum of terms of the form $e^{-\kappa|x|}$, and the κ of the most slowly decreasing term is the smallest positive root of $m^2 = 2(\cosh\kappa - 1) + 2c(\cosh 2\kappa - 1)$. This is the quadratic equation $4cw^2 + 2w - (2 + 4c + m^2) = 0$ in $w = \cosh\kappa$, so the root can be written by a formula. When κ is small the right side is about $(1 + 4c)\kappa^2$, so $\xi \approx \sqrt{1 + 4c}/m$.

Computed **Table 3** ξm for the chain with c added. The limit is $\sqrt{1+4c} = 1,\ 1.41421,\ 2.23607$ (also agreeing with the inverse matrix to 10 digits).

m	c = 0	c = 1/4	c = 1
0.4	1.00659	1.42586	2.24613
0.1	1.00042	1.41495	2.23670
0.025	1.00003	1.41426	2.23611

Figure 3 Moving the next-but-one coupling c changes the height of the limit $\sqrt{1+4c}$ (dashed) of the curve of ξm (horizontal axis m, logarithmic scale; the dots are the m of Table 3). That the curve becomes flat as $m \to 0$ ($\xi \propto m^{-1}$) does not depend on c. The thin line is $c = 0$. "Play" takes c from 0 to 1



c = 0.25, √(1+4c) = 1.41421, ξm: m = 0.4 → 1.42586, 0.1 → 1.41495, 0.025 → 1.41426

Computed The figure computes the root of the quadratic equation above on the spot. At $c = 0,\ 1/4,\ 1$ it agrees with Table 3 (and with python3's root search and inverse matrix).

The exponent ($\xi \propto m^{-1}$) does not depend on c, but the coefficient $\sqrt{1+4c}$ does. For a slowly varying field $(\phi_{x+2} - \phi_x)^2 \approx 4(\phi_{x+1} - \phi_x)^2$, so the next-but-one coupling looks merely like the neighbour coupling multiplied by $1 + 4c$, and disappears when the ruler of length is changed.

08

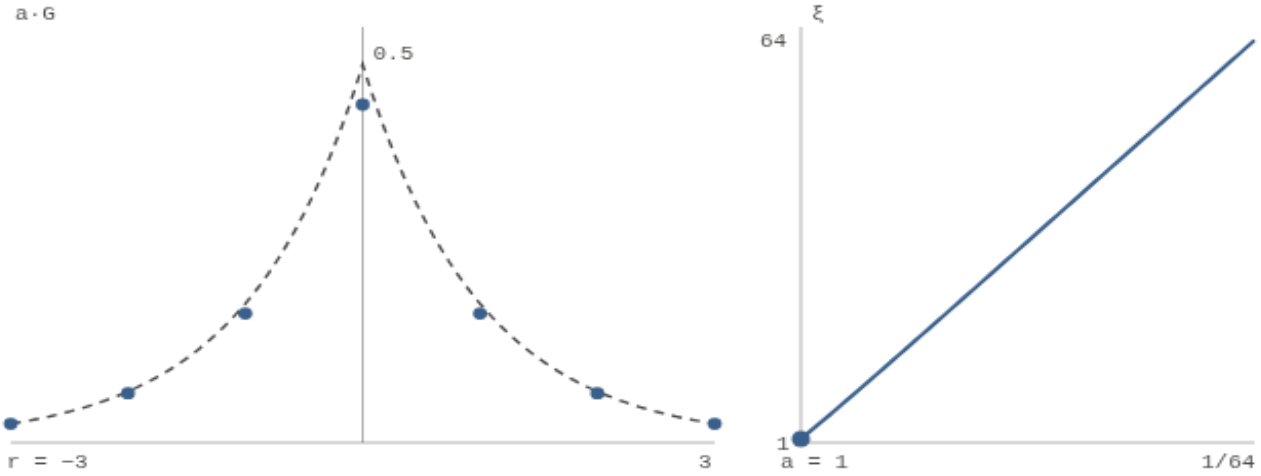
The continuum limit — the critical point is where the lattice disappears

Introduce the lattice spacing a and read the point x as the position $r = xa$. If the correlation length in lattice units is ξ (the $1/\kappa$ of section 03), the correlation length in physical units is $a\xi$. Keeping $a\xi$ fixed and letting $a \to 0$, ξ has no choice but to go to infinity. **Taking the continuum limit is approaching a critical point** (Chatterjee 2018 § 5) Physics . For the Gaussian chain everything is explicit: fixing $M > 0$ and putting $m = Ma$, from $\kappa/a \to M$, $a/(2\sinh\kappa) \to 1/(2M)$ the following holds Known .

$$a\,G\Big(\frac{r}{a}\Big) \;\longrightarrow\; \frac{e^{-M|r|}}{2M} \qquad (a \to 0) \quad (16)$$

The right side is the kernel of the inverse of $-d^2/dr^2 + M^2$ (the propagator, 1-11), and M is the mass in physical units.

Figure 4 Making the lattice spacing a smaller, the aG in physical units (dots) settles onto the continuous $e^{-M|r|}/2M$ (dashed, $M = 1$), and the correlation length in lattice units ξ on the right (log–log) diverges. "Play" takes a from 1 to 1/64



a = 1.00, ξ = 1.039, aξ = 1.0390, a·G(0) = 0.44721 (continuum 0.5)

Computed The figure computes (8)(16) on the spot. **Table 4** ξ (in lattice units) and aG (python3) as a is shrunk with $M = 1$. The bottom row is the right side of (16).

a	ξ (lattice units)	a·G(0)	a·G(r = 1)
1	1.0390	0.447214	0.170820
1/8	8.0052	0.499026	0.183701
1/64	64.0007	0.499985	0.183936
Continuum	—	0.5	0.183940

For Yang–Mills, in many non-abelian lattice gauge theories the critical point is believed to be at $\beta = \infty$ (the β here is the constant multiplying the action of lattice gauge theory, proportional to the inverse square of the coupling constant g, and plays the role of inverse temperature; 5-01), and the inverse of the correlation length corresponds to the mass of the lightest glueball (a particle made of gluons alone, 17-01) (Chatterjee 2018 § 5) Physics . One of the questions on the lattice side is "that correlations decrease exponentially for each $\beta > 0$ ($\xi(\beta)$ is determined in $(0, \infty)$), and that $\xi(\beta) \to \infty$ as $\beta \to \infty$", which Chatterjee lists in the same section as the lattice correlation-length problem (Problem 5.1) (5-09). Whether the mass stays positive in the continuum limit is a further question (5-05, 4-11). Weights, the correlation length, the critical point and universality correspond to the path integral of field theory (4-02), the inverse of the mass, the continuum limit and renormalisation, and for the Gaussian chain this is exact.

What this chapter can say

Content	Status
(1)–(4), analyticity of finite systems, (8) and the gap of the transfer matrix, the ξ of (15), (16), (11)(12)	Said Known (Tables 1, 3, 4 and Figures 1, 3, 4 Computed)
Existence of the thermodynamic limit	Theorem Known (literature bibliographic data only)
The facts for $d = 1, 2$ of the Ising model, the Lee–Yang theorem	Theorem Known (literature checked up to the abstracts; Ising 1925 and Peierls 1936 bibliographic data only)
The angles of the zeros in the detour (Figure 2)	Numbers for $N \leq 80$ Computed . The limits are predictions
Three-dimensional critical exponents, (14), universality in general	Numbers and the level of physics Physics (the * of Table 2 is arithmetic from (14) Computed)
The correspondence between correlation length and mass, the critical point $\beta = \infty$ of lattice gauge theory	At the level of physics Physics . Nothing is shown about the mass gap or the construction of the continuum limit

Sources and reproduction

Item	Kind	Source / tools
Tables 1–4, the numbers in section 03 and the detour, the ratio of eigenvalues of the transfer matrix, the check values for Figures 1–4 and the zeros of Figure 2	Computed	python3 (NumPy, SciPy; the zeros with 80-digit decimal arithmetic from the standard library). The figures by JavaScript in the browser
The form of (6), the continuum limit and the critical point, $\beta_c = \infty$, glueballs, the lattice correlation-length problem (Chatterjee's Problem 5.1)	Known	S. Chatterjee, arXiv:1803.01950. § 5 checked
(13)(14), mean-field and two-dimensional exponents, universality classes	Known	A. Pelissetto, E. Vicari, Phys. Rep. 368 (2002) 549–727, arXiv:cond-mat/0012164. § 1.1, § 1.2, § 3.1 and Table 13 checked
The ν, η of three-dimensional Ising	Physics	F. Kos, D. Poland, D. Simmons-Duffin, A. Vichi, JHEP 08 (2016) 036, arXiv:1603.04436. Abstract and introduction checked
(9), specific heat / duality / spontaneous magnetisation / zeros	Known	L. Onsager, Phys. Rev. 65 (1944) 117 / H. A. Kramers, G. H. Wannier, ibid. 60 (1941) 252 / C. N. Yang, ibid. 85 (1952) 808 / C. N. Yang, T. D. Lee, ibid. 87 (1952) 404 / T. D. Lee, C. N. Yang, ibid. 410. Abstracts checked

Item	Kind	Source / tools
Existence of the thermodynamic limit	Known	D. Ruelle, Statistical Mechanics: Rigorous Results (first edition 1969, reissued by World Scientific 1999) (bibliographic data only)
The view of grouping into blocks / the renormalisation group	Known	L. P. Kadanoff, Physics 2 (1966) 263 / K. G. Wilson, Phys. Rev. B 4 (1971) 3174. Abstracts checked
No phase transition in one dimension / the low-temperature argument	Known	E. Ising, Z. Phys. 31 (1925) 253 (bibliographic data only) / R. Peierls, Proc. Cambridge Philos. Soc. 32 (1936) 477 (bibliographic data only, not checked)
(2), Laplace's method, the covariance of Gaussian measures, the gap of the transfer matrix	Known	Standard facts (1-08, 2-07, 2-09, 4-11)

Next chapter to read: 2-09 Detour: the gap and the correlation length in the Ising model. For the relation to the mass gap see 4-11; for the relaxation time and the dynamical exponent, 2-10.

Detour: the gap and the correlation length in the Ising model — the ratio of eigenvalues of the transfer matrix sets the length over which correlations reach

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That the ratio of the two largest eigenvalues of the transfer matrix sets the correlation length, and $1/\xi$ is exactly the gap of $-\log T$. The three behaviours of the gap on a two-dimensional strip: high temperature, the critical point, low temperature. That the gap which closes at low temperature is not a sign that there is no mass gap.

Prerequisite chapters — 2-08 (the correlation length, critical points). Sections 04 and 08 are easier if you have read 2-05 (the gap, two levels, the double well) and 2-07 (imaginary time, the transfer matrix).

Said first — This is a detour, so you may skip it. The two-dimensional numbers are computations on strips of width up to 20. This chapter shows nothing about the mass gap of Yang–Mills.

Order of this chapter

1. The overall picture — the inverse of the correlation length is the gap
2. The one-dimensional Ising model and the transfer matrix
3. Correlations decrease by the ratio of eigenvalues (Figure 1, Table 1)
4. The transfer matrix is quantum mechanics in imaginary time
5. Two dimensions: the transfer matrix of a strip (Table 2)
6. The high-temperature side — the gap remains
7. The critical point (Figure 2) — closes like $1/L$
8. The low-temperature side (Table 3) — a gap that closes and a gap that remains
9. Back to the language of Yang–Mills (Detour: Z_2 gauge theory)

01

The overall picture — the inverse of the correlation length is the gap

The mass gap was the difference between the two lowest eigenvalues of the Hamiltonian (2-05), and the correlation length was the length over which the joint movement of two distant points disappears (2-08). We check in the simplest model that the two are a single number.

The key is the transfer matrix. If the eigenvalues of the matrix T that advances the lattice one column at a time are $\lambda_0 > \lambda_1 \geq \cdots$, $E_n = -\log \lambda_n$, then

$$\frac{1}{\xi} = \log \frac{\lambda_0}{\lambda_1} = E_1 - E_0 \quad (1)$$

(the conditions are in section 04). The left side is a statistical quantity, the right side is the gap of the "Hamiltonian" $H = -\log T$, and reading one tick as one step of imaginary time, this is the same content as "the rate of decay is the gap" stated by (8) of 2-05.

Example. On a two-dimensional strip of width L, $E_1 - E_0$ settles to a positive value on the high-temperature side, is proportional to $1/L$ at the critical point, and closes exponentially on the low-temperature side (sections 05–08).

02

The one-dimensional Ising model and the transfer matrix — the partition function is the trace of a matrix

Arrange N spins $\sigma_i = \pm 1$ on a ring ($\sigma_{N+1} = \sigma_1$), coupling neighbours. With $K = \beta J$, $B = \beta h$, and distributing the external-field term half and half to neighbouring pairs, the weight is a product of factors for each neighbouring pair.

$$e^{-\beta H} = \prod_{i=1}^{N} \exp\Big(K\sigma_i\sigma_{i+1} + \frac{B}{2}(\sigma_i + \sigma_{i+1}) \Big) \quad (2)$$

Each factor is determined by σ_i and σ_{i+1} alone, so it becomes a 2×2 matrix (rows are σ_i, columns σ_{i+1}, $+1$ is the first row), and since the sum goes once round the ring and returns, it is a trace.

$$T = \begin{pmatrix} e^{K+B} & e^{-K} \\ e^{-K} & e^{K-B} \end{pmatrix}, \qquad Z_N = \sum_{\sigma} \prod_{i} T_{\sigma_i \sigma_{i+1}} = \operatorname{tr} T^N = \lambda_0^N + \lambda_1^N \quad (3)$$

T is real symmetric, so its eigenvalues are real (1-01), and

$$\lambda_{0,1} = e^{K}\cosh B \pm \sqrt{e^{2K}\sinh^{2} B + e^{-2K}} \tag{4}$$

For large N, $Z_N \approx \lambda_0^{N}$. The method of finding the partition function from the largest eigenvalue of a matrix was carried out by Kramers–Wannier 1941 for the one-dimensional chain, and taken on to two dimensions Known .

Example. Computed For $N = 14$, $K = 0.7$, $B = 0.2$, the Z_N obtained by summing all 2^{14} configurations, and the correlations, agreed with the trace formula (3) within rounding error (differences in correlations below 10^{-13}).

03

Correlations decrease by the ratio of eigenvalues — there is no phase transition in one dimension

Write the orthonormal eigenvectors of T as v_0, v_1 and the spin as $S = \mathrm{diag}(1, -1)$. In $N \to \infty$ the rest of the ring is replaced by $\lambda_0^{N-r} v_0 v_0^{\mathsf{T}}$, and expanding the T^{r} in between in eigenvalues gives the correlation at distance r.

$$m = v_0^{\mathsf{T}} S v_0, \qquad G(r) = \langle \sigma_0 \sigma_r \rangle - m^{2} = (v_0^{\mathsf{T}} S v_1)^{2}\left(\frac{\lambda_1}{\lambda_0}\right)^{r} = (1 - m^{2})\left(\frac{\lambda_1}{\lambda_0}\right)^{r} \tag{5}$$

The last equality follows because putting $v_0 = (\cos\varphi, \sin\varphi)$, $v_1 = (-\sin\varphi, \cos\varphi)$ gives $m = \cos 2\varphi$, $v_0^{\mathsf{T}} S v_1 = -\sin 2\varphi$. The correlation length is as in (1). For $B = 0$, $\lambda_0 = 2\cosh K$, $\lambda_1 = 2\sinh K$, and

$$G(r) = (\tanh K)^{r}, \qquad \xi = \frac{1}{\log\coth K} \approx \frac{e^{2K}}{2} \quad (K \to \infty) \tag{6}$$

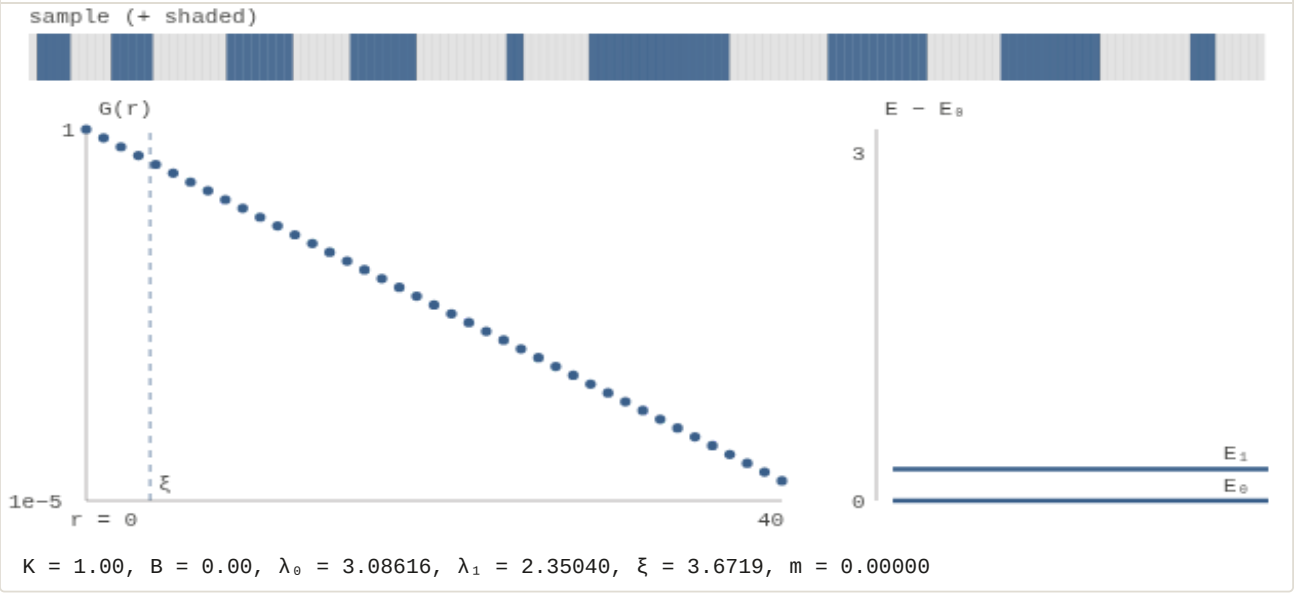
Computed **Table 1** Eigenvalues and correlation lengths for $B = 0$.

K	λ_0	λ_1	ξ	$e^{2K}/2$
1	3.086161	2.350402	3.671861	3.6945
2	7.524391	7.253721	27.296022	27.2991
3	20.135324	20.035750	201.713984	201.7144

For every $K < \infty$, $\lambda_0 > \lambda_1$, so ξ is finite, and switching off the external field gives $m = \sinh B / \sqrt{\sinh^{2} B + e^{-4K}} \to 0$. The quantity under the square root in (4) is positive, so λ_0 does not coincide with λ_1, and the free energy per site f is $\beta f = -\lim_{N\to\infty} \frac{1}{N}\log Z_N = -\log\lambda_0$, an analytic function of K and B. This is the reason, in terms of the transfer matrix, why there is no phase transition in one dimension (Ising 1925; for the

definition of a phase transition see section 04 of 2-08) Known . Even for a general square matrix with all entries positive, the largest eigenvalue does not coincide with another; this is called the Perron–Frobenius theorem (here we give only the name). Intuitively, the fraction of neighbours that disagree (the fraction of walls) $p = \frac{1-\tanh K}{2} \approx e^{-2K}$ does not become 0 however low the temperature, and the correlation changes sign each time it crosses a wall, so $\xi \approx 1/(2p)$.

Figure 1 Moving K and B (= βh) changes together the lengths of the clusters in the sample chain at the top, the slope of G(r) on the left (logarithmic scale), and the level spacing $E_1 - E_0 = 1/\xi$ on the right. "Play" increases K (cools)



K = 1.00, B = 0.00, λ₀ = 3.08616, λ₁ = 2.35040, ξ = 3.6719, m = 0.00000

Computed $G(r)$ and the levels are computed on the spot from (4)(5); the chain is a sample of the Markov chain built from v_0 (random numbers with a fixed seed). The readout agrees with python3 ($\xi = 2.9563$ at $K = 1$, $B = 0.1$).

04

The transfer matrix is quantum mechanics in imaginary time — the two levels of one spin

The T of $B = 0$ is the sum of e^{K} times the identity matrix and e^{-K} times $\sigma^x = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ (1-04). Comparing with $e^{K^*\sigma^x} = \cosh K^* + \sigma^x \sinh K^*$, it can be written as

$$T = \sqrt{2\sinh 2K}\; e^{K^*\sigma^x}, \qquad \tanh K^* = e^{-2K} \quad \left(\text{equivalent form } \sinh 2K \,\sinh 2K^* = 1\right)$$

We call K^* the dual coupling (the quantity of Kramers–Wannier duality). $T = e^{-H}$, $H = -K^*\sigma^x + \text{const.}$ is a single spin in a transverse magnetic field, that is, a two-level system (2-05), and its gap is $E_1 - E_0 = 2K^* = \log\coth K = 1/\xi$. The direction of the chain is imaginary time, and one tick is a step of 1 (2-07). The stronger the coupling (the lower the temperature), the smaller K^*, the closer the two levels, and the longer the correlations persist.

In general, if T is a positive symmetric matrix (in infinite dimensions, a positive self-adjoint operator) with eigenvectors $|n\rangle$, the correlation of an observable A is, by the same computation,

$$\langle A_0 A_r\rangle - \langle A\rangle^2 = \sum_{n\geq 1} |\langle 0|A|n\rangle|^2\, e^{-(E_n-E_0)r} \tag{8}$$

Known . Example. A level whose coefficient $\langle 0|A|n\rangle$ is 0 by symmetry does not appear in the correlation of that A. The gap is "the slowest decrease, valid for every observable", and individual correlations decrease at least that fast (the remark after Table 2 of 2-05). That the transfer matrix of lattice gauge theory is positive self-adjoint follows from reflection positivity Known (said to be Osterwalder–Seiler 1978; details in 5-03).

05

Two dimensions: the transfer matrix of a strip — a matrix of order 2^L

Put the two-dimensional Ising model on the square lattice ($B=0$, K both vertically and horizontally) on a strip of width L. The width direction is joined periodically (a cylinder), and the length direction is read as imaginary time. One column has 2^L states; writing the couplings within a column as the diagonal matrix $V_1 = \exp(K\sum_j \sigma_j \sigma_{j+1})$ and the couplings to the next column as V_2, using (7) at each point gives

$$T = V_1^{1/2} V_2 V_1^{1/2}, \qquad V_2 = (2\sinh 2K)^{L/2} \exp\Big(K^* \sum_{j=1}^{L} \sigma_j^x\Big) \tag{9}$$

Here σ_j^x is the matrix of order 2^L that flips only the j-th spin, and σ_j^z is the diagonal matrix that multiplies by the sign ± 1 of the j-th spin; neither does anything to the other spins. The exponent of V_1 is the sum of $\sigma_j^z\sigma_{j+1}^z$, and the exponent of V_2 the sum of σ_j^x. Putting half of V_1 on each side makes T a symmetric matrix, so (8) of section 04 can be used. T has the form of one step of imaginary time for a system of quantum spins in a row (the transverse-field Ising chain). Two-dimensional classical statistical mechanics has become a one-dimensional quantum system.

T commutes with the global flip $\sigma \to -\sigma$, so the eigenvectors split into even and odd; the largest even one is λ_0, the largest odd one λ_1, and the correlation length of the strip is $\xi_L = 1/(E_1 - E_0)$ (equation (1) of de Queiroz 2011). The critical point $\sinh 2K_c = 1$ ((9) of 2-08) is the self-dual point where, by (7), $K^* = K$.

Computed **Table 2** Level differences on strips of width L (E_2 is the third level). The bottom row gives the values of the formulas for $L\to\infty$ (sections 06–08).

L	$T/T_c = 1.5$ $E_1 - E_0$ To Figure 2	**1.2** $E_1 - E_0$ To Figure 2	**1.0** ξ_L / L To Figure 2	**0.8** $E_1 - E_0$ To Figure 2	**0.8** $E_2 - E_0$ To Figure 2
8	0.666788	0.326116	1.264745	7.731e-3	1.122679
12	0.665579	0.314285	1.269548	1.173e-3	0.969005
16	0.665513	0.311692	1.271177	1.924e-4	0.907939
20	–	0.311069	1.271924	3.287e-5	0.878022
Formula	0.665509	0.310856	1.273240	0	0.821788

$L = 20$ (order $2^{20} \approx 10^6$) too was found without building the matrix, only by applying the actions of V_1 and V_2 in turn. At the three temperatures the behaviour is completely different.

06

The high-temperature side — the gap remains, with a band of particles above it

For $T > T_c$ ($K < K^*$), $E_1 - E_0$ settles to a positive value even as L is widened. The eigenvalues of the strip have a closed form (Onsager 1944; summarised in the review of Domb 1960, with the form of the formulas in equations (6)–(8) of de Queiroz 2011), and its building block is the following $\gamma(\omega)$. The right-hand form follows from $\sinh 2K \sinh 2K^* = 1$.

$$\cosh\gamma(\omega) = \cosh 2K^* \cosh 2K - \sinh 2K^* \sinh 2K \cos\omega = \cosh\gamma(0) + (1 - \cos\omega) \quad (10)$$

As $\gamma(0) = 2|K^* - K|$, the $E_1 - E_0$ of Table 2 approaches $2(K^* - K)$ Computed (difference 4×10^{-6} at $T/T_c = 1.5$, $L = 16$). $\gamma(\omega)$ can be read as the energy of a single particle of momentum ω (the free-fermion view; Schultz–Mattis–Lieb 1964). On the strip the momentum is quantised to $2\pi n/L$, and the second odd level is close to $\gamma(2\pi/L)$, agreeing better the further from T_c (at $L = 16$: 0.763461 and 0.763457 for $T/T_c = 1.5$, 0.495433 and 0.494596 for 1.2) Computed . As $L \to \infty$ these come down to $\gamma(0)$, and above the gap there is a continuous band (a real example of section 06 of 2-05).

Example. Near the critical point, when γ, ω is small, expanding the right-hand form of (10) to second order gives

$$\gamma(\omega)^2 \approx \gamma(0)^2 + \omega^2 \quad (11)$$

which has the relativistic form $E^2 = m^2 + p^2$ (speed of light 1, 2-04). Precisely, (11) is an approximation only when $\gamma(0)$ and ω are small (for $T/T_c = 1.05$, $\omega = 0.1$: left side 0.017234, right side 0.017263) Computed .

07

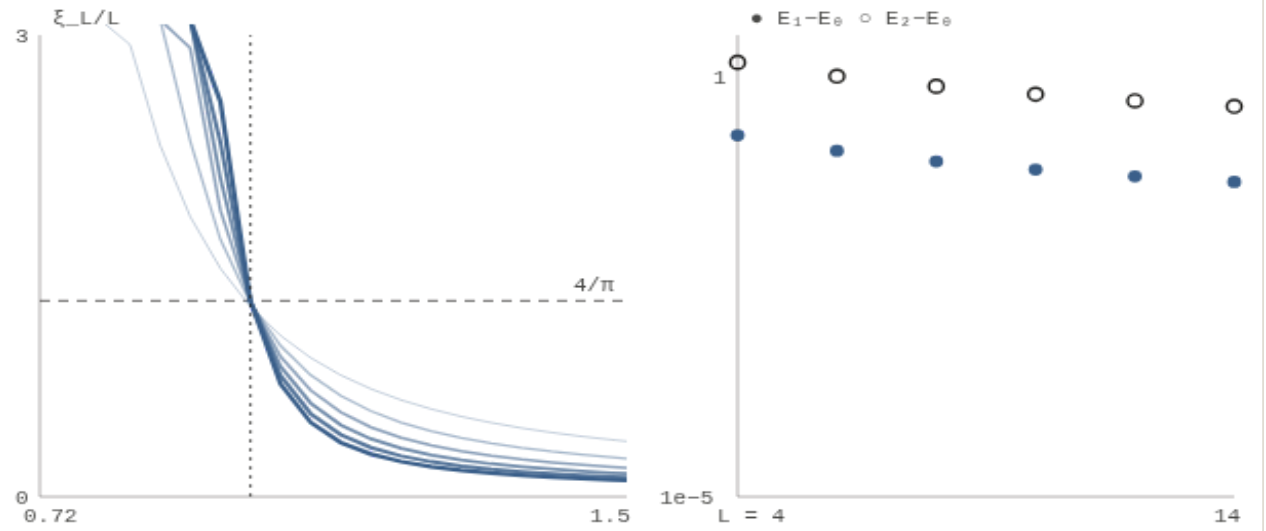
The critical point — the gap closes like 1/L

At $T = T_c$, $\gamma(0) = 0$, and the gap of an infinitely wide strip is 0. On a finite strip ξL grows in proportion to L, and the constant of proportionality is a universal quantity independent of the details of the model. The first formula of (12) below (a general relation) is a prediction from conformal invariance (the property, at a critical point, of not changing form even under transformations that preserve angles; here we give only the name) Physics (Cardy 1984; the form is equations (3)(5) of de Queiroz 2011). For the two-dimensional Ising model, the same value $4/\pi$ comes out of the closed form of the eigenvalues of the strip in section 06 Known .

$$E_1 - E_0 \approx \frac{2\pi x}{L}, \qquad \frac{\xi L}{L} \longrightarrow \frac{1}{2\pi x} = \frac{4}{\pi} = 1.273240 \quad \left(x = \tfrac{1}{8}\right) \quad (12)$$

x is the scaling dimension of the spin, with $G(r) \sim r^{-2x}$ at the critical point; in two dimensions it is half of $\eta = 1/4$ of 2-08. Example. Computed Extrapolating from the $\xi L/L$ of $L = 18,\ 20$, assuming the difference vanishes like c/L^2, gives 1.273248, differing from $4/\pi$ by 8×10^{-6}.

Figure 2 Moving T/T_c changes the vertical line on the left, and $E_1 - E_0$ (●) and $E_2 - E_0$ (○) for each width L on the right (logarithmic scale). On the left are ξ_L/L (L = 4–14, darker for wider) and 4/π (dashed). On the right, the dashed lines are the formulas for $L \to \infty$, and the dotted line is the low-temperature slope σ. "Play" cools. The temperature can also be chosen with "To Figure 2" in Table 2



T/Tc = 1.00, L = 14: E₁−E₀ = 0.0562 (ξ_L = 17.8), E₂−E₀ = 0.4469, L→∞: ξ_L/L → 4/π = 1.2732

Computed The numbers in the figure were computed on this machine ($L = 4$–14, $T/T_c = 0.72$–1.5 in steps of 0.04 (0.02 for the last only); python3 values embedded to 4 significant figures). The dashed and dotted lines on the right compute the values of (10)(13) on the spot.

The lines on the left gather almost at one point at T_c even for different widths, and open downwards on the high-temperature side and upwards on the low-temperature

side, the larger L.

08

The low-temperature side — a gap that closes, and a gap that remains

For $T < T_c$, $E_1 - E_0$ becomes exponentially small with L. On the strip there are two states, "almost all $+$" and "almost all $-$", and the largest even and odd eigenvectors are their sum and difference. To move from one to the other needs a wall of length L across the strip, whose weight is of order $e^{-\sigma L}$. σ is the free energy per unit length of the wall (the interface tension), and for the two-dimensional Ising model

$$\sigma = 2(K - K^*) \qquad (K > K_c) \quad (13)$$

is held to hold (the free energy of the boundary between orders of opposite direction was found by Onsager 1944). (13) becomes 0 at T_c (the self-dual point $K = K^*$), and as $K \to \infty$ it approaches $2K$, one disagreeing bond per unit length of wall (the weight e^{K} becomes e^{-K}). The relation (7) also holds with K and K^* exchanged ($\tanh K = e^{-2K^*}$), so an equivalent form is $\sigma = 2K + \log\tanh K$. This is the same picture as the double well of section 07 of 2-05. Precisely, what corresponds to the height of the hill is the free energy of the wall σL, which grows in proportion to the width.

On the other hand, the difference from the third level $E_2 - E_0$ does not close, and is held to approach two particles' worth of (10), $2\gamma(0) = 4(K - K^*)$; the extrapolation of Table 3 does not contradict this Physics. It is two particles' worth because, in one column joined periodically in the width direction, a disagreement of direction returns to the original after going once round, so only an even number of them fit, and one disagreement corresponds to one particle of (10).

Computed **Table 3** $T/T_c = 0.8$. The slope is $\frac{1}{2}\log\frac{(E_1-E_0)L-2}{(E_1-E_0)L}$; the correction assumes the prefactor $L^{-1/2}$; the extrapolation is from the $E_2 - E_0$ of $L - 2$ and L (Table 2), assuming the difference vanishes like c/L^2.

L	Slope	Corrected slope	Extrapolation of $E_2 - E_0$
12	0.46438	0.41880	0.838971
16	0.44854	0.41515	0.827635
20	0.43991	0.41357	0.824247
Formula	0.410894	0.410894	0.821788

The slope moves towards σ, and the extrapolation towards $4(K - K^*)$. The prefactor and the form of the correction are assumptions, so this is a check at the level of not contradicting (13). Conversely, fixing σ at the value of (13), taking the prefactor as L^{-p},

and solving back for p from two neighbouring L, we get 0.587, 0.564, 0.551 at $L = 12$, 16, 20, approaching $1/2$ Computed . Finite-size effects of first-order transitions on a finite strip (cylinder) are said to have been discussed by Privman–Fisher 1983. Lowering T/T_c below 1 on the right of Figure 2, you can see the ● line up parallel to the dotted line of slope σ.

This is an important reading. If the width L is finite, every entry of T is positive (the entries of V_2 are products of $e^{\pm K}$), so by the Perron–Frobenius theorem named in section 03, $\lambda_0 > \lambda_1$, that is, $E_1 - E_0 > 0$ (as in section 03, there is no phase transition). $E_1 - E_0 \to 0$ is the sign that in the limit where the width L is widened to infinity there are two ordered states, $+$ and $-$, and above each a positive gap (the limit of $E_2 - E_0$ above) is held to remain Physics . The gap that closes is a sign that "there are two vacua", not a sign that "there is no mass gap". When discussing the mass gap, one checks separately that the vacuum is unique (4-11).

09

Back to the language of Yang–Mills

We list the correspondences of this chapter. All are reinterpretations at the level of physics and show nothing about Yang–Mills Physics . The β in the table and in the paragraph below is the constant multiplying the action of lattice gauge theory (proportional to the inverse of g^2; 2-08 § 08), a different quantity from the inverse temperature β of this chapter.

Ising model	Lattice gauge theory	Chapter
$1/\xi = E_1 - E_0$	The mass of the lightest glueball (in lattice units)	5-08, 4-11
Widening the width L of the strip	Widening the spatial box	5-04
$\xi \to \infty$ as $K \to K_c$	$\xi(\beta) \to \infty$ as $\beta \to \infty$ (the continuum limit)	5-05, 5-09
Interface tension σ	String tension	5-02, 17-03

There are also differences. The critical point of four-dimensional non-abelian gauge theory is believed to be at $\beta = \infty$, and how it is approached is determined by asymptotic freedom (4-06). Nor is there a closed-form solution like that of the two-dimensional Ising model.

Detour: Z_2 gauge theory and the Ising model

A subsection you may skip.

Lattice gauge theory with group $\mathbb{Z}_2 = \{\pm 1\}$ (link variables ± 1 in the form of 5-01) is said to have been introduced by Wegner 1971. In three dimensions it is dual to the three-dimensional Ising

model, and the coefficient of the area law of its Wilson loops is held to correspond to the interface tension of the three-dimensional Ising model at the dual coupling Known (bibliographic data only). With dimension and group matched, the last row of the table above becomes an equality rather than a metaphor (this is about three dimensions, not the two-dimensional σ of this chapter itself). Lattice gauge theory with finite groups is in 5-10.

What this chapter can say

Content	Status
(2)–(7) in one dimension, the two-level reinterpretation, (8)	Said Known (Table 1 and full enumeration Computed)
Table 2 ($L \leq 20$), the high-temperature limit, (10)(11), the $4/\pi$ at the critical point	Formulas from the literature Known , numerical agreement Computed (the extrapolation assumes c/L^2). The general relation of (12) is a prediction from conformal invariance Physics
The interface tension (13), $E_2 - E_0 \to 4(K - K^*)$, the gap remaining above the ordered states	The form of the formulas not checked in the text of the literature ("is held to") Physics . The values at both ends of (13), Table 3 and the exponent of the prefactor are Computed (Table 3 and the exponent are checks of tendencies under assumptions)
The table of correspondences in section 09	Reinterpretation at the level of physics Physics . Nothing is shown about the mass gap or the construction of the continuum limit
Duality between $\mathbb{Z}_2$ gauge theory and the three-dimensional Ising model	Known (bibliographic data only, not checked)

Sources and reproduction

Item	Kind	Source / tools
Tables 1–3, full enumeration, the check of the variance, the exponent of the prefactor, the numbers in the figures	Computed	python3 (NumPy, SciPy's `eigsh`; the action of order 2^L applied without building the matrix). The figures by JavaScript in the browser
The transfer-matrix method, the one-dimensional chain	Known	H. A. Kramers, G. H. Wannier, Phys. Rev. 60 (1941) 252. Abstract checked
$\xi_L = 1/\log(\lambda_0/\lambda_1)$, (10), the dual coupling, the form of (12)	Known	S. L. A. de Queiroz, Phys. Rev. E 84 (2011) 031107, arXiv:1105.6248. Equations (1)–(9) of § I and § II checked. The review summarising the closed forms, C. Domb, Adv. Phys. 9 (1960) 149, bibliographic data only
The general relation of (12) (from conformal invariance)	Physics	J. L. Cardy, J. Phys. A 17 (1984) L385. Abstract checked

Item	Kind	Source / tools
The closed form of the eigenvalues of the strip, the free energy of the boundary (the origin of (13))	Known	L. Onsager, Phys. Rev. 65 (1944) 117. Abstract checked (that it treats the eigenvalue problem of the cylindrical strip and the free energy of the boundary; the form of formula (13) not checked)
Free fermions / no phase transition in one dimension	Known	T. D. Schultz, D. C. Mattis, E. H. Lieb, Rev. Mod. Phys. 36 (1964) 856 / E. Ising, Z. Phys. 31 (1925) 253. Bibliographic data only for both
Z_2 gauge theory and duality	Known	F. J. Wegner, J. Math. Phys. 12 (1971) 2259 (bibliographic data only, not checked)
The transfer matrix of lattice gauge theory and reflection positivity	Known	K. Osterwalder, E. Seiler, Ann. Phys. 110 (1978) 440 (bibliographic data only, not checked)
Finite-size effects of first-order transitions on a finite strip	Known	V. Privman, M. E. Fisher, J. Stat. Phys. 33 (1983) 385 (bibliographic data only, not checked)

Next chapter to read: 2-10 The relaxation time and the dynamical exponent z. To read alongside: 3-01 The gauge principle. The main treatment of the definition of the gap is in 4-11, and the transfer matrix of the lattice in 5-03.

The relaxation time and the dynamical exponent z — by what power of the correlation length does the approach to equilibrium slow down

> This chapter organises established knowledge. § 05 is this book's own investigation.

What this chapter shows — Measuring the relaxation time of dynamics that produce the same equilibrium distribution (Langevin, Glauber, heat bath) by the Poincaré constant C_P. That for the Gaussian chain and one-dimensional Ising the relaxation time is the square of the correlation length ($z = 2$), and that at the critical point of two-dimensional Ising it is slower than that. The definitions of the dynamical exponents z and $\eta_{\mathrm{dyn}} = 2 - z$, and known examples in which the pattern $z = 2$ breaks.

Prerequisite chapters — 2-08 (the correlation length, critical points, the Gaussian chain) and 2-09 (the transfer matrix of the Ising model). It is easier if you have read the Poincaré inequality in 1-10 § 07 and Brownian motion in 1-09 § 06.

Said first — Only the Gaussian chain and the one-dimensional chain are exact. The values of z (2.1665 and so on) are numerical results and predictions of physics, and the computations on this machine for two-dimensional Ising are on small boxes of side up to 5. This chapter shows nothing about the mass gap of Yang–Mills.

Order of this chapter

1. Adding time — a photograph of equilibrium, and the machine that makes it
2. Langevin dynamics — the inverse of the gap is the Poincaré constant
3. Glauber dynamics (Table 1) — in one dimension $\tau \approx 2\xi^2$
4. Langevin for the Gaussian chain (Figure 1, Table 2) — $C_P = \xi^2$ exactly in the continuum limit
5. Small boxes of two-dimensional Ising (Table 3, Figure 2) — C_P/L^2 grows at the critical point
6. The dynamical exponents z and η_dyn — $\eta_{\mathrm{dyn}} := 2 - z$ is defined here
7. Known examples with z ≠ 2 (Table 4, Detour: z = 4 for conserving dynamics)

8. Do not identify the heat bath with Langevin — constants, units, choice of variables
9. A bridge to Yang–Mills — what is known and what is not

01

Adding time — a photograph of equilibrium, and the machine that makes it

What 2-08 and 2-09 treated was the equilibrium distribution, that is, "which configurations appear with what probability". There is no time in it. But when a computer draws samples from an equilibrium distribution, or when a physical system in contact with a heat bath moves towards equilibrium, the configuration moves in time. This chapter measures the speed of that movement.

As a metaphor, the equilibrium distribution is a single photograph, and the dynamics is the machine that makes it. Among machines that make the same photograph, some are fast and some slow. Precisely, a Markov process (a random motion whose next move depends only on the current configuration) leaving a given probability measure μ invariant is not unique, and the way it approaches equilibrium depends on the choice.

The time s of this chapter is a different thing from the imaginary time of 2-09 (for the contrast see 18-03 § 05).

Imaginary time (2-07, 2-09)	The dynamical time s (this chapter)
One direction **within** the configuration. The correlation $\langle \sigma_0 \sigma_r \rangle$ decays in a way that sets the correlation length ξ, and $1/\xi$ is the gap of the transfer matrix.	An external time in which the **whole** configuration moves. The decay of autocorrelations sets the relaxation time τ, and $1/\tau$ is the gap of the generator.

The subject is the relation $\tau \approx \xi^z$, and z is called the dynamical exponent. The assumptions of the later parts (A-06 Index of names) have the form of placing $z = 2$ at the level of a power. Here we first line up examples where that form holds and examples where it breaks.

02

Langevin dynamics — adding fluctuations to a motion down a slope

Suppose we want a single real number X to follow the probability density $e^{-V(x)}/Z$. If we move it in the direction down the slope V and at the same time shake it with random numbers, the distribution stops where the two effects balance. This is Langevin dynamics.

$$dX = -V'(X)\,ds + \sqrt{2}\,dW(s) \quad (1)$$

W is Brownian motion (1-09 § 06), and dW is a Gaussian increment with mean 0 and variance ds over time ds. The coefficient $\sqrt{2}$ is chosen so that the distribution at which it stops is exactly e^{-V}.

The operator producing the time change of the expectation of a function f is called the generator, written $\mathcal{L}$. Expanding $f(X+dX)$ to second order, the first-order term in dX leaves $-V'f'\,ds$ in the direction of the slope, and the second-order term leaves $\frac{1}{2}f''\cdot 2\,ds$ since the mean of dW^2 is ds (orders higher than ds are dropped). Thus for (1),

$$\mathcal{L}f = f'' - V'f', \qquad \frac{d}{ds}\,\mathbb{E}\,f(X_s) = \mathbb{E}\,(\mathcal{L}f)(X_s), \qquad \int f\,(-\mathcal{L}g)\,d\mu = \int f'g'\,d\mu \quad (2)$$

($\mu = e^{-V}\,dx/Z$; the last equality is integration by parts). By the last equality $-\mathcal{L}$ is symmetric in $L^2(\mu)$ and non-negative. It is 0 only for constants, and we write the next eigenvalue (the gap) as $\mathrm{gap}(\mathcal{L})$.

The inverse of the gap is the Poincaré constant (1-10 § 07).

$$\mathrm{Var}_\mu f \;\le\; C_P \int |f'|^2\,d\mu \quad (\text{for all } f), \qquad C_P = \frac{1}{\mathrm{gap}(\mathcal{L})} \quad (3)$$

Running from equilibrium, the autocorrelation of every observable decreases faster than e^{-s/C_P}, and there is an observable that decreases at exactly this rate. That is, the slowest relaxation time $\tau_{\exp}$ equals C_P Known. This equality is a property of reversible dynamics in continuous time; for chains in discrete time it splits into three quantities (18-03 § 06).

Example. For $V = x^2/(2\sigma^2)$, (1) is the Ornstein–Uhlenbeck process, $f = x$ is an eigenfunction, and $\mathcal{L}x = -x/\sigma^2$. Hence $C_P = \sigma^2$: the larger the variance (the wider the tails), the slower the relaxation. With many variables, we set up (1) for the field ϕ_x at each point,

$$d\phi_x = -\frac{\partial S}{\partial \phi_x}\,ds + \sqrt{2}\,dW_x(s) \quad (4)$$

(W_x independent at each point), and replace the $|f'|^2$ of (3) by the squared length of the gradient $|\nabla f|^2$. The distribution at which it stops is e^{-S}/Z. In field theory this is called stochastic quantisation (7-09).

03

Glauber dynamics — flipping spins one at a time

The spins of the Ising model are ± 1, so they cannot move continuously down a slope. Instead, each spin carries an independent clock, and when the clock rings the spin is redrawn from its conditional distribution. This is called the heat-bath rule, and the dynamics doing this in continuous time is called Glauber dynamics (Glauber 1963 analysed a process of this type in detail on a closed one-dimensional chain) Known .

With K including the inverse temperature in the coupling (as in 2-09), let $\mu(\sigma) \propto \exp(K \sum_{\langle xy \rangle} \sigma_x \sigma_y)$. Writing the configuration with spin j flipped as σ^j and the sum of neighbours as $h_j = \sum_{y \sim j} \sigma_y$, the heat-bath flip rate is

$$w_j(\sigma) = \frac{1}{1 + e^{2K \sigma_j h_j}}, \qquad \mu(\sigma)\, w_j(\sigma) = \mu(\sigma^j)\, w_j(\sigma^j) \tag{5}$$

The right-hand formula is detailed balance: it says that "the probability flowing from σ to σ^j" equals the flow in the opposite direction. From this, μ is the distribution at which it stops, and the dynamics is reversible. The more aligned a spin ($\sigma_j h_j > 0$), the less readily it flips. The generator is $(\mathcal{L} f)(\sigma) = \sum_j w_j(\sigma)\big(f(\sigma^j) - f(\sigma)\big)$, the $\int |f'|^2$ on the right side of (3) is replaced by $\frac{1}{2} \sum_\sigma \mu(\sigma) \sum_j w_j(\sigma)\big(f(\sigma^j) - f(\sigma)\big)^2$, and $CP = 1/\mathrm{gap}(\mathcal{L})$ is the same.

The one-dimensional chain can be solved

In one dimension $h_j \in \{-2, 0, 2\}$, so we can write $w_j = \frac{1}{2}\big(1 - \frac{g}{2}\sigma_j(\sigma_{j-1} + \sigma_{j+1})\big)$, $g = \tanh 2K$. The expectation of a spin changes by 2 in one flip, so $\frac{d}{ds}\langle \sigma_j \rangle = -2\langle \sigma_j w_j \rangle$, and

$$\frac{d}{ds}\langle \sigma_j \rangle = -\langle \sigma_j \rangle + \frac{g}{2}\big(\langle \sigma_{j-1} \rangle + \langle \sigma_{j+1} \rangle\big) \tag{6}$$

closes. The mean of the total magnetisation $M = \sum_j \sigma_j$ decreases like $e^{-(1-g)s}$. On the other hand, the static correlation is $\langle \sigma_0 \sigma_r \rangle = (\tanh K)^r$ (2-09 § 03), and

$$\tau = \frac{1}{1 - \tanh 2K}, \qquad \xi = \frac{-1}{\log \tanh K}, \qquad \frac{\tau}{\xi^2} \xrightarrow[K \to \infty]{} 2 \tag{7}$$

The limits follow from $1-\tanh 2K \approx 2e^{-4K}$, $-\log\tanh K \approx 2e^{-2K}$. The colder it is, the longer the correlations, and relaxation slows by its square: for this dynamics $z=2$.

Computed Table 1 One-dimensional Glauber dynamics. We diagonalised the generator of the ring of $N=8,\ 10$ points on all states (2^N of them) and checked that the gap agrees with $1-\tanh 2K$ to 10 digits for all of $K=0.25,\ 0.5,\ 1,\ 1.5$. On these two rings the decay of magnetisation is the slowest relaxation. For general N, the decay rate of magnetisation $1-\tanh 2K$ gives an upper bound on the gap (since magnetisation is an eigenfunction of $\mathcal{L}$).

K	ξ	$\tau = 1/(1-\tanh 2K)$	τ/ξ^2
0.50	1.295	4.195	2.499
1.00	3.672	27.80	2.062
1.50	10.03	202.2	2.008
2.00	27.30	1491	2.001
3.00	201.7	81378	2.00002

04

Langevin for the Gaussian chain — the longer the wave the slower, and the relaxation time is the square of the correlation length

We apply (4) to the Gaussian chain of (7) of 2-08 (the $H(\phi)$ there; here written as the action S) $S=\frac{1}{2}\sum_x[(\phi_{x+1}-\phi_x)^2+m^2\phi_x^2]$. Since $\partial S/\partial\phi_x = (2+m^2)\phi_x-\phi_{x+1}-\phi_{x-1}$ is linear, the equations separate for each Fourier component $\hat\phi_q=\sum_x e^{-iqx}\phi_x$ of wave number q (1-11).

$$d\hat\phi_q = -\omega_q\,\hat\phi_q\,ds + (\text{noise}), \qquad \omega_q = m^2 + 4\sin^2\frac{q}{2} \tag{8}$$

Every component is the Ornstein–Uhlenbeck process of (1), decreasing at rate ω_q. The slowest is the uniform component $q=0$, with rate m^2. Therefore $C_P = 1/m^2$ (the same number as "the C_P of a Gaussian measure is the largest eigenvalue of the covariance" of 1-10 § 09) Known .

The correlation length was $\cosh(1/\xi) = 1+m^2/2$ by (8) of 2-08. Using $\cosh u - 1 = 2\sinh^2(u/2)$, the two are connected by a single formula.

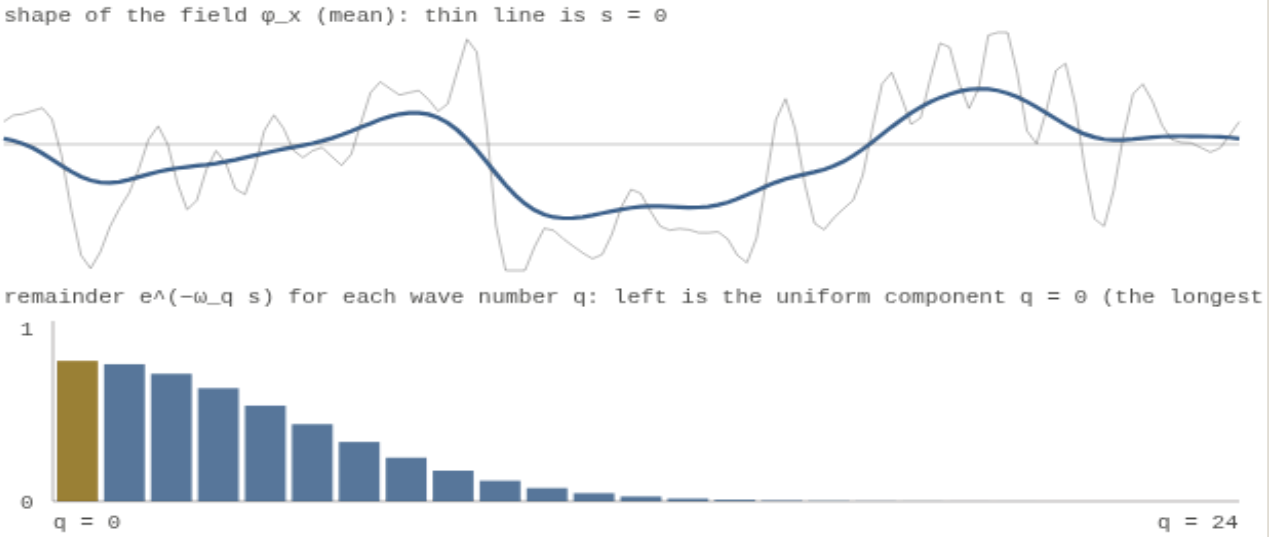
$$\frac{1}{C_P} = m^2 = 4\sinh^2\frac{1}{2\xi}, \qquad \frac{C_P}{\xi^2} = \Big(\frac{1/(2\xi)}{\sinh(1/(2\xi))}\Big)^2 \xrightarrow[\xi\to\infty]{} 1 \tag{9}$$

This is an exact equality; on the lattice C_P is slightly smaller than ξ^2, and becomes $C_P/\xi^2 \to 1$ in the continuum limit ($\xi\to\infty$). For the continuous free field $C_P=\xi^2$ holds exactly Known (an elementary derivation; the numbers in Table 2 are Computed).

Why the square

Why the square. The mean motion of (8), with the noise removed, is the heat equation with mass $\frac{d}{ds}\phi = \Delta_{\text{lat}}\phi - m^2\phi$. Diffusion takes a time of order ℓ^2 to smooth out bumps of length ℓ (as for a random walk), and the mass term removes waves longer than ξ at a uniform rate m^2. So the time is ξ^2. Precisely, the content of this explanation is that the ω_q of (8) approaches $m^2 + q^2$ for small q, and bottoms out at m^2 for $q \lesssim 1/\xi$.

Figure 1 The mean motion of the Gaussian chain (a ring of 128 points). Making m smaller lengthens, like $1/m^2$, the time for which the long waves (bars on the left) survive. Moving the time s smooths out the shape above, and the bars below (the remainder $e^{-\omega_q s}$ for each wave number) disappear in order from the right



m = 0.158, ξ = 6.33, C_P = 1/m² = 40.0, s = 10 (s/C_P = 0.25), remainder of the uniform component 0.779

Computed The curves in the figure are the exact solution $\hat{\phi}_q(s) = e^{-\omega_q s}\hat{\phi}_q(0)$ of (8), computed on the spot (the initial shape was made with random numbers from a fixed seed). This is the mean motion with the noise removed, not a sample of the fluctuations.

Computed Table 2 ξ and C_P of the Gaussian chain. The last two columns are the heat-bath lower bound of section 08 (determined by the decrease on linear functions)

m	ξ	$C_P = 1/m^2$	C_P/ξ^2	**Heat-bath lower bound** $(2+m^2)/m^2$	**Lower bound / Lan**
1.00	1.039	1	0.9263	3	1
0.50	2.020	4	0.9798	9	2
0.20	5.008	25	0.9967	51	2
0.10	10.004	100	0.99917	201	2
0.02	50.001	2500	0.999967	5001	2

A table set beside the gap of the transfer matrix is in Table 1 of 18-03, and a table measured with a checkerboard heat bath is in 5-07 § 07.

```
05
```

Small boxes of two-dimensional Ising — at the critical point C_P/ξ^2 grows

With interactions, the formulas do not close. So on $L \times L$ boxes (periodic boundary) of the two-dimensional Ising model we diagonalise the generator of Glauber dynamics (5) numerically and find C_P. There are 2^{L^2} states, about 34 million for $L = 5$. The slowest relaxation is expected to lie in the direction of magnetisation (the subspace invariant under translations and changing sign under the global flip). For $L = 3,\ 4$ the gap of the whole space and the gap of this subspace agreed to 8 digits at all seven temperatures below Computed . For $L = 5$ the whole space cannot be handled, so we found the gap restricted to this subspace (671,104 dimensions). That value is at least the gap of the whole space, so the C_P for $L = 5$ is a lower bound on the true C_P.

At the critical point $K_c = \frac{1}{2}\log(1+\sqrt{2})$ (2-08 § 04) the ξ of the infinite system is infinite, and the length over which correlations reach inside a box levels off at the box size L (2-09 § 07). So instead of ξ we use L as the ruler. If $\tau \approx \xi^z$ is right, then

$$C_P(L) \approx L^z, \qquad \frac{C_P(L)}{L^2} \approx L^{\,z-2}, \qquad z_{\text{eff}}(L, L') = \frac{\log\big(C_P(L')/C_P(L)\big)}{\log(L'/L)} \quad (10)$$

z_{eff} is an effective exponent read from neighbouring sizes, and is expected to approach z as $L \to \infty$.

Computed Table 3 Two-dimensional Ising, continuous-time heat bath (each spin redrawn at rate 1), $C_P = 1/\text{gap}$ with $L \times L$ periodic boundary. The unit of time is "one attempt per spin"

K/K_c	$L=3$	$L=4$	$L=5$	$z_{\text{eff}}(3,4)$	$z_{\text{eff}}(4,5)$	To Figure 2
0.6	5.759	6.267	6.474	0.29	0.15	Select
0.8	14.56	19.34	22.83	0.99	0.74	Select
0.9	24.56	38.47	52.26	1.56	1.37	Select
1.0	42.89	82.84	137.4	**2.29**	**2.27**	Select
1.1	77.18	191.0	408.0	3.15	3.40	Select
1.2	142.4	464.6	1324	4.11	4.69	Select
1.5	997.9	7906	56656	7.19	8.83	Select

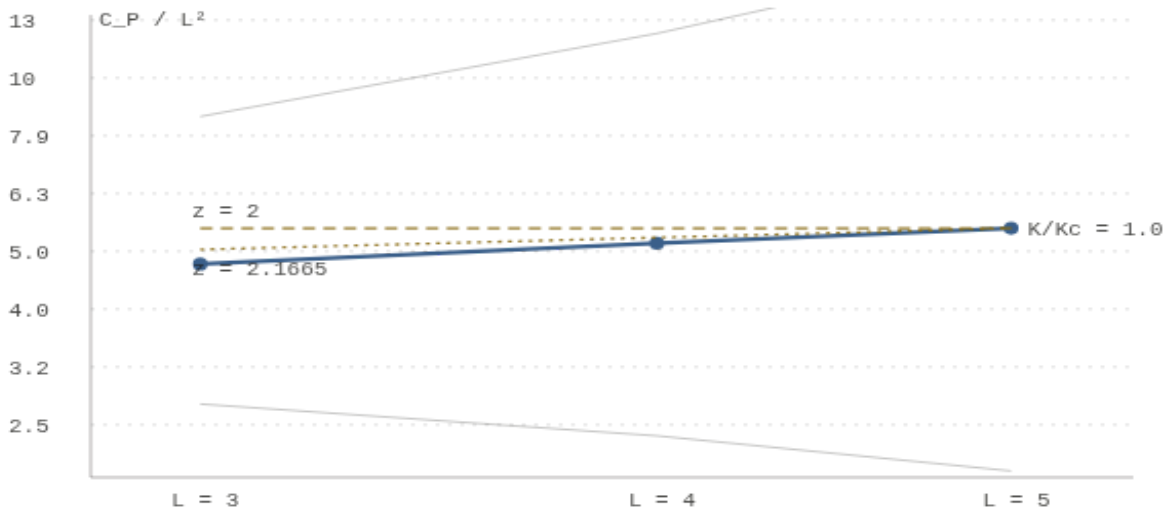
Three behaviours

Three behaviours can be read. On the high-temperature side ($K/K_c = 0.6$), C_P levels off around 6, and z_{eff} tends to 0: if ξ is finite, so is the relaxation time. On the low-temperature side ($K/K_c = 1.5$), $\log C_P$ increases with L by an almost constant difference (the differences are 2.07, 1.97), growing more like an exponential than a power. This is thought to be because moving from "almost all +" to "almost all −"

requires making a wall across the box, whose weight becomes small like an exponential e^{-cL} proportional to the length of the wall Physics . In a box with periodic boundary two walls are needed, so c is expected to be about twice the interface tension σ. However, the $2\sigma \approx 1.55$ at $K/K_c = 1.5$ and the difference of about 2 for $L \le 5$ do not yet agree, and what this box can say goes only as far as "a tendency to grow exponentially". A similar picture is the closing gap of 2-09 § 08.

At the critical point C_P/L^2 keeps increasing as $4.77,\ 5.18,\ 5.50$, and z_{eff} is $2.29,\ 2.27$. It lies on the side greater than 2. The value known from numbers on larger boxes (up to 15×15) is $z = 2.1665(12)$ (Nightingale–Blöte 1996, heat bath) Physics ; the values from the small boxes on this machine can be read as approaching it from above, but two points cannot decide this. What can be said goes only as far as "in the range $L \le 5$, C_P/L^2 does not settle to a constant".

Figure 2 C_P/L^2 for two-dimensional Ising (log–log; the vertical axis is magnified to suit the selected temperature). Moving the temperature changes the selected line (●) and the z_{eff} in the readout on the right. The dashed line is z = 2 (horizontal), the dotted line the slope of z = 2.1665 (both passing through the point at L = 5). "Play" changes the temperature in turn. You can also select with "Select" in Table 3



K/Kc = 1.0, C_P = 42.89, 82.84, 137.4 (L = 3, 4, 5), z_eff = 2.29 (3→4), 2.27 (4→5)

Computed The numbers in this figure were computed on this machine (the values of Table 3 embedded). L = 2 is excluded, as it is a special box in which neighbours coincide.

06

The dynamical exponents z and η_{dyn} — naming how far the power departs from 2

In physics one puts forward the hypothesis that as ξ grows the relaxation time grows as a power of ξ. It is called the dynamic scaling hypothesis, and is held to be the extension of static scaling (2-08 § 06) to correlations involving time (Ferrell et al. 1967,

Halperin–Hohenberg 1967, with a detailed development in 1969) Physics . Put in mathematical language for reversible continuous-time dynamics,

$$C_P \asymp \xi^{\,z}, \qquad \eta_{\mathrm{dyn}} := 2 - z, \qquad C_P \asymp \xi^{\,2-\eta_{\mathrm{dyn}}} \quad (11)$$

$A \asymp B$ means that the ratio A/B is bounded above and below by positive constants. In this book η_{dyn} is defined by this formula. The name comes from its being the quantity that measures, as a power, the departure from the $C_P \approx \xi^2$ of the free field.

It is a different quantity from the η of (13) of 2-08. The η of 2-08 is the departure in the power of the static correlation $G(x) \sim |x|^{-(d-2+\eta)}$ at the critical point, and is written η_s when it must be distinguished from the dynamic quantity. The two coincide only within the conventional theory (called by the name of van Hove) that holds "the relaxation time is proportional to the susceptibility"; for two-dimensional Ising, $\eta_s = 1/4 > 0$ while $\eta_{\mathrm{dyn}} \approx -0.17 < 0$, so they differ even in sign (Table 2 in 18-03 § 07).

Pattern z = 2	Too slow	Too fast
$z = 2$, $\eta_{\mathrm{dyn}} = 0$	$z > 2$, $\eta_{\mathrm{dyn}} < 0$	$z < 2$, $\eta_{\mathrm{dyn}} > 0$
Holds for Langevin on the Gaussian chain ((9)) and one-dimensional Glauber ((7))	Glauber for two- and three-dimensional Ising (numerical), conserving dynamics	Non-local updates (cluster methods). For local dynamics the lower bound $z \geq \gamma/\nu$ (γ, ν are the static exponents of the susceptibility and the correlation length) follows under two scaling hypotheses (18-03 § 07) Physics . For local Ising dynamics, $z \geq 2$ at the critical point is a theorem (§ 07)

The diffusive-relaxation-time assumption of Part 23 (23-01) is the statement $\lim \ln C_P / \ln \xi = 2$ ($C_P = \xi^{2+o(1)}$, $\eta_{\mathrm{dyn}} = 0$) that places this left-hand form at the level of a power, for the Langevin dynamics of lattice Yang–Mills. The band statement $C_P \asymp \xi^2$ is called the band form of the diffusive-relaxation-time assumption in 23-01, and is treated separately from the diffusive-relaxation-time assumption (A-06 Index of names). Whether it holds is an unproved assumption, and the physical counterpart is set out in the form of questions in 18-05 § 03. What this chapter wants to confirm is that the form itself is not trivial: change the dynamics once and it breaks. The example is the next section.

07

Known examples with $z \neq 2$ — three ways the pattern breaks

Table 4 lists known examples in which the pattern $CP \asymp \xi^2$ breaks. The ways of breaking divide into three.

Interaction shifts the power. For Glauber dynamics of two- and three-dimensional Ising, numerically z slightly exceeds 2. Even with the rule kept local, interaction moves the power. In mathematics it has been shown that for Glauber dynamics of two-dimensional critical Ising the inverse of the gap is bounded by a polynomial in L (Lubetzky–Sly) Known . From below, it has been shown that at the critical point with periodic boundary (and on the low-temperature side) the heat-bath gap is at most $C(\log L/L)^2$, that is, $z \geq 2$ (Masaoka–Soejima–Watanabe 2025; extending to local reversible single-spin-flip dynamics) Known . At the two-dimensional critical point $(L/\log L)^2 \lesssim CP(L) \leq \mathrm{poly}(L)$ is a theorem, and the "too fast" side of § 06 does not occur for local Ising dynamics. No theorem determining the value of the power z itself has been found.

Conserved quantities slow it down. In dynamics that do not change the total magnetisation (exchanging neighbouring spins), the uniform component cannot move, and long waves are carried by diffusion, so it is slower still. The prediction of physics is held to be $z = 4 - \eta_s$ (model B in the classification of Hohenberg–Halperin) Physics .

The way of updating speeds it up. The Swendsen–Wang cluster method is non-local dynamics that flips whole clusters, and its abstract says it breaks dynamic universality, giving 'unusually small values of the dynamical critical exponent' Physics . Even for the same equilibrium distribution, changing the machine changes z.

Physics Table 4 Example values of z (numerical and predicted values from the abstracts of primary sources; $\eta_{\mathrm{dyn}} = 2 - z$ computed on this machine)

Model, dynamics	z	η_{dyn}	Level
Gaussian chain, Langevin ((9)) / one-dimensional Ising, Glauber ((7))	2	0	Known (exact)
Two-dimensional Ising, heat bath (Glauber)	2.1665(12)	−0.1665	Physics numerical
Three-dimensional Ising, Glauber	2.032(3)	−0.032	Physics numerical
Conserving dynamics (model B), three-dimensional Ising class	3.972(2)	−1.972	Physics numerical (prediction $4 - \eta_s$)
Conserving dynamics, Gaussian chain (detour below; measured by the relaxation time of the component of wave number $1/\xi$)	4	−2	Computed (exact)

Model, dynamics	z	η_{dyn}	Level
Pure-gauge HMC, square of the topological charge	About 5	About −3	Physics numerical (18-03 § 08)

The last row is a slowness that actually appears in computations of lattice gauge theory. HMC (hybrid Monte Carlo; an update method that adds fictitious momenta and moves deterministically; 5-07 § 08) is dynamics different from Langevin (next section), and whether the same slowness occurs for Langevin is treated in 23-02 Roads tried on the four-dimensional lattice, and walls.

Detour: $z = 4$ for conserving dynamics

You may skip this. In the Gaussian chain, if the drive $-\partial S/\partial\phi$ of (4) is passed once more through the lattice Laplacian Δ_{lat}, and the noise too is put in the form of differences between neighbours, then $\sum_x \phi_x$ is conserved. The decay rate of (8) is replaced by

$$\omega_q^{(\mathrm{B})} = 4\sin^2\frac{q}{2}\left(m^2 + 4\sin^2\frac{q}{2}\right) \tag{12}$$

and $q = 0$ stops (being a conserved quantity, it does not decrease). The relaxation time of the component $q = 1/\xi$ with wavelength about the correlation length is $1/\omega^{(\mathrm{B})}_{1/\xi} \approx \xi^4/2$. The computation on this machine gave $\tau/\xi^4 = 0.510,\ 0.5004,\ 0.50002$ for $m = 0.5,\ 0.1,\ 0.02$ Computed . On the infinite chain the components with $q \to 0$ are as slow as we like, so the gap is 0 and $C_P = \infty$. The pattern $C_P \asymp \xi^2$ loses its meaning here.

08

Do not identify the heat bath with Langevin — the same photograph, different machines

"The relaxation time" is not determined until we say whose dynamics' time it is. The heat bath and Langevin have different generators for the same distribution.

We compare them on the Gaussian chain. For the continuous-time heat bath (each point, at rate 1, redrawn from the conditional distribution $\mathcal{N}\big((\phi_{x-1}+\phi_{x+1})/(2+m^2),\ 1/(2+m^2)\big)$ with its neighbours fixed), the mean of the field values moves by

$$\frac{d}{ds}\,\mathbb{E}\,\phi_x = -\phi_x + \frac{\phi_{x-1}+\phi_{x+1}}{2+m^2} = -\frac{1}{2+m^2}\,\frac{\partial S}{\partial \phi_x} \tag{13}$$

Linear functions are eigenfunctions of this generator, and the decay rate of the uniform component is $m^2/(2+m^2)$. Therefore the C_P of the heat bath is at least $(2+m^2)/m^2$ Computed (checked with the matrix of a ring of 64 points). This lower bound is about twice ξ^2, so for the heat bath $z \geq 2$. As the right two columns of Table 2 show,

the ratio of the lower bound to the CP of Langevin approaches 2 as $\xi \to \infty$. Even if the powers are the same, the constants differ.

Differences other than the constant

The difference is not only in the constant.

- **The choice of variables.** The heat bath is the same process even if the variable at each point is relabelled ($\psi_x = c\,\phi_x$ and so on), because the operation of drawing from a conditional distribution does not depend on coordinates. The CP of Langevin is measured by the length of the gradient, so measured in $\psi = c\,\phi$ it becomes c^2 times as large. This is why in lattice gauge theory CP changes by a factor β depending on whether one measures by "the angle of the link ϑ or the field $A = \sqrt{\beta}\,\vartheta$". Since $\beta \sim \ln \xi$, the power statement (the diffusive-relaxation-time assumption) does not change, but the band statement does (1-10 § 09, 23-01 The line of construction and the diffusive-relaxation-time assumption).
- **The unit of time.** The time of the heat bath is "the number of attempts per point", and the time of Langevin is the s of (1). The heat bath that advances one sweep at a time in discrete time (5-07 § 07) is yet another quantity, and CP, the relaxation time and $\tau_{\exp}$ split apart (18-03 § 06).
- **The exponent.** In physics, two local reversible dynamics with the same conserved quantities (both of model A type: not conserving magnetisation) are expected to have the same z (the same dynamic universality class) Physics . The conserving dynamics of § 07 has different conserved quantities (model B), so it is not a counterexample to this expectation. For the Gaussian chain, what can be said by formulas is $z = 2$ for Langevin, and, for the heat bath, $z \geq 2$ from the lower bound from linear functions. That the gap of the heat bath is determined by linear functions, with equality, is not shown in this chapter. In general, this chapter has no theorem showing that the z of two dynamics are equal.

When this book writes CP, unless stated otherwise it is the inverse of the gap of the generator of Langevin (4), with the variable whose gradient it is measured by stated alongside. Reading numbers from the heat bath or HMC as the CP of Langevin is an assumption.

09

A bridge to Yang–Mills — where it is used, and what is not known

Langevin dynamics (4) can be set up for lattice Yang–Mills too. Since the variables are group elements on the links, the gradient is taken as a derivative on the group (1-03).

Doing the same for a continuous field is stochastic quantisation (7-09), and drawing samples from the same distribution on a computer is the Monte Carlo method (5-07).

At strong coupling (the side where β is small), the Poincaré inequality and the log-Sobolev inequality (an inequality stronger than (3), controlling the decrease of entropy) have been shown for lattice Yang–Mills (Shen–Zhu–Zhu 2023) Known . This is a result on the side where ξ is finite. In four dimensions there is no theorem determining the power of C_P as $\xi \to \infty$.

From this chapter onwards

Putting the pattern $z = 2$ for Langevin dynamics of lattice Yang–Mills as the statement at the level of powers $\lim \ln C_P / \ln \xi = 2$ ($C_P = \xi^{2+o(1)}$, $\eta_{\mathrm{dyn}} = 0$) is 23-01 The line of construction and the diffusive-relaxation-time assumption. The band statement $C_P \asymp \xi^2$ is called the band form of the relaxation-time assumption in 23-01 and treated separately. The metaphorical entrance is 23-00 (the smearing time and the ruler), the reconciliation with the vocabulary of physics is 18-05 § 02 (fixing the η of the record to η_{dyn}), and the dictionary of the two gaps ξ and C_P is 18-03 § 05– § 07. This chapter shows nothing about the mass gap.

What this chapter can say

Content	Status
The generator (2) of Langevin (1)(4), $\tau_{\exp} = C_P$ (reversible dynamics in continuous time), detailed balance (5) of the heat bath	Said Known
One-dimensional Glauber dynamics (6)(7), the decay rate of magnetisation $1 - \tanh 2K$, $\tau/\xi^2 \to 2$	Said Known ((6) was derived in this chapter). The gap $= 1 - \tanh 2K$ is Computed for rings with $N = 8, 10$. For general N the decay rate of magnetisation gives an upper bound on the gap
(8)(9) of the Gaussian chain, $C_P/\xi^2 \to 1$, (12) of conserving dynamics and $\tau \approx \xi^4/2$, (13) of the heat bath and the lower bound $C_P \geq (2+m^2)/m^2$ (for the heat bath, as far as $z \geq 2$)	Said (the formulas are exact) Known (elementary derivations). The numbers in Table 2 and (12) are Computed
Table 3: C_P and z_{eff} of two-dimensional Ising with $L \leq 5$	Computed (exact diagonalisation; the restriction to the magnetisation subspace is checked against the whole space for $L \leq 4$, and for $L = 5$ the subspace value is a lower bound on the true C_P). It does not determine z; it is a check of tendencies

Content	Status
Dynamic scaling $C_P \asymp \xi^z$, the values of z in Table 4, the $4-\eta_s$ of model B, that two local dynamics have the same z	Physics (hypotheses, numbers, predictions). For two-dimensional critical Ising, $(L/\log L)^2 \lesssim C_P(L) \le \mathrm{poly}(L)$ ($z \ge 2$ from below, polynomial from above) is Known ; the value of the power has not become a theorem
$\eta_{\mathrm{dyn}} := 2 - z$	This book's definition (common to the later chapters)
The dependence of C_P on ξ for Yang–Mills	Not said. There is no theorem determining it in four dimensions, and both the relaxation-time assumption (the statement at the level of powers, 23-01 § 05) and the band statement (the band form of the relaxation-time assumption) remain assumptions. This chapter shows nothing about the mass gap or the construction of the continuum limit

Sources and reproduction

Item	Kind	Source / tools
Tables 1, 2 and 3, the numbers in (9)(12)(13), Figures 1 and 2	Computed	python3 (NumPy, SciPy; for Table 3 the generator was made a sparse matrix on the subspace reduced by translations and the global flip, and the smallest eigenvalue found by the Lanczos method). The figures by JavaScript in the browser. The scripts are kept in the record of computations
Glauber dynamics, the one-dimensional chain	Known	R. J. Glauber, "Time-dependent statistics of the Ising model", J. Math. Phys. 4 (1963) 294–307. Abstract checked (analyses the Markov process of a closed chain of N points in contact with a heat bath, and finds the expectations and correlations of the spins explicitly). The text has not been read; (6) was derived in this chapter
The Poincaré constant, the difference between continuous and discrete time	Known	D. A. Levin, Y. Peres, E. L. Wilmer, *Markov Chains and Mixing Times* 2nd ed., § 12.2 (the same as in the table of sources of 18-03)
The dynamic scaling hypothesis	Physics	R. A. Ferrell, N. Menyhárd, H. Schmidt, F. Schwabl, P. Szépfalusy, Phys. Rev. Lett. 18 (1967) 891–894. Abstract checked (derives the dispersion of second sound $\omega \propto k^{3/2}$ from the absence of a characteristic length at the lambda point). B. I. Halperin, P. C. Hohenberg, Phys. Rev. Lett. 19 (1967) 700–703. Abstract checked (generalises the Widom–Kadanoff scaling laws to dynamic phenomena under an assumption on time-dependent correlation functions). The detailed development is B. I. Halperin, P. C. Hohenberg, Phys. Rev. 177 (1969) 952 (abstract only). The classification (model A, B) is P. C. Hohenberg, B. I. Halperin, Rev. Mod. Phys. 49 (1977) 435–479. Abstract checked (identifies and characterises the main dynamic universality classes; the names model A and B are not in the abstract, and the text has not been read)

Item	Kind	Source / tools
$z = 2.1665(12)$ of two-dimensional Ising	Physics	M. P. Nightingale, H. W. J. Blöte, Phys. Rev. Lett. 76 (1996) 4548. Abstract checked (heat bath, random choice of sites, up to 15×15)
$z = 2.032(3)$ of three-dimensional Ising	Physics	Z. Liu, E. Vatansever, G. T. Barkema, N. G. Fytas, Phys. Rev. E 108 (2023) 034118, arXiv:2307.01837. Abstract checked
$z = 3.972(2)$ of conserving dynamics and the prediction $z = 4 - \eta$	Physics	C. Chattopadhyay, J. Ott, T. Schäfer, V. Skokov, Phys. Rev. D 108 (2023) 074004, arXiv:2304.07279 (volume and article number checked with Crossref). Text checked (the prediction is attributed to Hohenberg–Halperin 1977)
The small z of cluster methods	Physics	R. H. Swendsen, J.-S. Wang, Phys. Rev. Lett. 58 (1987) 86. Abstract checked
The inverse of the gap of two-dimensional critical Ising is polynomial in L	Known	E. Lubetzky, A. Sly, Commun. Math. Phys. 313 (2012) 815–836, arXiv:1001.1613. Abstract checked (the inverse of the gap is polynomial for any boundary condition)
$z \geq 2$ for Ising in two or more dimensions (gap $\leq C(\log L/L)^2$, periodic boundary, $\beta \geq \beta_c$)	Known	R. Masaoka, T. Soejima, H. Watanabe, J. Stat. Phys. 192 (2025) 76, arXiv:2502.09908. Abstract checked
The slowness of the topological charge in pure-gauge HMC	Physics	S. Schaefer, R. Sommer, F. Virotta, Nucl. Phys. B 845 (2011) 93 (abstract only; the same as 18-03 § 08)
Poincaré and log-Sobolev for lattice Yang–Mills at strong coupling	Known	H. Shen, R. Zhu, X. Zhu, Commun. Math. Phys. 400 (2023) 805 (abstract only; the same as 18-03)

Next chapter to read: 3-01 The gauge principle. This chapter continues in 5-07 § 07 (critical slowing down of the Monte Carlo method), 18-03 § 05– § 07 (the dictionary of the two times) and 23-01 (the relaxation-time assumption).

The gauge principle — demanding a local symmetry requires a tool for comparison, and that tool carries the force

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That demanding a (local) symmetry that turns the phase point by point breaks the derivative, and that a tool for comparing neighbouring points (a connection) repairs it. That this tool becomes a field, couples to the conserved charge, and carries a force. That for SU(2) the gauge field itself is also turned.

Prerequisite chapters — 2-03 (gauge transformations, phase factors) and 1-04 (SU(2), the adjoint representation). Beyond these we use only Noether's theorem from 2-01 and the indices of 2-04 § 07. We take $\hbar = c = 1$, $\varepsilon_0 = 1$ (Heaviside–Lorentz units).

Said first — The gauge principle is a guide for choosing the form of a theory, not a theorem (section 07). The non-abelian field strength and self-interaction are left to 3-02. Only classical fields are treated, and this chapter shows nothing about the mass gap of Yang–Mills.

Order of this chapter

1. Global symmetry — charge is conserved
2. Turning locally — the derivative breaks
3. A tool for comparison — the connection and the covariant derivative (Figure 1, Table 1)
4. Making A a dynamical field — the field strength, and the mass term that cannot be written
5. A force appears — coupling to the conserved charge (Figure 2)
6. The same demand for SU(2) — a field that is itself turned, and the problem of mass
7. What the principle decides, and what it does not (Table 2)
8. Detour: "gauge" was originally a measuring rod
9. Where this book is going

01

Global symmetry — turning everything by the same amount

In 2-03 the potentials $\mathbf{A}$ came first, and the replacement of phase (G') was found afterwards. This chapter goes the other way: it derives the electromagnetic field by demanding that "the theory does not change even if the phase is turned arbitrarily point by point". This procedure is the **gauge principle**.

The starting point is a field $\psi(t, \mathbf{x})$ with complex values, the simplest model of a charged particle. Take the Lagrangian density (2-01, 2-02) to be

$$\mathcal{L}_0 = |\partial_t \psi|^2 - |\nabla \psi|^2 - V(|\psi|^2) \quad (1)$$

(for example $V = M^2|\psi|^2$, where M is the mass of the particle). Turning by a constant α as $\psi \mapsto e^{i\alpha}\psi$ leaves $\mathcal{L}_0$ invariant (a **global U(1) symmetry**; U(1) is the group of complex numbers of absolute value 1, 1-02). Noether's theorem (2-01) gives a conservation law, written with a constant q multiplied in.

$$\rho = iq\big(\psi^*\partial_t\psi - \psi\,\partial_t\psi^*\big), \qquad \mathbf{J} = -iq\big(\psi^*\nabla\psi - \psi\nabla\psi^*\big), \qquad \partial_t\rho + \nabla\cdot\mathbf{J} = 0 \ (2)$$

Example. For a plane wave $\psi = e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)}$ ($\omega > 0$), $\rho = 2q\omega$, $\mathbf{J} = 2q\mathbf{k}$. q becomes the charge in section 05. A global symmetry says that "nothing changes if the reference of phase at every point is shifted all at once".

02

Turning locally breaks the derivative

Let the angle of turning change from point to point, $\psi \mapsto e^{i\alpha(t,\mathbf{x})}\psi$ (a **local** transformation). $|\psi|^2$ does not change, but the derivative acquires an extra term (∂_μ are the components of ∂_t and ∇, $\mu = 0,1,2,3$).

$$\partial_\mu\big(e^{i\alpha}\psi\big) = e^{i\alpha}\big(\partial_\mu\psi + i(\partial_\mu\alpha)\,\psi\big)$$

Example. The gradient energy $|\nabla\psi|^2$ of a constant field $\psi = 1$ is 0, but turning by $\alpha = kx$ gives $\psi = e^{ikx}$, $|\nabla\psi|^2 = k^2$. Only the reference was changed, yet (1) counts it as a different state.

The derivative is the limit of the difference $\psi(x+\varepsilon n) - \psi(x)$ of two neighbouring points divided by ε (n is the unit vector of the direction). If the two values are turned by different angles, the difference mixes "the change in the field" with "the change in the reference". Values at different points cannot be subtracted as they stand.

Why demand local transformations? The abstract of Yang and Mills (1954) says that the usual invariance under isospin rotations (section 06) is not consistent with the idea of localised fields (the content of the abstract is in 3-02 § 01) Known . This is a demand (a principle), not a theorem, and whether nature follows it is decided by experiment (section 07).

03

A tool for comparison — the connection and the covariant derivative

To compare values at different points, we need a correspondence: "which reference of phase at the point y corresponds to the reference at the point x". We call this the **comparator** $U(y,x) \in U(1)$, write $U(y,x)\psi(x)$ for $\psi(x)$ carried to y, and stipulate that under transformations

$$U(y,x) \mapsto e^{i\alpha(y)}\,U(y,x)\,e^{-i\alpha(x)} \ (3)$$

Carrying in the opposite direction is the inverse element, $U(x,y) = U(y,x)^{-1}$. Then the difference $\psi(y) - U(y,x)\psi(x)$ is merely multiplied by $e^{i\alpha(y)}$. Carry before comparing: that is the point.

For two nearby points the comparator is close to 1, and we name the rate of its deviation the field A_μ: $U(x+\varepsilon n, x) = \exp\big(iq\varepsilon\, n^\mu A_\mu(x)\big) + O(\varepsilon^2)$. The opposite direction is its inverse, so $U(x, x+\varepsilon n) = \exp\big(-iq\varepsilon\, n^\mu A_\mu(x)\big) + O(\varepsilon^2)$. Carrying the value at $x+\varepsilon n$ to x with this before subtracting, and taking the limit of the difference quotient with n in the direction μ,

$$D_\mu\psi = \lim_{\varepsilon\to 0}\frac{U(x,x+\varepsilon n)\,\psi(x+\varepsilon n) - \psi(x)}{\varepsilon} = \big(\partial_\mu - iqA_\mu\big)\psi \ (4)$$

This is the **covariant derivative**. Expanding (3) in $y = x + \varepsilon n$ determines how A_μ changes.

$$\psi \mapsto e^{i\alpha}\psi, \qquad A_\mu \mapsto A_\mu + \frac{1}{q}\,\partial_\mu \alpha, \qquad D_\mu\psi \mapsto e^{i\alpha} D_\mu \psi \tag{5}$$

Example. Putting $A_0 = -\phi$, $(A_1, A_2, A_3) = \mathbf{A}$ (the notation of 2-04), $\alpha = q\chi$, (5) becomes $\mathbf{A} \mapsto \mathbf{A} + \nabla\chi$, $\phi \mapsto \phi - \partial_t \chi$, $\psi \mapsto e^{iq\chi}\psi$. These are the gauge-transformation formula of 2-03 and the phase transformation of the wave function (G'). From the demand we have arrived at the electromagnetic potentials and their gauge transformations. Multiplying comparators along a path gives the phase factor of 2-03 § 05, in geometric language parallel transport (1-07 § 03).

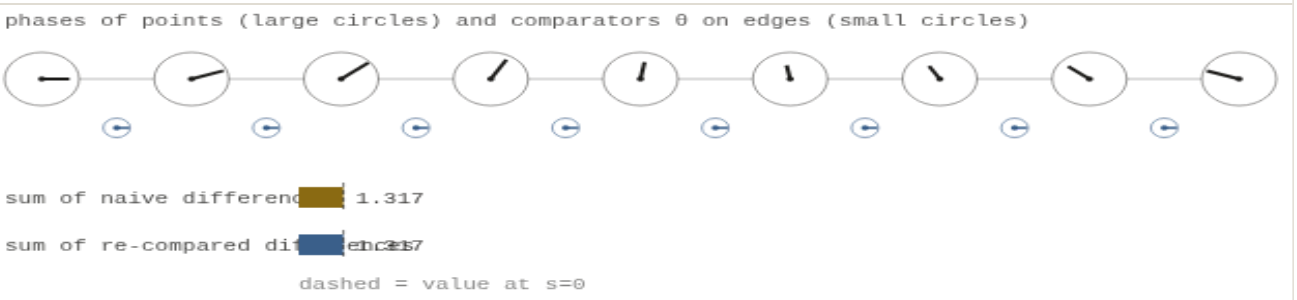
Seen on a chain

If space is cut into a sequence of points $j = 0, 1, 2, \ldots$, the comparator becomes a $U_j = U(j{+}1, j) = e^{i\theta j}$ on each edge (2-03 § 08, the link variables of 5-01), and instead of the derivative we use the **re-compared difference**. (3) becomes $\theta_j \mapsto \theta_j + \alpha_{j+1} - \alpha_j$, and

$$e^{-i\theta'_j}\psi'_{j+1} - \psi'_j = e^{i\alpha_j}\left(e^{-i\theta_j}\psi_{j+1} - \psi_j\right)$$

The sum of squared absolute values does not change, while the sum of naive differences $\psi_{j+1} - \psi_j$ does.

Figure 1 Changing s turns the phase at each point (large circles). With "Connection: on" the comparators on the edges (small circles) also turn by (3), and the sum of re-compared differences does not change. With "Connection: off" it changes



Computed The numbers in the figure are computed on the spot. Point by point with the connection on, moving s from 0 to 1 in steps of 0.25, the sum of naive differences is 1.317, 8.070, 19.536, 22.541, 17.700, while the sum of re-compared differences stays 1.317 (with the connection off both are the former; turning everything by the same amount, both are 1.317).

Computed **Table 1** A ring of 12 points (random field and comparators) before and after transformation. W is the product of the comparators once round (SU(2) is in section 06), and W' is the value after the transformation. "Max. entry of $W' - W$" is the maximum absolute value of the entries of the difference matrix.

Group	Turning	Sum of naive differences before → after	Sum of re-compared differences before → after	Max. entry of $W' - W$
U(1)	point by point	65.474 → 50.856	83.239 → 83.239	0
SU(2)	same everywhere	82.514 → 82.514	75.704 → 75.704	1.39
SU(2)	point by point	82.514 → 85.310	75.704 → 75.704	0.239

The change in the sum of re-compared differences is within 3×10^{-14} in every case (rounding error). Turning U(1) by the same amount everywhere, no column changes.

04

Making A a dynamical field — the field strength, and the term that cannot be written

To make A_μ a field obeying its own equation, we add to the action a term containing derivatives of A that does not change under (5). The material is the comparator of a closed path, in which, by (3), all the α cancel. Going from x in the direction μ, then in the direction ν, once round a square of side ε, and back to x,

$$U_{\square} = \exp\bigl(iq\varepsilon^2 F_{\mu\nu}\bigr) + O(\varepsilon^3), \qquad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \quad (6)$$

The same quantity also appears in $[D_\mu, D_\nu]\psi = -iqF_{\mu\nu}\psi$. The comparator carries from start to end, like the parallel transport of 1-07, and rewriting in the geometric notation $A_{\text{geom}} = -iqA$, $F_{\text{geom}} = -iqF$, (6) becomes $\exp(-\varepsilon^2 F_{\text{geom}})$, agreeing with (10) of 1-07 and $U_p \approx \exp(-a^2 F_{\mu\nu})$ down to the sign (in 5-01, which reads the lattice links in the opposite direction, the sign of the exponent is reversed). **Example.** The components are $F_{0i} = -E_i$, $F_{ij} = \varepsilon_{ijk}B_k$ (2-04), and F is the electric and magnetic fields themselves. The electric and magnetic fields were the **curvature** of the comparator (the deviation after going round a small loop, 1-07 § 05). Listing the invariant terms with the lowest order of derivatives,

$$\mathcal{L} = |D_t\psi|^2 - |\mathbf{D}\psi|^2 - V(|\psi|^2) + \tfrac{1}{2}\bigl(|\mathbf{E}|^2 - |\mathbf{B}|^2\bigr), \qquad D_t = \partial_t + iq\phi, \;\; \mathbf{D} = \nabla - iq\mathbf{A} \quad (7)$$

The last term is $-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}$ ((8) of 2-04). The way of coupling that merely replaces the ∂ of (1) by D is called **minimal coupling**.

The term that cannot be written. The term $-\frac{1}{2}m^2 A_\mu A^\mu = \frac{1}{2}m^2(\phi^2 - |\mathbf{A}|^2)$ that would give A a mass changes under (5) by terms in $-\frac{m^2}{q}\bigl(\phi\,\partial_t\alpha + \mathbf{A}\cdot\nabla\alpha\bigr)$ and the second order of α. A mass term cannot be written for a field that comes from the gauge principle (as in 2-03 § 04) Known . This affects the distance a force reaches in section 05, and becomes a difficulty in section 06.

05

A force appears — coupling to the conserved charge

Expanding (7) in q, the quantities of section 01 appear as they are.

$$\mathcal{L} = \mathcal{L}_0 - \phi\,\rho + \mathbf{A}\cdot\mathbf{J} + q^2\bigl(\phi^2 - |\mathbf{A}|^2\bigr)|\psi|^2 + \tfrac{1}{2}\bigl(|\mathbf{E}|^2 - |\mathbf{B}|^2\bigr) \quad (8)$$

The gauge field couples to the conserved quantity (2) of the global symmetry, with the same q that appeared in the transformation. This is exactly the form of (L) of 2-03.

Varying (8) with respect to ϕ gives Gauss's law.

$$\nabla\cdot\mathbf{E} = \rho - 2q^2\phi\,|\psi|^2 = iq\bigl(\psi^* D_t\psi - \psi\,(D_t\psi)^*\bigr)$$

The right side is not the ρ of (2) itself but that with its ∂_t replaced by D_t (terms containing ϕ are added). For a point charge Q at rest at the origin, we put this right side equal to $Q\,\delta^3(\mathbf{x})$. In the static case $\mathbf{E} = -\nabla\phi$, and from $-\nabla^2\phi = Q\,\delta^3(\mathbf{x})$, $\phi = Q\,G_0(r)$. The energy of a charge Q' at distance r is

$$\mathcal{U}(r) = QQ'\,G_m(r), \qquad G_0(r) = \frac{1}{4\pi r}, \qquad G_m(r) = \frac{e^{-mr}}{4\pi r} \;\; \text{(with a mass term added provisionally)} \quad (9)$$

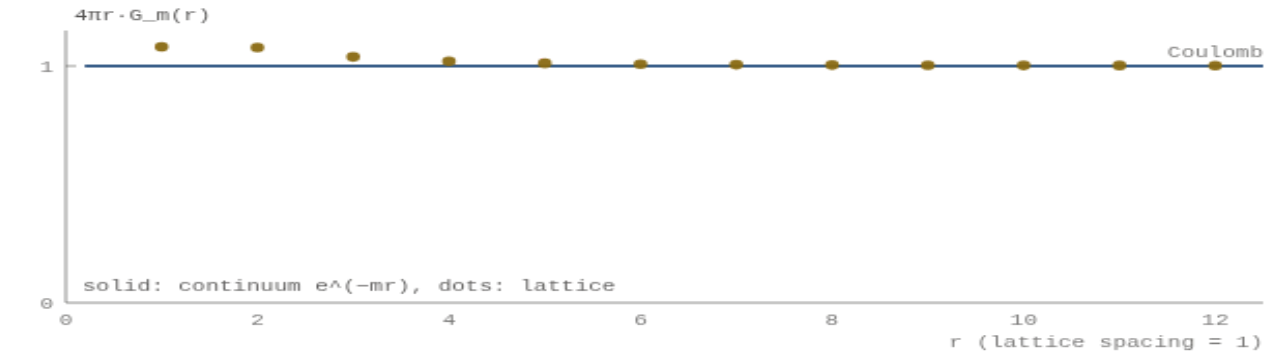
Example. For $m = 0$, charges of the same sign repel, and the force reaches far away like $1/r^2$. Merely demanding a local symmetry has produced a force between distant charges. In the theory (7) (when the minimum of V is at $\psi = 0$), adding the forbidden mass term provisionally would give $(-\nabla^2 + m^2)\phi = \rho$,

and G_m (Yukawa type) does not reach much further than about $1/m$. Without it, the force is of Coulomb type and reaches far.

However, gauge invariance does not guarantee that a force reaches far. Even classically, with a Higgs field the distance reached becomes finite while gauge invariance is kept (3-07) Known . In terms of (8), when the minimum of V is at $|\psi|^2 = v^2 > 0$ and $|\psi|^2$ stays there, the term $q^2(\phi^2 - |\mathbf{A}|^2)|\psi|^2$ has the same form as the mass term $\frac{1}{2}m^2(\phi^2 - |\mathbf{A}|^2)$ of $m^2 = 2q^2v^2$ ((9) of 3-07 puts the minimum of the size of the field at $v/\sqrt{2}$, so the same mass is written $m = qv$).

Figure 2 is $(-\Delta_{\text{lat}} + m^2)G_m = \delta$ found numerically on a cubic lattice (spacing 1) ($\Delta_{\text{lat}} f$ is the sum of the values at the six neighbours minus $6f$).

Figure 2 Changing m changes 4πr·G_m(r) (1 for Coulomb). The solid line is the continuous e^(−mr), the dots the cubic lattice. m > 0 is the case with a mass term added by hand (not gauge invariant)



m = 0.00, rough reach 1/m = ∞, r = 4: lattice 1.020, continuum 1.000

Computed The dots are computations on this machine, the solid line is computed on the spot. Static values: for $r = 1,\ 4,\ 12$ the lattice $4\pi r\, G_0$ is 1.082, 1.020, 1.002, approaching Coulomb further than a few lattice spacings away. $4\pi r\, G_{0.5}$ is 0.670, 0.147, 0.0027 (the continuous $e^{-0.5r}$ is 0.607, 0.135, 0.0025), and the rate of decay along the lattice axis is $\mu = 0.4949$, from $\cosh\mu = 1 + m^2/2$.

06

The same demand for SU(2)

The proton and the neutron have almost equal masses and feel the nuclear force in almost the same way. The isospin view gathers the two into a two-component field $\psi = (\psi_1, \psi_2)$, and the theory of the form (1) does not change under the mixing rotations $\psi \mapsto \Omega\psi$ ($\Omega \in SU(2)$, 1-04; to distinguish it from the V of (1), the matrix is written Ω, as in 3-02). Yang–Mills (1954) imposed the demand of section 02 on this symmetry.

Making the comparator an SU(2) matrix with $U(y,x) \mapsto \Omega(y)\,U(y,x)\,\Omega(x)^{-1}$, and for two nearby points $U(x + \varepsilon n, x) = \exp(ig\varepsilon\, n^\mu A_\mu) + O(\varepsilon^2)$, $A_\mu = A_\mu^a T^a$ ($T^a = \sigma^a/2$; relative to the geometric notation of 1-07, $A_{\text{geom}} = -igA = -i\hat{A}$; $\hat{A} = gA$ is the "coupling absorbed" notation of 3-02 § 02),

$$D_\mu = \partial_\mu - igA_\mu, \qquad A_\mu \mapsto \Omega A_\mu \Omega^{-1} - \frac{i}{g}(\partial_\mu\Omega)\Omega^{-1}, \qquad D_\mu\psi \mapsto \Omega\, D_\mu\psi \quad (10)$$

Example. For $\Omega = e^{i\alpha}$ (1×1) we return to (5) (g playing the role of q), and in geometric notation this is the $A \mapsto \Omega A\Omega^{-1} - (d\Omega)\Omega^{-1}$ of 1-07 § 07 (1-07 writes the transformation matrix as g, but here g is the coupling constant). There are three differences from U(1).

1. **The number of fields.** As many as the dimension of the Lie algebra are needed: 1 for U(1) (the photon), 3 for SU(2) (A^1, A^2, A^3), 8 for SU(3) (gluons), $N^2 - 1$ for SU(N).

2. **The gauge field itself is turned.** The $\Omega A \Omega^{-1}$ of (10) turns A even when Ω is constant (the adjoint representation, 1-04 § 06). Unlike the photon, it carries itself the "charge" it couples to. This is why in Table 1 the W of SU(2) changes as $\Omega_0 W \Omega_0^{-1}$ even under a global rotation, and only $\mathrm{tr}\, W$ remains; self-interaction comes from here (3-02).
3. **There is one coupling constant.** For U(1) a different q can be put freely for each field. For a simple non-abelian group (1-03 § 08) there is only one coupling constant g, and the strength of the coupling to each matter field is determined by g and the generators T^a of that field's representation ($\sigma^a/2$ for a two-component field, $(T^a)_{bc} = -i\varepsilon_{abc}$ for a three-component field in the adjoint representation) Known . The outline of the reason is that the non-linear term of the field strength $-ig[A_\mu, A_\nu]$ (section 09, 3-02) ties the normalisation of A to g and fixes it, and that if the group is simple the invariant form of the kinetic term is unique up to a constant multiple (2-04 § 07, 1-03).

The problem of mass. For SU(2) too, no mass term can be written for the gauge field. In Yang's recollection, at a seminar in Princeton in late February 1954, Pauli asked "what is the mass of this field?" (the story is in the detour of 3-02). The nuclear force is short-range, which does not fit the Coulomb-type reach of section 05. § 1 of the Jaffe–Witten problem statement also writes that the absence of mass in the classical waves was a serious obstacle Known , and says it was overcome for the weak force by adding a Higgs field (3-07), and for the strong force by a property of the quantum theory itself (asymptotic freedom, 4-06) Physics . Showing as mathematics the mass gap expected in the latter (4-11) is the task of the problem statement.

07

What the principle decides, and what it does not

The gauge principle is a strong constraint, but it does not determine the theory uniquely (**Table 2**).

What it decides	What it does not decide
That a connection A is needed, and its transformation laws (5)(10)	Which group (choosing U(1), SU(2), SU(3) is up to experiment)
That derivatives can enter only through D, and that the current coupled to is conserved	The ratios of charges for U(1) (if the group is a circle, the n of the representation $e^{i\alpha} \mapsto e^{in\alpha}$ is an integer and charges are in integer ratios; whether it is a circle or the line $\mathbb{R}$ lies outside the principle)
That no mass term can be written for the gauge field	That the kinetic term is F^2 (F^4 is also invariant; the reason only low-dimension terms are kept is in 4-04)
That configurations carried into each other by gauge transformations represent the same physics	Choosing minimal coupling (merely replacing ∂ by D). The principle does not forbid non-minimal ways of coupling ($\lvert\psi\rvert^2 F_{\mu\nu}F^{\mu\nu}$ is also invariant)

An example for the second row of the table. If the group is a circle, the charge of every field is an integer multiple of a single unit. If the group is the line $\mathbb{R}$, the ratios of charges can be any real numbers, and the principle alone gives no reason for charges to line up as integer multiples of one unit. An example for the third row. A term like F^4 is of higher order in the fields than F^2, and in four dimensions its coefficient has the dimension of length to the fourth power, so it is thought to have less effect the longer the distance (2-04 § 08, 4-04 § 06) Physics .

A local "symmetry" differs from an ordinary symmetry in producing a constraint (Gauss's law) and a redundancy of description (2-01 § 08). How to divide out the redundancy is the subject of 3-04 and 4-08.

There is one more difference from an ordinary symmetry. A symmetry is **spontaneously broken** when the action (in statistical mechanics, the energy) is symmetric but the equilibrium state (the expectation values) is not. An example is the spontaneous magnetisation of 2-08 § 04. Elitzur (1975) argued that in lattice gauge theory without gauge fixing a local symmetry is not spontaneously broken (the expectation values of

quantities that are not gauge invariant are 0), and showed it concretely in simple systems on abelian lattices Known .

08

Detour: "gauge" was originally a measuring rod

A section you may skip.

The words "gauge transformation" and "gauge invariance" first appear in physics in Weyl's 1918 paper attempting to derive gravity and electromagnetism from a single geometry, where they meant changing the standard of length (calibration; the German Eich-, the verb eichen) (O'Raifeartaigh–Straumann § I) Known . Weyl voiced his dissatisfaction with Riemannian geometry, in which the magnitudes of vectors at two distant points can be compared directly (English translation, ibid. § II).

> A true infinitesimal geometry should, however, recognize only a principle for transferring the magnitude of a vector to an infinitesimally close point ...

(A truly infinitesimal geometry should recognise only a principle for transferring the magnitude of a vector to an infinitely near point.) This is the same idea as the comparator of section 03, except that what is carried is length, not phase. Weyl's theory has a real-valued field w_μ expressing at what rate the standard of length changes from point to point, and carrying a vector along a path γ from a point p to a point q changes its length ℓ as

$$\ell(q) = \exp\Big(-\int_\gamma w_\mu \, dx^\mu\Big)\, \ell(p)$$

((2.2) of O'Raifeartaigh–Straumann; the sign depends on conventions). Going once round a closed path and back, the length is multiplied by a real factor $\exp\big(-\oint w_\mu \, dx^\mu\big)$. As with (6), in a region without holes this factor is 1 for every loop when $\partial_\mu w_\nu - \partial_\nu w_\mu = 0$, and Weyl regarded w_μ as the electromagnetic potential (ibid. § II).

Einstein objected at once: the frequency of an atomic clock would change with the path it had followed, contradicting the stability of atomic spectra Known .

When quantum mechanics appeared, London (1927) realised that making the exponent imaginary turns it into a change of the phase of the wave function (that invariance had been found by Fock (1926)). In his 1928 book and 1929 paper Weyl raised this as "the principle of gauge invariance", and wrote that the electromagnetic field necessarily accompanies the wave field of matter (Jackson–Okun § III.B) Known . The name stayed, derived as it was from the calibration of length. As a metaphor it is "a theory in which the measuring rod can be chosen freely at each point", but precisely, what can be chosen at each point is not a scale of length but a complex phase (in general, a group element). The first non-abelian model was Klein (1938), but it was forgotten, and today's line is held to start from Yang–Mills (1954) (ibid. § V).

09

Where this book is going

In 3-02 the square of (6) is turned with SU(N). Because comparators depend on the order of multiplication, a commutator is added to the field strength, and self-interaction appears.

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - ig[A_\mu, A_\nu]$$

Imposing Gauss's law in the quantum theory is 4-01. On the lattice of 5-01 the comparators U_j are the variables, and the action is built from the products around the squares of (6) (plaquettes). That g becomes dimensionless in four dimensions is in 2-04 § 08 and the four-dimensions article.

To sum up: the demand for a local symmetry calls in a gauge field, that field couples to the conserved charge and carries a force, and classically its mass term cannot be written by hand (adding a Higgs field

changes matters; 3-07). The problem statement asks to show that, in the four-dimensional non-abelian case, a mass gap nevertheless arises in the quantum theory (0-03).

What this chapter can say

Item	What is said	What is not said
Sections 01–05: the covariant derivative (4), the transformation law (5), the field strength (6), that the mass term is not invariant, coupling to the conserved current, the two forms of (9)	Standard derivation for classical fields Known . The numbers in Table 1 and Figures 1 and 2 Computed	The lattice values are for spacing 1; the continuum limit is not taken
The distance a force reaches	In the theory (7) (minimum of V at $\psi = 0$), that without a mass term the force is of Coulomb type and reaches far Known	That the force reaches far does not follow from gauge invariance alone. The content of the Higgs mechanism is in 3-07
Section 06: the SU(2) transformation law (10), the number of fields, being turned in the adjoint, that there is one coupling constant	(10), the number of fields, the turning in the adjoint Known . Only an outline of why there is one coupling constant (2-04 § 07, 1-03)	The formula for self-interaction (3-02)
The problem of mass	That no mass term can be written by hand for the classical gauge field Known . How it was overcome for the strong force is Physics	The mass gap of the quantum theory (unsolved; this chapter shows nothing)
Section 07: the standing of the principle, Elitzur's result	That the principle is a guide for choosing the form of a theory. Elitzur's result at the level of the abstract (argued, and shown in simple systems on abelian lattices)	Deriving the principle as a theorem. The content of Elitzur's argument for general groups (the text has not been checked)
The history in section 08	What is written in the two reviews Known	The texts of the original papers (bibliographic data only)

Sources and reproduction

Item	Kind	Source / tools
Table 1, Figure 1	Computed	python3, NumPy (`covariant_diff.py`). The figure is computed on the spot
The dots of Figure 2, the values in section 05	Computed	python3, SciPy (`lattice_green.py`: $G_m(r) = \int_0^\infty e^{-(m^2+6)t} I_r(2t) I_0(2t)^2 dt$, modified Bessel functions; reproduces the identity $G_0(0) - G_0(1) = 1/6$ to 10 digits). Cross-check by FFT, `check_fft.py`
Motivation, Pauli's question	Known	C. N. Yang, R. L. Mills, Phys. Rev. 96 (1954) 191 (abstract checked). C. N. Yang, Selected Papers 1945–1980 with Commentary (1983) p. 525 (bibliographic data only; checked through the quotation in O'Raifeartaigh–Straumann § VII)
The obstacle of massless classical waves, and overcoming it by the Higgs field and asymptotic freedom	Known	A. Jaffe, E. Witten, Quantum Yang–Mills Theory (the Clay problem statement) § 1 (text checked)
Local symmetries are not spontaneously broken (Elitzur)	Known	S. Elitzur, Phys. Rev. D 12 (1975) 3978 (abstract checked)

Item	Kind	Source / tools
The history in section 08	Known	J. D. Jackson, L. B. Okun, Rev. Mod. Phys. 73 (2001) 663, arXiv:hep-ph/0012061 (§ III.B and § V checked). L. O'Raifeartaigh, N. Straumann, Rev. Mod. Phys. 72 (2000) 1, arXiv:hep-ph/9810524 (§ I, § II, § V and § VII of the arXiv version checked). Original papers bibliographic data only (from the reference lists of the two reviews): H. Weyl, Sitzungsber. Akad. Wiss. Berlin (1918) 465 / V. Fock, Z. Phys. 39 (1926) 226 / F. London, Z. Phys. 42 (1927) 375 / H. Weyl, Gruppentheorie und Quantenmechanik (Hirzel, Leipzig, 1928) / H. Weyl, Z. Phys. 56 (1929) 330 / O. Klein, in New Theories in Physics (conference in Warsaw, 1938)

Next chapter: 3-02 Non-abelian gauge fields.

Non-abelian gauge fields — when the order of multiplication matters, the field becomes its own source

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — Allowing the $\mathrm{SU}(N)$ transformation to be chosen point by point determines the covariant derivative and the transformation law, and a commutator term is added to the field strength F. Because of that term the gauge field carries its own charge, and cubic and quartic self-interactions and non-linear equations arise. For fields uniform in space, the commutator alone makes the x^2y^2 valley.

Prerequisite chapters — 3-01 (the gauge principle; especially the SU(2) case of § 06). We use the structure constants of 1-03, the adjoint representation of 1-04, and 2-04 (raising and lowering indices, and the signature of the metric). The same content written in the language of geometry is 1-07.

Said first — This is a classical chapter, treating neither quantum theory nor the mass gap. The action and the equations of motion are left to 3-03, and the energy density is used first by analogy with electromagnetism. Pauli's letters and Yang's recollection in § 01 are quoted from reviews.

Order of this chapter

1. Turning at each point — from the SU(2) of 3-01 to SU(N); the abstract of Yang and Mills. Detour: the spring of 1954
2. Conventions — generators, and translating between three notations
3. The covariant derivative — the two terms of the transformation law; seen on a chain (Figure 1)
4. The field strength — the commutator of covariant derivatives; pure gauge
5. Colour electric and colour magnetic fields — recounting the measurable quantities

6. Self-interaction — quadratic, cubic and quartic terms
7. Fields uniform in space — the x^2y^2 valley (Figure 2); the classical theory has no scale
8. The problem of mass — why it cannot be added by hand

01

Turning at each point — the question of Yang and Mills

In 3-01 § 06, we imposed the demand that turning point by point be allowed on the symmetry (isospin) that mixes, by SU(2) matrices, the two-component field ψ combining the proton and the neutron, and obtained the covariant derivative and the transformation law (10). This chapter extends that to $\mathrm{SU}(N)$ and follows in turn what arises from the commutator.

The abstract of the paper of Yang and Mills (Phys. Rev. 1954) points out that the usual invariance, turning by the same matrix at every point, is not consistent with the idea of localised fields, and continues: “The possibility is explored of having invariance under local isotopic spin rotations.” Rephrasing the first sentence of the abstract in this book's reading: that the choice of which direction to call "proton" at one point should fix the naming at distant points as well does not fit the laws of fields, which are written in terms of relations between nearby points. So we allow the transformation matrix $\Omega = \Omega(x)$ to be chosen point by point.

According to the abstract, the consequence is a **b** field that plays the same role for isospin as the electromagnetic field, obeys non-linear differential equations, and whose quanta are particles of spin 1, isospin 1 and charge $\pm e$ or 0 Known . § 03– § 06 turn this summary into formulas. The group carrying the strong force later turned out to be SU(3) (Jaffe–Witten's problem description: “the gauge group is G = SU(3)”). However, what was gauged was not isospin but SU(3) of colour, a different degree of freedom (isospin remains an approximate symmetry, not turned point by point). Mathematically, one just replaces the group (3-07, the detour of 1-04).

Detour: the spring of 1954

A subsection you may skip.

According to an essay by Straumann, Pauli had obtained the same field strength from the framework of extra dimensions (17-06), in notes attached to a letter to Pais in July 1953 and in two seminars in Zürich in November. But in a letter to Pais in December he wrote “one will always obtain vector mesons with rest mass zero”, and did not publish. Yang recalls that at a

seminar in Princeton in late February 1954, just after Yang had written $(\partial_\mu - i\epsilon B_\mu)\psi$ on the blackboard, Pauli asked "What is the mass of this field B_μ?" (quoted from the commentary in Yang's selected papers by Straumann's essay and the review of O'Raifeartaigh–Straumann). The same review mentions that the theory was found independently by other authors, and places Shaw (1954–55) and Utiyama (1954–55, 1956) in its family-tree figure Known . The question of mass is answered in § 08.

02

Conventions — generators, and three notations

In Part 3, $\hbar = c = 1$, and the signature of the metric is $(-,+,+,+)$ (2-04). $\mu, \nu = 0, \dots, 3$ are spacetime indices, $i, j, k = 1, 2, 3$ spatial, and a, b, c indices in the direction of the group (colour); indices appearing twice are summed. The group is $G = \mathrm{SU}(N)$, the generators are Hermitian T^a $(a = 1, \dots, N^2 - 1)$, and the structure constants and normalisation are as follows (1-04).

$$[T^a, T^b] = i f^{abc} T^c, \qquad \mathrm{tr}(T^a T^b) = \tfrac{1}{2}\delta^{ab}$$

Example. For SU(2), $T^a = \sigma^a/2$ (half the Pauli matrices) with $f^{abc} = \varepsilon^{abc}$; for SU(3), $T^a = \lambda^a/2$ (half the Gell-Mann matrices). The gauge transformation matrix is written $\Omega(x) \in G$ (the same letter as in 3-01), and the letter g is kept for the coupling constant. A matter field $\psi(x) \in \mathbb{C}^N$ is taken to change as $\psi \mapsto \Omega(x)\psi$ (the fundamental representation).

This chapter follows the physicists' style, writing the Hermitian $A_\mu = A_\mu^a T^a$ and the coupling g separately. The ϵ and B_μ on Yang's blackboard in the detour of § 01 correspond to this g and A_μ (the conventions for coefficients depend on the paper). The other notations in the book can be translated by the following lines.

Notation	Covariant derivative	Field strength	Chapter
This chapter	$D_\mu = \partial_\mu - igA_\mu$	$F_{\mu\nu}$ (formula (6))	3-02, Part 4
Geometric (anti-Hermitian)	$D = d + \mathcal{A}$, $\mathcal{A} = -igA$	$\mathcal{F} = d\mathcal{A} + \mathcal{A}\wedge\mathcal{A} = -igF$	1-07, 3-04
Coupling absorbed	$D_\mu = \partial_\mu - i\hat{A}_\mu$, $\hat{A} = gA$	$\hat{F} = gF$	2-04

03

The covariant derivative — the two terms of the transformation law

We extend (10) of 3-01 to $\mathrm{SU}(N)$ and follow the derivation again. The trouble is the derivative. If Ω depends on the point, the derivative of the product produces an extra

term.

$$\partial_\mu(\Omega\psi) = \Omega\,\partial_\mu\psi + (\partial_\mu\Omega)\,\psi \quad (1)$$

The second term appears because values at neighbouring points with different frames were subtracted as they stand (1-07 § 01). So, as in 3-01, we demand that the derivative with a correcting field A_μ added change in the same way as ψ.

$$D_\mu = \partial_\mu - igA_\mu, \qquad D'_\mu(\Omega\psi) = \Omega\,D_\mu\psi \quad (2)$$

The left side is $\Omega\partial_\mu\psi + (\partial_\mu\Omega)\psi - igA'_\mu\Omega\psi$, the right side $\Omega\partial_\mu\psi - ig\,\Omega A_\mu\psi$. Equality for every ψ determines A'_μ.

$$A'_\mu = \Omega A_\mu \Omega^{-1} - \frac{i}{g}\,(\partial_\mu\Omega)\,\Omega^{-1} \quad (3)$$

The first term is the rotation by the adjoint representation (1-04 § 06), and the second is an inhomogeneous term coming from the slope of Ω. Differentiating $\Omega\Omega^\dagger = 1$, $(\partial\Omega)\Omega^{-1}$ is anti-Hermitian, and from $\det\Omega = 1$ its trace is 0, so A' also stays in the same Lie algebra. Example. For $\Omega = e^{i\alpha(x)}$ (U(1)) the first term stays A_μ, and $A'_\mu = A_\mu + \partial_\mu\alpha/g$, putting $\chi = \alpha/g$, is exactly the gauge-transformation formula of 2-03 (with the convention $A_\mu = (-\phi, \mathbf{A})$ of 2-04). What is new in the non-abelian case is the first term, which sandwiches A.

Infinitesimal transformations — the gauge field itself turns

Putting $\Omega = 1 + i\alpha^a T^a$ (α small) and expanding (3) to first order, in components

$$\delta A^a_\mu = \frac{1}{g}\,\partial_\mu\alpha^a + f^{abc}A^b_\mu\,\alpha^c = \frac{1}{g}\,(D_\mu\alpha)^a \quad (4)$$

The $(D_\mu\phi)^a = \partial_\mu\phi^a + gf^{abc}A^b_\mu\phi^c$ at the right end is the covariant derivative acting on fields in the adjoint representation. Example. For SU(2), taking α constant with $\alpha^3 = \theta$ (others 0), $\delta A^1 = \theta A^2$, $\delta A^2 = -\theta A^1$, $\delta A^3 = 0$. Even under a rotation independent of position the components of the gauge field turn into each other, so **the gauge field itself carries non-abelian charge** (this is "the gauge field itself is turned" of 3-01 § 06 seen in components). This is in contrast to the photon, which carries no charge, and it is what "isospin 1" in the abstract means.

Seen on a chain

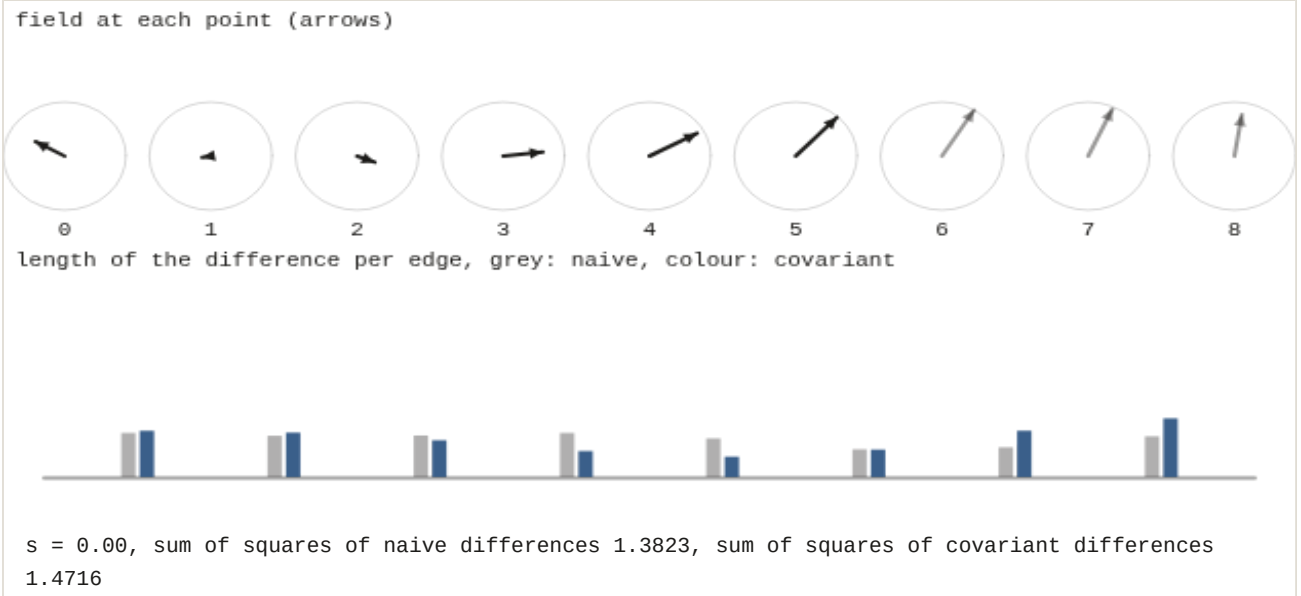
Arranging points in a row makes the meaning of (2) visible. At each point n put a field in the adjoint representation (a real three-component vector) ϕ_n, and on each edge put

the rotation U_n that carries the value at the point $n+1$ into the frame of the point n. This is the non-abelian version of the U(1) elements on the edges of 2-03, and a component of the lattice of 5-01. Written with the comparator of 3-01 it is $U_n = U(n, n{+}1)$, the opposite direction from the $U_j = U(j{+}1, j)$ of the chain of 3-01. The Ω_n and U_n here denote, by the same letters, the rotations of $\mathbf{3} \times \mathbf{3}$ given by SU(2) elements in the adjoint representation (1-04 § 06).

$$\phi'_n = \Omega_n \phi_n, \qquad U'_n = \Omega_n U_n \Omega^{-1}_{n+1}, \qquad U'_n \phi'_{n+1} - \phi'_n = \Omega_n \big(U_n \phi_{n+1} - \phi_n \big) \tag{5}$$

The "covariant difference" at the right end is multiplied by only one rotation, so its length does not change. The naive difference $\phi'_{n+1} - \phi'_n$ has different rotations at the two points, so its length changes. With the spacing of the points h, putting $U_n \approx 1 - igh A$ and letting $h \to 0$, the second formula of (5) becomes (3), and the covariant difference becomes $h D_\mu \phi$. Here A acts as $A = A^a T^a_{\text{adj}}$ using the matrices $(T^a_{\text{adj}})_{bc} = -i f^{abc}$ of the adjoint representation, and the components of $D_\mu \phi$ have the same form as the right end of (4).

Figure 1 The field at nine points (arrows above) and the length of the difference on each edge (bars below). Changing the strength of the gauge transformation with the slider s, the grey bars (naive differences) change, and the coloured bars (covariant differences) do not.



Computed The numbers in this figure were computed on this machine (the same formulas and the same configuration as Table 1, computed in the browser). If JavaScript is disabled, Table 1 is the content of the figure.

Table 1 The strength of the gauge transformation s, and the sums of squares of the differences on the eight edges

s	Naive difference	Covariant difference
0.00	1.3823	1.4716
0.25	3.0524	1.4716

s	Naive difference	Covariant difference
0.50	7.6091	1.4716
0.75	13.5267	1.4716
1.00	18.6570	1.4716

The change in the covariant difference on each edge was at most 3×10^{-16} (rounding error) Computed . That the two differ at $s = 0$ is because U_n is taken to be a small rotation.

04

The field strength — the commutator of covariant derivatives

In electromagnetism $\partial_\mu A_\nu - \partial_\nu A_\mu$ was a quantity unchanged by gauge transformations. In the non-abelian case, because of the first term of (3), this combination is not even covariant. The correct combination comes from the difference of covariant derivatives multiplied in the two orders.

$$[D_\mu, D_\nu]\,\psi = -ig\,F_{\mu\nu}\,\psi, \qquad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - ig\,[A_\mu, A_\nu] \tag{6}$$

Expanding the left side, the terms in $\partial_\mu\partial_\nu\psi$ and $A\,\partial\psi$ are symmetric in $\mu\nu$ and cancel, so no derivatives of ψ remain. $F_{\mu\nu}$ is a matrix (an element of the Lie algebra) at each point, and in components it is as follows.

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g\,f^{abc} A^b_\mu A^c_\nu \tag{7}$$

For SU(2), regarding the three components $(A^1_\mu, A^2_\mu, A^3_\mu)$ as a vector in the internal space, the last term is the cross product $g\,A_\mu \times A_\nu$. (2) means $D'_\mu = \Omega D_\mu \Omega^{-1}$ as an operator, so the commutator is sandwiched in the same way.

$$F'_{\mu\nu} = \Omega\,F_{\mu\nu}\,\Omega^{-1} \tag{8}$$

F is not invariant; it turns at each point by the adjoint representation (it is covariant). This is a large difference from electromagnetism, where the electric and magnetic fields were themselves measurable quantities.

Example 1 (pure gauge). The $A_\mu = -\frac{i}{g}(\partial_\mu\Omega)\Omega^{-1}$ obtained by transforming $A = 0$ by (3) has $F = 0$, by (8). Conversely, if $F = 0$ on a simply connected region such as $\mathbb{R}^4$, then A can be written in this form Known . The notes Pauli attached to his letter to Pais of July 1953, as presented by Straumann, state this as “The vanishing of the field strength is necessary and sufficient for the $A^a_\mu(x)$ in the whole space to be

transformable to zero." It fails on regions with holes, and the Aharonov–Bohm effect of 2-03 is an example (the global story is in 3-04).

Example 2 (commutator only). For constant fields $A_x = a\,T^1$, $A_y = b\,T^2$, $F_{xy} = -ig\,ab\,[T^1, T^2] = g\,ab\,T^3 \neq 0$. In electromagnetism a constant potential makes no field. A figure showing this as the deviation after going round a small square is Figure 2 of 1-07.

Computed Giving smooth A and Ω for SU(2) and SU(3) and comparing by finite differences, the difference between the two sides of (8) was 5×10^{-9} (the error of step 10^{-4}); dropping the commutator gave 0.78, 1.74; and the F of a pure gauge was at most 10^{-9}.

05

Colour electric and colour magnetic fields — recounting the measurable quantities

Following electromagnetism, we define electric and magnetic fields for each direction of the group. With the same assignment as $F_{0i} = -E_i$, $F_{ij} = \varepsilon_{ijk} B_k$ of 2-04, the three spatial components are written in bold ($\mathbf{A}^a = (A^a_1, A^a_2, A^a_3)$).

$$E^a_i = F^a_{i0}, \qquad B^a_i = \tfrac{1}{2}\varepsilon_{ijk} F^a_{jk} = (\nabla \times \mathbf{A}^a)_i + \tfrac{g}{2}\,\varepsilon_{ijk} f^{abc} A^b_j A^c_k \tag{9}$$

Taking $A_0 = 0$ (temporal gauge), $E^a_i = -\partial_t A^a_i$. The second term of the magnetic field is the commutator part, and a magnetic field arises even if the potential is spatially uniform.

By (8), $\mathbf{E}^a$ and $\mathbf{B}^a$ turn in the colour index at each point. So "component 3 of the colour electric field" changes value with the gauge and is not a measurable quantity. What does not change are squared lengths such as $\sum_a |\mathbf{E}^a|^2 = 2\,\mathrm{tr}(\mathbf{E}\cdot\mathbf{E})$. Following electromagnetism, in this chapter we take the following as the energy density (deriving it from the action is in 3-03).

$$\mathcal{E} = \tfrac{1}{2}\sum_a \left(|\mathbf{E}^a|^2 + |\mathbf{B}^a|^2\right) \tag{10}$$

Example. Typical quantities that do not change with the gauge are $\mathrm{tr}(F_{\mu\nu}F^{\mu\nu})$ and the trace of parallel transport around a closed path (the Wilson loop, 1-07 § 06).

06

Self-interaction — quadratic, cubic and quartic terms

Both the energy (10) and the $F^a_{\mu\nu}F^{a\,\mu\nu}$ put in the action in 3-03 are squares of F. Putting in (7), they split into terms quadratic, cubic and quartic in A.

$$\begin{aligned} F^a_{\mu\nu}F^{a\,\mu\nu} = {} & (\partial_\mu A^a_\nu - \partial_\nu A^a_\mu)(\partial^\mu A^{a\,\nu} - \partial^\nu A^{a\,\mu}) \\ & + 4g\, f^{abc}\,(\partial_\mu A^a_\nu)\, A^{b\,\mu} A^{c\,\nu} \\ & + g^2 f^{abe} f^{cde}\, A^a_\mu A^b_\nu A^{c\,\mu} A^{d\,\nu} \end{aligned} \qquad (11)$$

The first line is Maxwell's $F_{\mu\nu}F^{\mu\nu}$ repeated for the number of colours N^2-1, and on its own it is a collection of Maxwell fields that do not act on each other. The second and third lines contain f^{abc} and vanish for U(1). Put in the action (3-03), they are interactions among gauge fields, and in perturbation theory (4-03) they become vertices where three or four gauge fields meet. In the classical Maxwell equations light does not scatter off light, but the Yang–Mills field scatters off itself.

The consequence is the non-linearity of the equations. Jaffe–Witten write "The Yang–Mills equations are nonlinear—in contrast to the Maxwell equations." Known . Adding two solutions does not in general give a solution. Computed For random A^a_μ, $\partial_\mu A^a_\nu$ the two sides of (11) agreed to 10 digits. The non-zero f^{abc}, counted as ordered, number 6 for SU(2) and 54 for SU(3).

Table 2 Classical differences between electromagnetism and Yang–Mills

	Electromagnetism U(1)	**Yang–Mills SU(N)**
Number of gauge fields	1	N^2-1 (8 for SU(3))
Field strength	Invariant	Covariant (8)
Charge of the gauge field	None	Carries it (4)
Constant potential	Makes no field	Makes a field if the commutator is not 0
Equations	Linear	Non-linear (11)
Classical waves	Mass 0	Mass 0 (§ 08)

07

Fields uniform in space — a valley made by the commutator alone

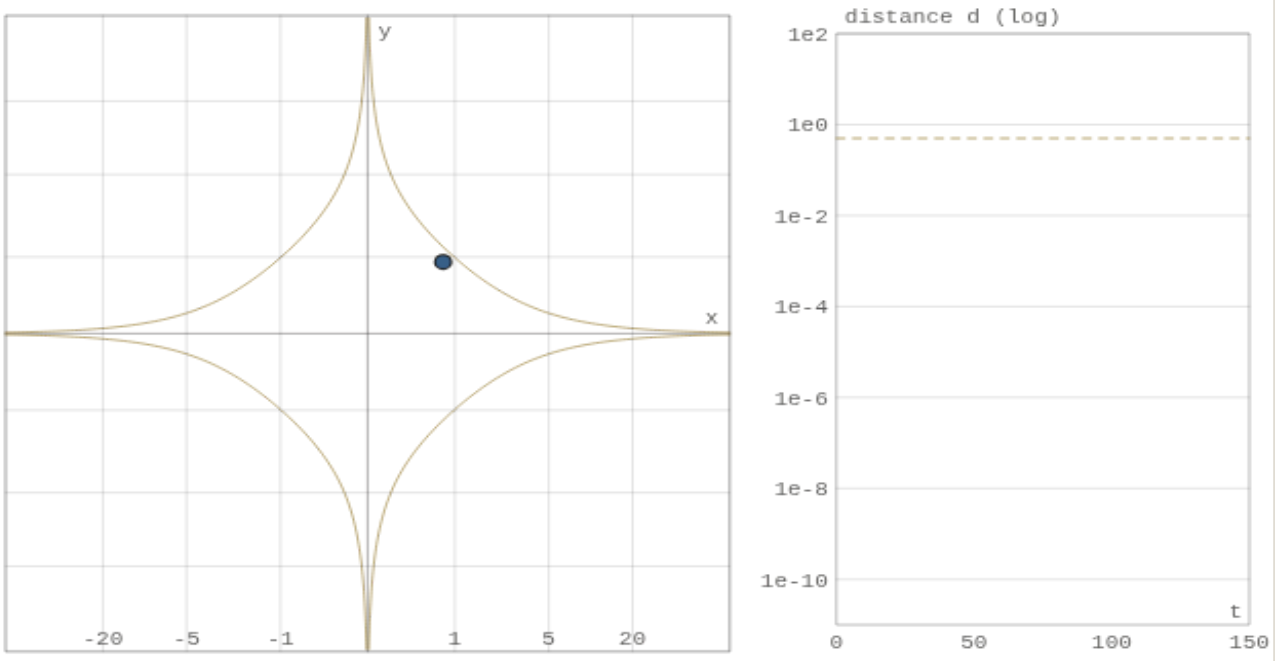
We look at self-interaction in its smallest form. For SU(2) with $A_0 = 0$, and A^a_i functions of time only, the magnetic field of (9) consists of the commutator term alone. Further putting $A^1_1 = x(t)$, $A^2_2 = y(t)$ and the rest 0, the non-zero fields are $E^1_1 = -\dot{x}$, $E^2_2 = -\dot{y}$, $B^3_3 = g\,xy$, and (10) becomes the energy of a particle in the plane.

$$H = \tfrac{1}{2}\left(\dot{x}^2 + \dot{y}^2\right) + \tfrac{1}{2} g^2 x^2 y^2, \qquad \ddot{x} = -g^2 x y^2, \quad \ddot{y} = -g^2 x^2 y \tag{12}$$

For uniform fields, the Yang–Mills equations of motion (3-03) become the equations of a particle with 9 components A_i^a moving in $\frac{1}{2}\sum (B_i^a)^2$, and this assignment is consistent with them Computed (by numerical differentiation, the force on the other 7 components is 0, and the Gauss constraint $\varepsilon^{abc} A_i^b E_i^c = 0$ also holds). The Gauss constraint is the condition left by the equation for A_0 lost in the temporal gauge, the non-abelian version of $\nabla \cdot \mathbf{E} = 0$ in electromagnetism. For uniform fields the derivative terms vanish, giving this form (4-01). For comparison, putting the two components in the same colour direction ($A_1^1 = x$, $A_2^1 = y$), the commutator is 0 and $V = 0$, and the particle moves in a straight line, as in electromagnetism.

The non-abelian $V = \frac{1}{2} g^2 x^2 y^2$ is 0 on the coordinate axes, so the valleys go on for ever, narrowing the further in (the width available at energy E is $|y| \le \sqrt{2E}/(g|x|)$). Figure 2 runs two particles at the same time with velocities differing by δ.

Figure 2 Left: orbits of (12) ($g = 1$, initial values $(x, y, \dot{x}, \dot{y}) = (0.8, 0.9, 0.6, -0.3)$, $E = 0.4842$; the scale compresses distant parts, and the thin curve is the contour $V = E$). Right: the distance d between the two particles (vertical axis logarithmic). The slider k changes the initial offset $\delta = 10^{-k}$, and "Make abelian" switches to $V = 0$.



t = 0.00, d = 0.0e+0, time exceeding 0.5 —, drift in E 0e+0 (non-abelian, δ = 1e-6)

Computed The numbers in this figure were computed on this machine (the same integration rule as Table 3: fourth-order Runge–Kutta, step $\min(0.01,\ 0.02/(g\max(|x|, |y|)))$, distance checked every 0.05). If JavaScript is disabled, Table 3 is the content of the figure.

Table 3 The initial offset δ, and the time at which the distance exceeds 0.5

δ	non-abelian	Abelian ($0.5/\delta$)

δ	non-abelian	Abelian ($0.5/\delta$)
1e-2	7.45	50
1e-4	25.05	5e3
1e-6	37.90	5e5
1e-8	54.55	5e7
1e-10	77.15	5e9

In the non-abelian case, each time δ is made 1/10, the separation time grows by only about 8 (a straight-line fit to the 9 points 10^{-2}–10^{-10} gives slope 8.04 (Table 3 shows 5 of them, one per decade); the increments are uneven because separation slows while the orbit is deep in a valley). This means the distance grows exponentially, and on the right it rises almost as a straight line. In the abelian case the distance is $\delta\,t$. The drift of the energy was at most 5×10^{-9} at $t \le 120$ Computed . This system has been studied as classical Yang–Mills mechanics. Savvidy (1983) argued that dynamics of this kind is a K-system (a dynamical system with strong mixing), but Dahlqvist–Russberg (1990) found a family of stable periodic orbits in this x^2y^2 system, refuting earlier claims that the motion is completely ergodic (according to the abstract). What the computation on this machine shows goes only as far as the separation visible in finite time for the initial values above.

The classical theory has no scale

(12) has no constant that could serve as a unit of length or time (g only changes the scale of time). If $x(t), y(t)$ is a solution, then for any $\lambda > 0$

$$x_\lambda(t) = \lambda\, x(\lambda t), \qquad y_\lambda(t) = \lambda\, y(\lambda t), \qquad E_\lambda = \lambda^4 E \quad \text{(13)}$$

is also a solution (substituting, λ^3 appears on both sides). Example. Making the energy 1/16, the motion of the same shape happens twice as slowly. Bringing the energy towards 0, the motion becomes as slow as we like, and the frequencies have no lower bound. Computed Solving the initial values $\lambda = 0.5,\ 0.25,\ 0.125$ separately with the same rule and dividing the position at time $5/\lambda$ by λ, all agreed with $(-0.401421,\ 0.616689)$ to 6 digits.

The same holds for the whole field: keeping only the quadratic terms of (11) around $A = 0$ gives a collection of Maxwell fields, and the frequencies $\omega = |\mathbf{k}|$ continue down to 0. Jaffe–Witten write that the classical equations describe “massless waves that travel at the speed of light” Known . Quantising, deep in the valley the lowest energy of the transverse oscillation $g|x|/2$ acts as a wall, and the spectrum of the operator obtained by quantising (12) is discrete, with a positive lowest eigenvalue Known (Simon 1983, the

detour of 2-05). This is a theorem about a model with finitely many degrees of freedom keeping only uniform fields.

That quantum Yang–Mills as a field theory has a gap is widely accepted in physics, but has not become a theorem Physics .

08

The problem of mass — why it cannot be added by hand

We answer Pauli's question in the detour of § 01 with formulas. For a vector field of mass m, the usual thing is to add $-\frac{1}{2}m^2 A^a_\mu A^{a\,\mu}$ to the Lagrangian density (for the electromagnetic field, the Proca field). But computing the change by (4),

$$\delta\left(A^a_\mu A^{a\,\mu}\right) = \frac{2}{g}\,A^{a\,\mu}\partial_\mu \alpha^a + 2 f^{abc} A^{a\,\mu} A^b_\mu\, \alpha^c = \frac{2}{g}\,A^{a\,\mu}\partial_\mu \alpha^a \quad (14)$$

The second term vanishes because f^{abc} is antisymmetric in a, b while $A^{a\,\mu} A^b_\mu$ is symmetric. The remaining first term comes from the inhomogeneous term of (3), and is not 0 if α depends on the point. The mass term breaks the demand that turning at each point be allowed Known . Computed For random A and small α, the two sides of (14) agreed to 4 digits or more for both SU(2) and SU(3).

So as long as gauge invariance is kept, the waves of the classical Yang–Mills field have no mass. Jaffe–Witten write “The massless nature of classical Yang–Mills waves was a serious obstacle to applying Yang–Mills theory to the other forces”. That the way past it differed for the weak force (the Higgs field, 3-07) and the strong force (asymptotic freedom, 4-06) we saw in 3-01 § 06.

As the first of the properties a theory of the strong force must have at the quantum level, Jaffe–Witten list the mass gap: “there must be some constant $\Delta > 0$ such that every excitation of the vacuum has energy at least Δ” (0-03, 4-11). The classical field has no scale (§ 07), and the gap is thought to arise only on quantisation Physics . Showing existence and the mass gap as theorems for pure Yang–Mills theory is what the problem statement asks, and this chapter shows nothing about it.

What this chapter can say

Item	What is said	What is not said
(2)–(8), (11)	Standard derivations, and numerical checks for SU(2) and SU(3) Known Computed	The equations of motion and the Bianchi identity (3-03), the geometry of the space of connections (3-04)

Item	What is said	What is not said
(12) the x^2y^2 system of uniform fields	Consistency of the assignment, and the exponential separation visible in finite time for these initial values Computed	A proof of irregularity, the value of the Lyapunov exponent (the exponent of the rate at which nearby orbits separate)
(13) that the classical theory has no scale	Exact scaling and numerical checks Computed	A proof, as a field theory, that quantisation produces a scale
(14) the mass term	That it breaks local gauge invariance Known	The content of the Higgs mechanism (3-07)
The mass gap	—	That quantum Yang–Mills theory has a gap (unsolved; this site shows nothing) Physics
The history in § 01	What is written in the abstract, the reviews and the problem description	The text of Yang–Mills, the original wording of Pauli's letters, the content of Shaw's work

Sources and reproduction

Item	Kind	Source / tools
The quotation and summary of the abstract in § 01	Primary source (abstract checked; text not read)	C. N. Yang, R. L. Mills, “Conservation of Isotopic Spin and Isotopic Gauge Invariance”, Phys. Rev. 96 (1954) 191 (received 28 June 1954)
Pauli's 1953 work, the phrase from his letter, the quotation from his notes, Yang's recollection	Review (text checked)	N. Straumann, “On Pauli's invention of non-abelian Kaluza-Klein Theory in 1953”, arXiv:gr-qc/0012054 (Yang's recollection is cited with C. N. Yang, Selected Papers p. 525 as the source). The collected letters (W. Pauli, Wissenschaftlicher Briefwechsel IV/II, Springer 1999) bibliographic data only
The recollection of the February 1954 seminar, Shaw and Utiyama	Review (text checked)	L. O'Raifeartaigh, N. Straumann, “Early History of Gauge Theories and Kaluza-Klein Theories, with a Glance at Recent Developments”, arXiv:hep-ph/9810524 § VII and Figure 1 (published title “Gauge theory: Historical origins and some modern developments”, Rev. Mod. Phys. 72 (2000) 1–23). The original source, C. N. Yang, Selected Papers 1945–1980 with Commentary (1983) p. 525, bibliographic data only
Bibliographic data of Utiyama (1956)	Primary source (abstract checked)	R. Utiyama, “Invariant Theoretical Interpretation of Interaction”, Phys. Rev. 101 (1956) 1597
The quotations on SU(3), non-linearity, massless waves, the serious obstacle and the mass gap	Primary source (text checked)	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory”, the official problem description of the Clay Mathematics Institute, § 1

Item	Kind	Source / tools
The x^2y^2 system: the claim that it is a K-system	Bibliographic data only (checked with Crossref)	Savvidy, “The Yang-Mills classical mechanics as a Kolmogorov K-system”, Phys. Lett. B 130 (1983) 303–307
The x^2y^2 system: a family of stable periodic orbits	Primary source (abstract checked)	P. Dahlqvist, G. Russberg, “Existence of stable orbits in the x^2y^2 potential”, Phys. Rev. Lett. 65 (1990) 2837–2838
Discrete spectrum of the quantum x^2y^2 system	Known (bibliographic data only)	B. Simon, “Some quantum operators with discrete spectrum but classically continuous spectrum”, Ann. Phys. 146 (1983) 209–220 (reference [43] in § 6.6 of Jaffe–Witten, cited as an example in which bound states arise even for potentials with flat directions)
Derivations of (1)–(14)	Known	Standard material. The translation of notation is the table in § 02
Table 1, Figure 1, the numbers in § 04, § 06 and § 08	Computed on this machine	Python 3.12, NumPy 1.26, SciPy 1.11 (`chain_covariance.py`, `formula_checks.py`). Figure 1 computes the same formulas in plain JavaScript
Consistency in § 07, Table 3, the check of (13), Figure 2	Computed on this machine	Python 3.12 (`ym_mechanics.py`). Figure 2 integrates with the same rule in plain JavaScript. Differences in rounding are amplified by irregular motion, so when t goes well beyond 100 the orbits of the figure and the table can diverge

Next chapter: 3-03 The action and the equations ($\mathrm{tr}\, F^2$, the Bianchi identity, the equations of motion, self-duality). The geometry continues in 3-04.

The action and the equations — from one action come the equations, an identity, the energy and self-duality

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — Varying the action gives the equation $D_\mu F^{\mu\nu} = 0$, and the way F is built gives the Bianchi identity. In four dimensions the trace of the stress is 0, and it follows that there is no static lump of finite energy. The action in imaginary time is at least $8\pi^2|Q|/g^2$, with equality for self-dual fields.

Prerequisite chapters — 3-02 (the covariant derivative, the field strength, the colour electric and magnetic fields). For variations, 2-01; for the Maxwell action, 2-03 and 2-04; for the Hodge star, 1-06 and 1-12.

Said first — This is a classical chapter, treating neither quantum theory nor the mass gap. § 06 is a known fact about classical fields, and a different matter from whether the quantum theory has a gap. That Q is an integer, and the construction of the BPST solution, are left to 3-05.

Order of this chapter

1. Conventions and the action — why $\operatorname{tr} F^2$ is chosen
2. The equations of motion — the field becomes its own source
3. Writing it with the colour electric and magnetic fields — Gauss's law is a constraint
4. The Bianchi identity — the other half of the equations
5. Energy and momentum — the trace vanishes in four dimensions
6. There is no static lump — Detour: no assumption on the rate of decay is needed, the spherically symmetric field (Figure 1), Detour: a solution of infinite energy
7. The action in imaginary time and self-duality — a lower bound on the action
8. A stretched instanton — Q is unchanged and the action grows (Figure 2)

9. What the classical theory says, and what it does not

01

Conventions and the action — why tr F² is chosen

The conventions are the same as in 3-02 § 02: $\hbar = c = 1$, the metric $(-,+,+,+)$, Hermitian T^a ($\operatorname{tr} T^a T^b = \frac{1}{2}\delta^{ab}$), and $D_\mu = \partial_\mu - igA_\mu$. On a field X with values in the Lie algebra, $D_\mu X = \partial_\mu X - ig[A_\mu, X]$ acts (in components, $\partial_\mu X^a + g f^{abc} A^b_\mu X^c$), and the colour electric and magnetic fields are $E^a_i = F^a_{i0}$ and $B^a_i = \frac{1}{2}\varepsilon_{ijk} F^a_{jk}$ (3-02 § 05). The action replaces the F of the Maxwell action ((8) of 2-04) with the non-abelian field strength and takes the trace.

$$S[A] = -\frac{1}{2}\int \operatorname{tr}\left(F_{\mu\nu}F^{\mu\nu}\right) d^4x = -\frac{1}{4}\int F^a_{\mu\nu} F^{a\,\mu\nu}\, d^4x = \frac{1}{2}\int \sum_a \left(|\mathbf{E}^a|^2 - |\mathbf{B}^a|^2\right) d^4x \tag{1}$$

The form is almost fixed by four requirements: gauge invariance (the rotation by $F' = \Omega F \Omega^{-1}$ disappears under the trace), Lorentz invariance, equations that close at second order in time derivatives, and no length scale in the coefficients of the action (only the terms with mass dimension $[F^2] = 4$; counting dimensions in 2-04 § 08). The last condition is needed because adding a higher-order invariant in F such as $(\operatorname{tr} F^2)^2$ still leaves the equations second order. Its coefficient needs the fourth power of a length, which brings in a scale. Imposing all four, and if the group is simple, the quadratic invariants in F are only linear combinations of (1) and the θ term (3-05), which does not change the equations (2-04 § 07) Known . Example. For the uniform field $A^1_1 = x(t)$, $A^2_2 = y(t)$ (3-02 § 07) the action per volume is $L = \frac{1}{2}(\dot{x}^2 + \dot{y}^2) - \frac{1}{2}g^2x^2y^2$, and its Euler–Lagrange equations (2-01) are (12) of 3-02.

Rescaling as $\hat{A} = gA$, (1) becomes $-\frac{1}{2g^2}\int \operatorname{tr} \hat{F}_{\mu\nu}\hat{F}^{\mu\nu}$ (the same form as the $\frac{1}{4g^2}\int \operatorname{Tr} F \wedge *F$ of Jaffe–Witten, up to conventions of normalisation and sign Known). g has no effect on the classical equations; in the quantum theory, as the combination $g^2\hbar$, it becomes the variable of the perturbative expansion (4-03).

02

The equations of motion — the field becomes its own source

Shifting A by δA, (6) of 3-02 gives $\delta F_{\mu\nu} = D_\mu \delta A_\nu - D_\nu \delta A_\mu$. By the antisymmetry of $F^{\mu\nu}$, and because integration by parts moves the commutator term onto the other

factor by $\operatorname{tr}([A,X]Y) = -\operatorname{tr}(X[A,Y])$ (δA is 0 far away),

$$\delta S = -2\int \operatorname{tr}\big(F^{\mu\nu} D_\mu \delta A_\nu\big)\, d^4x = 2\int \operatorname{tr}\big((D_\mu F^{\mu\nu})\, \delta A_\nu\big)\, d^4x$$

follows. The stationarity condition, that this is 0 for every δA, is the Yang–Mills equation Known .

$$D_\mu F^{\mu\nu} = 0, \qquad \partial_\mu F^{a\,\mu\nu} = -g\, f^{abc} A^b_\mu F^{c\,\mu\nu} \quad (2)$$

Unlike Maxwell, the product of A and F left on the right-hand side is the colour current carried by the gauge field itself (as announced in 3-02 § 06). (2) is a non-linear equation containing terms up to third order in A, and Jaffe–Witten write "The Yang–Mills equations are nonlinear—in contrast to the Maxwell equations." Known . In differential forms it is $d_A{*}F = 0$ (1-06).

The current is conserved only covariantly

When there is a matter field, a colour current appears on the right-hand side, with the same sign as in Maxwell: $D_\mu F^{\mu\nu} = -j^\nu$. On the left-hand side there is an identity that holds for any A (each term is a commutator of a matrix with itself, so it is 0).

$$D_\nu D_\mu F^{\mu\nu} = \tfrac{1}{2}[D_\nu, D_\mu] F^{\mu\nu} = -\tfrac{ig}{2}\,[F_{\nu\mu}, F^{\mu\nu}] = 0$$

So $D_\nu j^\nu = 0$ is required, and what is conserved in the ordinary sense is the sum of the matter current and the field current, $j^{a\nu} + g f^{abc} A^b_\mu F^{c\,\mu\nu}$ (the difference from the conservation of charge in 2-03).

03

Writing it with the colour electric and magnetic fields — the non-abelian Maxwell equations

Putting $F^{i0} = -E_i$ and $F^{ij} = \varepsilon_{ijk} B_k$ into $D_\mu F^{\mu\nu} = -j^\nu$ gives the laws of Gauss and Ampère–Maxwell (the Maxwell equations with sources of 2-03) with ∇ replaced by the covariant derivative $(\mathbf{D}_i X)^a = \partial_i X^a + g f^{abc} A^b_i X^c$ ($\rho = j^0$).

$$(\mathbf{D}\cdot\mathbf{E})^a = \rho^a, \qquad (\mathbf{D}\times\mathbf{B})^a - (D_t\mathbf{E})^a = \mathbf{j}^a \quad (3)$$

Gauss's law contains no second time derivative; it is a constraint at each time. With $A_0 = 0$, if it holds at the start, it keeps holding by the identity of § 02. The constraint $\varepsilon^{abc} A^b_i E^c_i = 0$ checked in 3-02 § 07 is this equation for a uniform field.

This constraint tells us how many components can be chosen freely. For each colour, A_μ has 4 components. Since no time derivative of A_0 appears in (1) (F_{i0} contains $\partial_0 A_i$ but not $\partial_0 A_0$), A_0 is not a dynamical variable and can be made $A_0 = 0$ by a gauge transformation. Of the 3 remaining components of A_i, the longitudinal one along the direction of propagation, seen in small waves, is removed by Gauss's law and by time-independent gauge transformations that preserve $A_0 = 0$. So $4 - 1 - 1 = 2$ remain for each colour, the same number as the two polarisations of light. For $SU(N)$ there are $N^2 - 1$ colours, so $2(N^2 - 1)$; for $SU(3)$ it is 16 Known (the quantum version is in 4-01).

04

The Bianchi identity — the other half comes from the construction

The source-free half of the Maxwell equations came from building the field out of a potential (the $dF = 0$ of 1-06). In the non-abelian case too, the following holds just because F is the curvature of A Known .

$$D\lambda F_{\mu\nu} + D_\mu F_{\nu\lambda} + D_\nu F_{\lambda\mu} = 0 \quad (4)$$

The reason is the Jacobi identity for operators, $[D_\lambda, [D_\mu, D_\nu]] + [D_\mu, [D_\nu, D_\lambda]] + [D_\nu, [D_\lambda, D_\mu]] = 0$. Putting in $[D_\mu, D_\nu] = -igF_{\mu\nu}$ ((6) of 3-02) and $[D_\lambda, F_{\mu\nu}] = D_\lambda F_{\mu\nu}$, which comes from the product rule, the left-hand side is $-ig$ times (4). In the geometric notation it is (12) of 1-07, and in Jaffe–Witten $0 = d_A F$. In electric and magnetic fields it is two equations, paired with (3).

$$\mathbf{D} \cdot \mathbf{B} = 0, \qquad \mathbf{D} \times \mathbf{E} + D_t \mathbf{B} = 0 \quad (5)$$

Unlike electromagnetism, the ordinary divergence $\nabla \cdot \mathbf{B}^a = -g f^{abc} \mathbf{A}^b \cdot \mathbf{B}^c$ is not necessarily 0; looking at one colour alone, magnetic field lines may end partway. More precisely, $\nabla \cdot \mathbf{B}^a$ is a quantity that changes under gauge transformations, and the gauge-independent statement is the $\mathbf{D} \cdot \mathbf{B} = 0$ of (5). The flux through the six faces of a small cube is measured in a different colour frame on each face; when they are parallel transported (1-07) to one common point before being added, the term of the order of the volume vanishes (the form of the first equation of (5) integrated over the cube). For a random field, the left-hand side of (4) was 0 within the error of the differences (4×10^{-7}), and the cyclic sum with ordinary derivatives was 2.1–2.6 Computed .

05

Energy and momentum — the trace vanishes in four dimensions

The Noether conserved quantity for translations (2-01), tidied into a gauge-invariant form, is the energy–momentum tensor Known .

$$T^{\mu\nu} = F^{a\,\mu\alpha} F^{a\,\nu}{}_{\alpha} - \tfrac{1}{4}\,\eta^{\mu\nu} F^{a}_{\alpha\beta} F^{a\,\alpha\beta} \quad (6)$$

$T^{00} = \frac{1}{2}\sum_a (|\mathbf{E}^a|^2 + |\mathbf{B}^a|^2)$ is the energy density ((10) of 3-02) and $T^{0i} = \sum_a (\mathbf{E}^a \times \mathbf{B}^a)_i$ the momentum density; since they are summed over colours they do not change under gauge transformations. The energy density is at least 0, and it is 0 only where $F = 0$ (pure gauge, 3-02 § 04). Taking the divergence and using (4), two derivative terms of F cancel, and

$$\partial_\mu T^{\mu\nu} = (D_\mu F^{\mu\alpha})^a\, F^{a\,\nu}{}_{\alpha} \quad (7)$$

remains. It is 0 on solutions, and energy and momentum are conserved. If there is matter the right-hand side is $-j^{a\alpha} F^{a\,\nu}{}_{\alpha}$, and the field loses as much as the colour Lorentz force does on the matter. From $\eta_{\mu\nu}\eta^{\mu\nu} = d$ (the dimension of spacetime), the trace is

$$T^{\mu}{}_{\mu} = \left(1 - \frac{d}{4}\right) F^{a}_{\alpha\beta} F^{a\,\alpha\beta} \quad (8)$$

and it vanishes identically only for $d = 4$. This is the stress side of saying that the classical action has no length scale (2-04, 1-12) Known . In the quantum theory, perturbative calculation gives the trace a F^2 term proportional to the β function (the rate at which the coupling constant moves with the scale; 4-06) (the trace anomaly), and this is taken to be the same phenomenon as the dimensional transmutation of 4-07 Physics . A figure that places the coefficient of the trace, $1 - d/4$, next to other quantities that switch at four is the dimension knob of 1-12 § 10.

For random SU(2) and SU(3) fields, the difference between $D_\nu D_\mu F^{\mu\nu}$ and the two sides of (7) was within the error of the differences (within 2.2×10^{-5}), T^{00}, T^{0i}, the trace and the form of (3) were within rounding error, and the coefficient in (8) was $d = 3, 5, 6$ also for $1 - d/4$ Computed .

06

There is no static lump — trace 0 and the balance of forces

If particles were to be represented by classical fields, the candidates would be solutions whose energy stays in one place (static lumps). For pure Yang–Mills fields in $3 + 1$ dimensions, the only time-independent solution that is smooth everywhere and

has finite energy is the vacuum Known . We follow the argument. If $T^{\mu\nu}$ is time-independent, (7) gives $\partial_k T^{kj} = 0$ ($T^{\mu\nu}$ is symmetric by (6), so the divergence is the same whichever index it is taken on). Together with $\partial_k x^i = \delta_{ik}$ this gives $\partial_k(x^i T^{kj}) = T^{ij}$, so integrating over a ball of radius R and using the divergence theorem moves the volume integral to a surface integral (usable because the field is smooth inside the ball).

$$\int_{|\mathbf{x}|<R} T^{ij}\, d^3x = \oint_{|\mathbf{x}|=R} x^i T^{kj} n_k \, dS$$

The right-hand side is of the order of the product of the area of the sphere $4\pi R^2$, the length of the arm R, and $|T|$ on the sphere. For example, if T decays faster than r^{-3}, it goes to 0 as $R \to \infty$ and $\int T^{ij} d^3x = 0$. In fact finite energy alone is enough (the detour below). Together with $-T^{00} + \sum_i T^{ii} = 0$ from (8),

$$E = \int T^{00}\, d^3x = \sum_i \int T^{ii}\, d^3x = 0 \quad (9)$$

follows. The energy density is a continuous function that is at least 0, so it is 0 everywhere, and $F = 0$ everywhere. In this form the outward pressure of the lump, $\sum_i T^{ii}$, equals the energy, and there is nothing to push back. Papers that put related conclusions in their titles are Deser (1976) “Absence of static solutions in source-free Yang–Mills theory” and Coleman (1977) “There are no classical glueballs” (bibliographic data only; the methods of proof have not been checked in the text). Adding a constant with the dimension of mass (a Higgs field and so on) makes the trace non-zero, and this argument no longer applies.

Detour: no assumption on the rate of decay is needed

A subsection you may skip.

Writing (6) in electric and magnetic fields, for each colour $T^{ij} = \frac{1}{2}\delta_{ij}(|\mathbf{E}|^2 + |\mathbf{B}|^2) - E_i E_j - B_i B_j$. Sandwiching between unit vectors $\mathbf{n}$ gives $\frac{1}{2}(|\mathbf{E}|^2 + |\mathbf{B}|^2) - (\mathbf{n}\cdot\mathbf{E})^2 - (\mathbf{n}\cdot\mathbf{B})^2$, and even summed over colours it lies between $-T^{00}$ and T^{00}. Since it is a symmetric matrix, $|T^{kj} n_k| \le T^{00}$, and the size of the surface integral is at most $R f(R)$. Here $f(R) = \oint_{|\mathbf{x}|=R} T^{00} dS$ satisfies $\int_0^\infty f(R)\, dR = E$.

If $R f(R) \ge c > 0$ always held for large R, $\int^\infty c/R\, dR$ would diverge, contradicting the finiteness of E. So a sequence of radii $R_n \to \infty$ with $R_n f(R_n) \to 0$ can be chosen, and along that sequence the surface integral tends to 0. The left-hand side also converges from $|T^{ij}| \le T^{00}$ to $\int T^{ij} d^3x$, so (9) holds with only "smooth everywhere and of finite energy" Known .

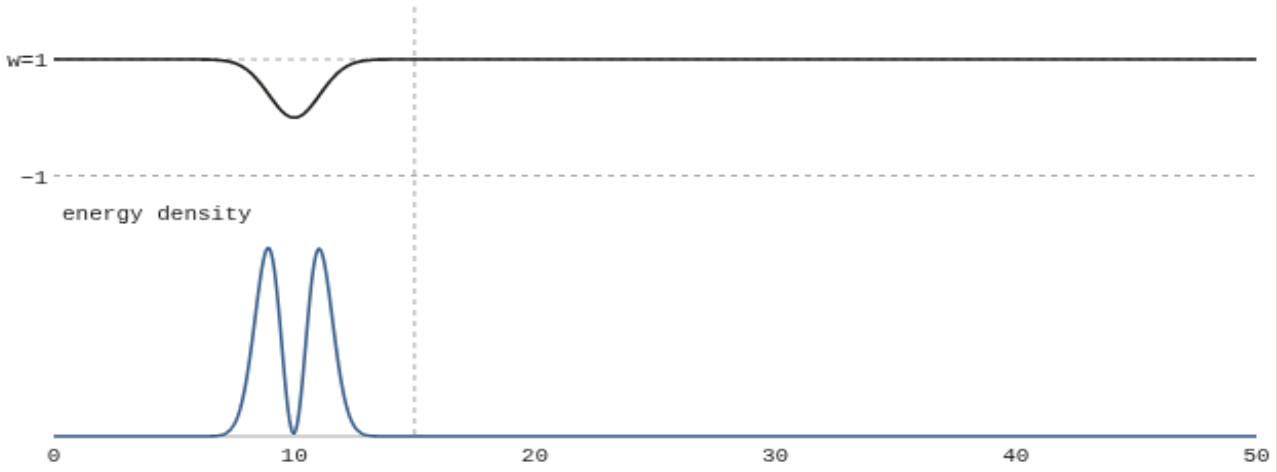
Seen in a spherically symmetric field

In SU(2), taking the form $A_0 = 0$, $A_i^a = \varepsilon_{aij}x_j\,(1 - w(r,t))/(gr^2)$, which is unchanged when space and colour are rotated together, the problem becomes one in a single variable $r = |\mathbf{x}|$.

$$E = \frac{4\pi}{g^2}\int_0^\infty \Big[w_t^2 + w_r^2 + \frac{(1-w^2)^2}{2r^2}\Big]\,dr, \qquad w_{tt} = w_{rr} + \frac{w(1-w^2)}{r^2} \quad (10)$$

$w = \pm 1$ is a pure gauge with energy 0. On a three-dimensional field with differences, we checked that the energy density (to accuracy 10^{-8}), Gauss's law and the rest of (3) reduce to (10) Computed .

Figure 1 The knob a changes the depth of the dip ($w(r,0) = 1 - a\,e^{-((r-10)/1.5)^2}$), and "Play" advances time. Top: w; bottom: the energy density per dr; the dotted line is $r = 15$



a = 1, t = 0.00, E(t)/E(0) = 1.0000, fraction in r<15 = 1.000

Computed The numbers in this figure were computed on this machine. It solves the same differences as Table 1 ($h = 0.05$, leapfrog with $\Delta t = h/2$) in the same order, and agrees with python3 to 12 digits. If JavaScript is disabled, Table 1 is the content of the figure.

Table 1 Time evolution of $a = 1.0$. The position of a peak is the maximum of the energy density per dr. The drift in the total energy $0.84922 \times 4\pi/g^2$ is at most 2×10^{-4} relative, for $t \le 40$

t	Fraction in $r < 15$	Outgoing peak	Incoming → reflected peak
0	1.0000	–	–
10	0.4996	18.95	1.10
20	0.4983	28.95	10.90
30	0.0000	38.90	20.90

The dip splits in two and travels at the speed of light, 1; the incoming peak reflects at the origin at $t \approx 10$, and by $t = 30$ nothing is left in $r < 15$. The same held for $a = 0.3$, 1.8 Computed . What is shown covers only these initial values and a finite range of time.

Detour: there is a static solution, but its energy is infinite

A subsection you may skip.

The $w = 0$ of (10) is a time-independent solution, and its magnetic field has the form of a point magnetic charge, $|\mathbf{B}| = 1/(gr^2)$ Computed . It is singular at the origin and the energy $\int 4\pi r^2\, dr/(2g^2r^4)$ diverges at the origin, so it lies outside the assumptions of (9) (smooth everywhere, finite energy). It is the solution said to be called by the names of Wu and Yang.

07

The action in imaginary time and self-duality — a lower bound on the action

We move to imaginary time $t = -i\tau$, $x_4 = \tau$ (2-04 § 09). From $A_0\, dt = A_4\, d\tau$ we get $A_0 = iA_4$, and also $\partial_t = i\partial_4$, so $F_{i0} = iF_{i4}$. Rewriting as $E_i^a = F_{i4}^a$, the sign of $|\mathbf{E}^a|^2$ in (1) flips, and iS becomes $-S_E$. The Euclidean action S_E is a sum of positive quantities, and in the quantum theory it enters the weight e^{-S_E} (4-02).

$$S_E[A] = \frac{1}{4}\int F^a_{\mu\nu}F^a_{\mu\nu}\, d^4x = \frac{1}{2}\int \sum_a \left(|\mathbf{E}^a|^2 + |\mathbf{B}^a|^2\right) d^4x \;\geq 0 \qquad (11)$$

The Hodge star $(*F)_{\mu\nu} = \frac{1}{2}\varepsilon_{\mu\nu\rho\sigma}F_{\rho\sigma}$ ($\varepsilon_{1234} = 1$, 1-12) exchanges $\mathbf{E}$ and $\mathbf{B}$ in this notation, so self-dual $F = *F$ means $\mathbf{E}^a = \mathbf{B}^a$, and anti-self-dual means $\mathbf{E}^a = -\mathbf{B}^a$. At each point $\frac{1}{2}(|\mathbf{E}|^2 + |\mathbf{B}|^2) = \frac{1}{2}|\mathbf{E} \mp \mathbf{B}|^2 \pm \mathbf{E}\cdot\mathbf{B}$, so

$$S_E = \frac{1}{2}\int \sum_a |\mathbf{E}^a \mp \mathbf{B}^a|^2\, d^4x \pm \frac{8\pi^2}{g^2}Q \;\geq\; \frac{8\pi^2}{g^2}|Q|, \qquad Q = \frac{g^2}{8\pi^2}\int \sum_a \mathbf{E}^a\cdot\mathbf{B}^a\, d^4x \qquad (12)$$

with equality only when self-dual ($Q > 0$) or anti-self-dual ($Q < 0$) Known . For a solution of the Yang–Mills equations with finite action, it is said that it extends to a field on S^4 and Q is an integer (Uhlenbeck 1982, bibliographic data only). General fields that are not solutions are treated in 3-05 § 03. In the geometric notation of 3-05 (anti-Hermitian A and c_2) it is $Q = -c_2$ (the detour of 3-05).

A self-dual field is automatically a solution, because $D_\mu F_{\mu\nu} = \pm\frac{1}{2}\varepsilon_{\mu\nu\rho\sigma}D_\mu F_{\rho\sigma}$ is 0 by the cyclic sum (4). Instead of the second-order (2), it is enough to solve the first-order $F = \pm *F$, and its solutions have the least action within the same Q. Jaffe–Witten § 2 write that classical non-abelian gauge theory has played a large role in the study of three- and four-dimensional manifolds Known . The representative of this in four dimensions is the first-order (anti-)self-dual equation, and Donaldson (1983) used the space of its solutions for four-dimensional topology (3-06) Known . Real self-dual fields exist only in imaginary time (1-06).

08

A stretched instanton — Q is unchanged and only the action grows

As an example we borrow the BPST solution (Belavin–Polyakov–Schwartz–Tyupkin 1975), constructed in 3-05 Known . In 't Hooft's symbols ($\eta_{aij}=\varepsilon_{aij}$, $\eta_{ai4}=-\eta_{a4i}=\delta_{ai}$), the solution of size ρ has the following form.

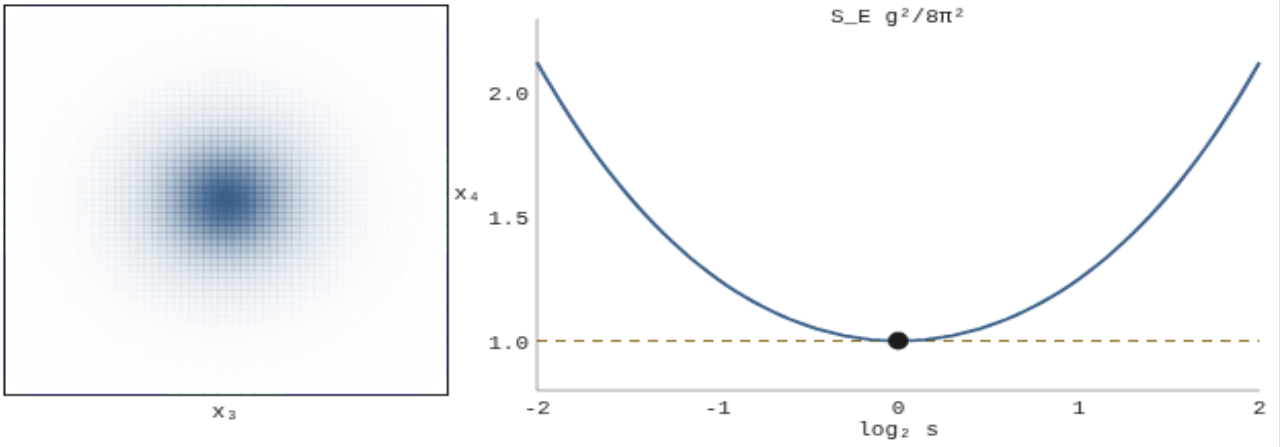
$$A^a_\mu=\frac{2}{g}\,\frac{\eta_{a\mu\nu}x_\nu}{|x|^2+\rho^2},\qquad F^a_{\mu\nu}=-\frac{4}{g}\,\frac{\eta_{a\mu\nu}\,\rho^2}{(|x|^2+\rho^2)^2}\quad(13)$$

Since η_a is self-dual, $\mathbf{E}^a=\mathbf{B}^a$, $S_E=8\pi^2/g^2$, $Q=1$ (Table 1 of 1-12). With differences we checked the right-hand formula of (13), $F=*F$ and $D_\mu F_{\mu\nu}=0$ to within 6×10^{-10} Computed . We deform this in the direction of imaginary time. Pulling back by $M=\operatorname{diag}(1,1,1,s)$ as $A^{(s)}_\mu(x)=M_{\nu\mu}A_\nu(Mx)$, for $s>1$ the shape of the field shrinks by a factor $1/s$ in the x_4 direction, and for $s<1$ it stretches by a factor $1/s$. $\mathbf{E}^{(s)}=s\,\mathbf{E}(Mx)$, $\mathbf{B}^{(s)}=\mathbf{B}(Mx)$, and only the components $Fi4$ of $\mathbf{E}$ that have x_4 as an index are multiplied by s (because s multiplies the derivative in x_4 and A_4). For $s\ /=1$ it is not a solution. Substituting $y=Mx$ ($d^4x=d^4y/s$) gives the following.

$$S_E^{(s)}=\frac{8\pi^2}{g^2}\cdot\frac12\Big(s+\frac1s\Big),\qquad Q^{(s)}=1\quad(14)$$

The action is least at $s=1$ and grows whether stretched or shrunk (the inequality of arithmetic and geometric means). Q does not change because the integrand $\operatorname{tr}F\wedge F$ is a four-form that does not use the metric, and its integral does not change under orientation-preserving changes of coordinates (1-06). The action uses the Hodge star, that is, the metric, so it changes.

Figure 2 Moving $\log_2 s$ with the knob stretches and shrinks the cross-section of the action density on the left ($x_1=x_2=0$, vertical axis imaginary time x_4) (vertically shrunk for $s>1$), and the point on the right moves along the curve $\frac12(s+1/s)$. The dashed line is the lower bound $Q=1$



s = 1.000 |E|/|B| = 1.000 S_E·g²/8π² = 1.0000 Q = 1

Computed The numbers in this figure were computed on this machine. The figure computes (14) and the closed formula for the action density on the spot (agreeing with Table 2).

Table 2 The instanton stretched in the x_4 direction, integrated numerically in four-dimensional polar coordinates in x without a change of variables ($g = \rho = 1$; 96 radial points, $48 \times 64 \times 64$ angular points) Computed . S_E agrees with (14) to 6 digits, and $\geq |Q| = 1$ for every s

s	0.25	0.5	0.8	1	1.25	2	4
$SEg^2/8\pi^2$	2.125000	1.250000	1.025000	1.000000	1.025000	1.250000	2.125000
Q	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

09

What the classical theory says, and what it does not

The classical Yang–Mills field has no static lump of finite energy (§ 06), no lower bound on the frequency of waves (3-02 § 07), and in imaginary time the minimum of the action in each Q is attained by self-dual fields. None of these is a mass gap. Jaffe–Witten § 2 call the mass gap and confinement "quantum properties that do not have classical analogs" Known . Since the classical theory has no scale, the scale of the gap is thought to enter in the process of quantisation (4-07) Physics . This chapter shows nothing about the quantum theory or about the mass gap.

A bridge to the chapters ahead. In 3-04 we study the shape of the space in which fields related by gauge transformations are regarded as the same (the gauge orbit space), and the reason why the gauge cannot be fixed in one way globally. That the Q of § 07 is an integer and the construction of the BPST solution are in 3-05, and the story of how the space of solutions of self-dual fields distinguishes four-dimensional topology is the detour of 3-06. The trace and scale of § 05 continue on the quantum side in 4-06 (asymptotic freedom) and 4-07 (dimensional transmutation).

What this chapter can say

Matter	What is said	What is not said
(1)–(8)	Derivations and numerical checks Known Computed	Matter fields (3-07), the quantum trace Physics
(9)	The argument about time-independent solutions that are smooth everywhere and of finite energy Known	Removing the smoothness assumption (the $w = 0$ in the detour of § 06 is singular)
(10), Figure 1, Table 1	The scattering seen over finite time for three initial values Computed	A proof for general initial values

Matter	What is said	What is not said
(12)–(14), Table 2	Derivations and 6-digit numerical checks Known Computed	That Q is an integer (3-05)
The mass gap	—	That the quantum Yang–Mills theory has a gap (open; this site shows nothing) Physics

Sources and reproduction

Item	Kind	Source / tools
Quotations and mentions in § 01, § 02, § 04, § 07, § 09	Primary source (text checked)	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory” (2000), the official problem description of the Clay Mathematics Institute, § 1 and § 2. Collected in *The Millennium Prize Problems* (2006)
Non-existence of static solutions, the BPST solution, extension of solutions of finite action	Bibliographic data only (Crossref)	S. Deser, “Absence of static solutions in source-free Yang–Mills theory”, Phys. Lett. B 64 (1976) 463–464 / S. Coleman, “There are no classical glueballs”, Commun. Math. Phys. 55 (1977) 113–116 / A. A. Belavin, A. M. Polyakov, A. S. Schwartz, Yu. S. Tyupkin, “Pseudoparticle solutions of the Yang-Mills equations”, Phys. Lett. B 59 (1975) 85–87 / K. Uhlenbeck, “Removable singularities in Yang-Mills fields”, Commun. Math. Phys. 83 (1982) 11–29
Donaldson in § 07	Bibliographic data only (quoted in the text of 3-06)	S. K. Donaldson, “An application of gauge theory to four-dimensional topology”, J. Differential Geom. 18 (1983) 279–315
The name of the $w = 0$ solution	Not checked	The primary source has not been checked. Candidate: T. T. Wu, C. N. Yang, “Some solutions of the classical isotopic gauge field equations”, in H. Mark, S. Fernbach (eds.), *Properties of Matter Under Unusual Conditions* (1969), p. 349 ff. (bibliographic data only; whether it contains this solution has not been checked)
Uniqueness of the action (with the scale condition), the variation, (4)–(9), (12)	Known	Standard material. Derivations in the text. Conventions as in 3-02 § 02
Numbers in § 03–§ 05, § 06 and § 08, Table 1, Table 2, Figure 1, Figure 2	Computed on this machine	Python 3.12, NumPy 1.26 (`identity_checks.py`, `radial_reduction.py`, `radial_sim.py`, `instanton_stretch.py`). Figure 1 solves the same differences in plain JavaScript (checked against Node.js)

Next chapter to read: 3-04 Gauge theory as geometry. For instantons, 3-05; for the lattice action, 5-01.

Gauge theory as geometry — divide a straight space by orbits, and representatives can no longer be chosen continuously

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — The set of all connections $\mathcal{A}$ is straight, but the orbit space, in which configurations related by gauge transformations are identified, is curved and has corners. Gauge fixing is choosing a slice of the orbit space, and a slice meets the same orbit many times (Gribov copies). By Singer's theorem, for semisimple groups such as $SU(N)$ over S^4 (and S^3) there is no continuous slice.

Prerequisite chapters — 1-07 (connections, gauge transformations, holonomy) and 3-02 (the non-abelian field strength). Haar measure (§ 04), homotopy groups and principal bundles (§ 07) are defined where they appear.

Said first — This is a chapter of classical geometry, treating neither quantum theory nor the mass gap. Singer's theorem is presented as a statement and the outline of its proof. The figure and tables are computations in lattice models with finitely many edges; they do not count copies in the four-dimensional continuum theory.

Order of this chapter

1. The space of configurations — each whole field as one point
2. The space of connections — straight, with no holes
3. The gauge group and orbits — tangents, and the elements that fix
4. The orbit space — curved, with corners (Table 1). Detour: the orbit space and the gap
5. Gauge fixing is a slice — copies and the horizon (Figure 1, Table 2)
6. Going once round leads to another copy (Table 3)
7. Singer's theorem — the outline in four steps (Table 4)

8. How it has been avoided

01

The space of configurations — the whole field seen as one space

In mechanics, the set of all positions of a particle was seen as one space (the configuration space) (2-01). For fields too, we regard one connection A as one point and write the set of all of them as $\mathcal{A}$ (infinite-dimensional). Configurations related by gauge transformations represent the same physics (3-01), so the true configuration space is the collection of bundles of related configurations. We call one such bundle an **orbit**, and the set of all of them the **orbit space** $\mathcal{A}/\mathcal{G}$. The gauge transformation is (4) of 1-07 and its lattice version ((7) of 1-07).

$$g \cdot A = gAg^{-1} - (dg)\,g^{-1}, \qquad (g \cdot U)\ell = g_y\, U_\ell\, g_x^{-1} \quad (\ell : x \to y) \ (1)$$

The U_ℓ in (1) is the parallel transport carrying from x to y along the edge ℓ ((7) of 1-07). The $U_{x,\mu}$ of 5-01 is read in the direction carrying from $x + a\hat{\mu}$ to x, so in the notation of this chapter it corresponds to U_ℓ ($\ell : x + a\hat{\mu} \to x$) ($U_\ell = U_{x,\mu}$; on a reversed edge, the inverse).

Example (lattice). On a lattice with group elements on the edges (5-01), a configuration is a tuple G^E of E elements, one per edge, and a gauge transformation is a tuple G^V of V elements, one per vertex; both are finite-dimensional. There are two questions. What shape is the orbit space? Can one representative be chosen from each orbit continuously (§ 07)?

02

The space of connections — straight, with no holes

Conventions. As in 1-07, $G = SU(N)$, $\mathfrak{g}$ is the anti-Hermitian traceless matrices, and a connection is a $\mathfrak{g}$-valued 1-form $A = \sum_\mu A_\mu\, dx^\mu$ (the $\mathcal{A} = -igA$ of the table in 3-02 is written A in this chapter, and $\mathcal{A}$ is used for the set of all connections). The base space M is $\mathbb{R}^4$, S^4, S^3 or the circle. The difference of two connections is a $\mathfrak{g}$-valued 1-form, and adding a 1-form to a connection gives a connection again (1-07 § 03). So $\mathcal{A}$ is a vector space with no fixed origin (an affine space).

$$A_t = (1-t)A_0 + tA_1 \in \mathcal{A}, \qquad g \cdot A_t = (1-t)\, g \cdot A_0 + t\, g \cdot A_1 \ (2)$$

The right-hand formula follows because the $-(dg)g^{-1}$ in (1) appears only once, with weight $(1-t)+t=1$. Moving t from 1 to 0 shrinks the whole of $\mathcal{A}$ continuously to one point A_0. We call this **contractible**, and it is the starting point of § 07. On the tangent directions a, b we put an inner product.

$$\langle a, b\rangle = -\int_M \sum_\mu \operatorname{tr}\left(a_\mu b_\mu\right) d^d x \quad (3)$$

For anti-Hermitian matrices $-\operatorname{tr}(a_\mu a_\mu) \geq 0$, and the gauge transformation $a \mapsto gag^{-1}$ preserves this inner product. **Example.** The lattice $U(1)^E$ is a torus and has loops that cannot be shrunk, so it is not contractible (that the configuration space of a lattice is not contractible matters in § 08).

03

The gauge group and orbits — tangents to orbits, and the elements that fix

The set of all gauge transformations $\mathcal{G}$ (smooth $g: M \to G$, with pointwise product) is an infinite-dimensional group. The tangent to an orbit comes from putting $g = e^{t\xi}$ ($\xi : M \to \mathfrak{g}$) into (1) and differentiating.

$$\left.\frac{d}{dt}\right|_{t=0} e^{t\xi}\cdot A = -D_A\xi, \qquad D_A\xi = d\xi + [A,\xi] \quad (4)$$

D_A is the covariant derivative in the adjoint representation (3-02 § 03). Integrating (3) by parts (the inner product of two ξ is $-\int \operatorname{tr}(\xi\eta)\, d^d x$, and boundary terms are dropped) determines the adjoint D_A^*.

$$\langle D_A\xi, a\rangle = \langle \xi, D_A^* a\rangle, \qquad D_A^* a = -\sum_\mu \left(\partial_\mu a_\mu + [A_\mu, a_\mu]\right) \quad (5)$$

That a direction a is orthogonal to the orbit is the same as $D_A^* a = 0$, and at $A = 0$ it is the Landau condition $\partial_\mu a_\mu = 0$ Known . **The elements that fix.** Multiplying $g\cdot A = A$ by g on the right gives $gA - dg = Ag$, that is,

$$g\cdot A = A \iff dg + [A, g] = 0 \quad (6)$$

which is the equation of parallel transport (1-07 § 04). g is determined by its value at one point m, and $g(m)$ must commute with the holonomy of A (Singer 1978 § 2) Known . So the based gauge group $\mathcal{G}_m = \{g : g(m) = 1\}$ moves every connection (acts freely), and the only elements that fix an **irreducible** connection, one whose holonomies have no common invariant subspace, are the centre $\mathbb{Z}_N$ (1-04). A **reducible** connection such

as $A = 0$ is fixed by a larger group (all constant G); its orbit becomes thinner, and it is a corner of the orbit space.

04

The orbit space — curved, with corners

On a lattice of L edges forming a ring, the only gauge-invariant quantity is the conjugacy class of the holonomy round the ring, $U = U_{L-1} \cdots U_0$ (1-07 § 06). For $U(1)$ it is the angle Φ ($U = e^{i\Phi}$), and the orbit space is a circle. For $SU(2)$ it is fixed by the eigenvalues $e^{\pm i\theta}$ ($0 \le \theta \le \pi$), and the orbit space is the segment $[0, \pi]$; the ends $U = \pm 1$ are corners where the elements that fix jump up to the whole of $SU(2)$. Seen as $SU(2) \cong S^3$ (unit quaternions, 1-12), the conjugacy class with angle θ is a sphere of radius $\sin\theta$, which collapses to a point at the ends. The measure of weight 1 on a group that does not change under multiplication is called **Haar measure**. For $SU(2)$ it is the uniform distribution on S^3 (1-08 § 03). Pushed down to the orbit space, it gets the weight of the area of the orbit (numerator) divided by the volume of S^3 (denominator) (the Weyl integration formula) Known .

$$\rho(\theta)\, d\theta = \frac{4\pi \sin^2\theta}{2\pi^2}\, d\theta = \frac{2}{\pi} \sin^2\theta\, d\theta \quad (7)$$

Computed **Table 1** The distribution of the angle of the holonomy θ for a ring whose 4 edges are each chosen at random with the Haar measure of $SU(2)$ (2 million samples; sampling error at most 0.0004).

Interval of θ	Samples	Integral of (7)
0°–30°	0.0289	0.0288
30°–60°	0.1666	0.1667
60°–90°	0.3047	0.3045
90°–120°	0.3044	0.3045
120°–150°	0.1667	0.1667
150°–180°	0.0288	0.0288

It agrees with (7), and the same holds for a single edge (a product of Haar-distributed elements is again Haar-distributed). Orbits near the corners are small, and the probability of landing there is small too. This weight plays the same role as the Faddeev–Popov determinant multiplied in when the gauge is fixed (the determinant of M in (9) of § 05). More precisely, when conjugation in a finite-dimensional group is fixed to diagonal form, the Faddeev–Popov determinant becomes a constant times $\sin^2\theta$ (the Weyl factor). The determinant in infinite-dimensional gauge fixing is a formal analogue of it (4-08 § 02).

Detour: the orbit space and the gap

A subsection you may skip.

In quantum theory the kinetic energy corresponds to the Laplacian on the orbit space (4-01). Karabali–Nair (1996) summarise Feynman's (1981) point as "for a finite-dimensional compact manifold the Laplacian has a discrete spectrum, the gap between the first and second eigenvalues being the analogue of the mass gap", and present a (2+1)-dimensional argument that a gap appears if gauge-invariant distances do not grow without bound Physics . Strictly this is not a theorem, and the same paper cautions that it would "have to be understood with regularizations and all that".

05

Gauge fixing is a slice — Gribov copies and the horizon

Fixing the gauge means choosing a map $s : \mathcal{A}/\mathcal{G} \to \mathcal{A}$ that picks one representative from each orbit (a **slice**). A rule often used is to choose the point on the orbit closest to the origin. From (4)(5),

$$\frac{d}{dt}\Big|_{t=0} \|e^{t\xi} \cdot A\|^2 = -2\langle A, D_A\xi\rangle = -2\langle D_A^* A, \xi\rangle \quad (8)$$

and this is 0 for every ξ when $\partial_\mu A_\mu = 0$, the Landau gauge. Singer calls the surface $\mathcal{S}_A = \{A + \tau : D_A^*\tau = 0\}$ through A orthogonal to the orbits the generalised Coulomb gauge, and states that near A the surface meets each orbit exactly once (a local slice) Known . Far away the surface becomes tangent to orbits. The tangent $D_{A+\tau}\xi$ to the orbit at $A + \tau$ lies in the surface when the following operator has a non-zero solution.

$$M(\tau)\,\xi = D_A^* D_{A+\tau}\,\xi \quad (9)$$

There the Faddeev–Popov determinant (the determinant of $M(\tau)$) is 0. For $A = 0$ it is $M(\tau) = -\partial_\mu D_{\tau,\mu}$. The range where M is positive is called the **Gribov region**, its boundary the **horizon**, and each point where an orbit meets the surface a **Gribov copy**. Singer's Theorem 8 shows that in every direction of the surface, far enough out, there is a point where it is tangent to an orbit. Gribov (1978) found, for $SU(2)$ in $\mathbb{R}^4$, that the Coulomb gauge meets the orbit of the origin once more far away (as presented by Singer) Known . How to count them in quantum theory is in 4-08 § 05 and § 06.

Example (Figure 1). A ring of L edges of $U(1)$. With $g_0 = 1$ at the base point and the rest $g_x = e^{i\alpha x}$, the angle of an edge transformed by (1) is $\psi_\ell = \theta_\ell + \alpha_{\ell+1} - \alpha_\ell$ (indices go round with L). We measure the distance from the origin with the same function as Mehta et al. (2014).

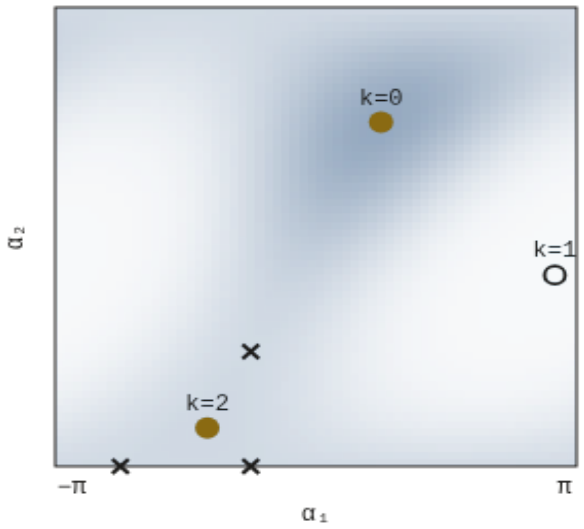
$$E(\alpha)=\sum_{\ell=0}^{L-1}\bigl(1-\cos\psi_\ell\bigr),\qquad \frac{\partial E}{\partial\alpha_j}=\sin\psi_{j-1}-\sin\psi_j \tag{10}$$

For small ψ, $E\approx\frac12\sum_\ell\psi_\ell^2$, the lattice version of (8). At stationary points all the $\sin\psi_\ell$ are equal; there are the uniform points $\psi_\ell=\varphi$ and points where only some edges have $\pi-\varphi$. The uniform points are the following L.

$$\varphi_k=\frac{\Phi+2\pi k}{L}\quad(k=0,\dots,L-1),\qquad E=L\,(1-\cos\varphi_k) \tag{11}$$

The matrix of second derivatives there (the lattice Faddeev–Popov matrix) is $\cos\varphi_k$ times the Laplacian of the ring, so for $\cos\varphi_k>0$ it is a minimum (a copy inside the Gribov region), and for negative values a maximum. Non-uniform points are not minima ($L=3$–9) Computed .

Figure 1 Changing Φ with the knob moves the shading of E (darker is smaller) on the orbit (a torus) of $L=3$ and its critical points (● minimum, × saddle, ○ maximum; k is the label in (11)). "Play" takes Φ once round, and the representative $k=0$ arrives at the original position of $k=1$. For $L=3$ it crosses the horizon on the way, so where it arrives is a maximum (§ 06).



Φ = 135°, minima: 2, E = 0.879, 2.224

Φ	Minimum	Saddle	Maximum	E at the minima	φ of the minima
0°	1	3	2	0.0000	0°
99°	2	3	1	0.4840, 2.8430	33°, −87°
135°	2	3	1	0.8787, 2.2235	45°, −75°
180°	2	3	1	1.5000, 1.5000	60°, −60°

Computed **Table 2** Critical points of the orbit of the representative $\theta=(0,0,\Phi)$, listed by (10) and classified by second derivatives. The numbers in this figure were computed on this machine (the figure computes the same formulas on the spot). If JS is disabled, this table is the content of the figure.

Only when Φ is between 90° and 270° are $\cos\varphi_0 > 0$ and $\cos\varphi_2 > 0$ both possible, and there are two copies inside the Gribov region. When Φ crosses 90° going up, the point $k = 2$ and three saddle points merge into one, and the second derivative becomes 0. Beyond this, $k = 2$ becomes a minimum. This is the horizon. At $\Phi = 180^\circ$ the E of the two minima are equal, and the boundary of the region that collects only the closest points (the fundamental modular region) is here. That the signed sum of critical points is 0 (Neuberger's 0/0) is in 7-11 § 05.

06

Going once round leads to another copy — keep choosing representatives continuously and they disagree

Taking Φ once round in Figure 1, the uniform representative of (11) moves continuously from $\varphi = \Phi/L$ to $(\Phi + 2\pi)/L$ and arrives at the representative $k = 1$. The orbit has returned to where it was, but the chosen representative is another copy. For $L = 3$ in Figure 1, it crosses the horizon at $\Phi = 270^\circ$ on the way, and where it arrives is a maximum (a copy outside the Gribov region). Arriving at another copy while staying inside the region happens for $L \geq 5$ (Table 3). The two copies are related by the gauge transformation that turns the point x by the angle $\alpha_x = 2\pi x/L$ (the transformation of winding number 1 on the continuous circle, transferred to the lattice) (on the last edge too, $\alpha_0 - \alpha_{L-1} \equiv 2\pi/L$).

$$\psi_\ell \;\mapsto\; \psi_\ell + \alpha_{\ell+1} - \alpha_\ell \equiv \psi_\ell + \frac{2\pi}{L} \pmod{2\pi}, \qquad \varphi_k \mapsto \varphi_{k+1} \tag{12}$$

Keep choosing representatives continuously, and after one round they disagree: this is the smallest form of the mechanism that prevents a continuous global gauge fixing. Whether the path hits the horizon ($\cos\varphi = 0$, that is, $\Phi = \pi L/2$) on the way depends on L.

Computed **Table 3** The result of following the minimum $\Phi = 0$ once round along Φ, and the number of copies inside the Gribov region (minimum–maximum over 4000 values of Φ).

L	Copies inside the region	Φ at which the horizon is hit	Representative after one round
3	1–2	270°	— (stops being a minimum on the way)
4	2	360° (end of the round)	—
5	2–3	Not hit	$\varphi = 2\pi/5$ ($k = 1$)
8	4	Not hit	$\varphi = 2\pi/8$ ($k = 1$)

For $L \geq 5$ the representative stays inside the Gribov region during the round and arrives at the copy $k = 1$; for $L \leq 4$ it hits the horizon on the way or at the end. In

neither case does it return to the original representative Computed .

Example (the continuous circle). For a $U(1)$ connection $A = i\,a\,dx$ on a circle of length ℓ, the Landau condition makes a constant, and the gauge transformation $e^{2\pi i n x/\ell}$ of winding number n takes a to $a - 2\pi n/\ell$, so there are infinitely many copies. Only $a\ell \bmod 2\pi$ is gauge-invariant; choosing a representative continuously means choosing a real number continuously from an angle (taking a continuous logarithm), which is impossible because after one round it is off by 2π Known .

07

Singer's theorem — a contractible space cannot be a product with a group that has holes

Singer (1978) takes as base space S^4 ($\mathbb{R}^4$ with one point at infinity added; this corresponds to the condition that gauge transformations approach a constant value far away) or S^3 (space at one time), and for compact connected semisimple groups (groups whose Lie algebra is a direct sum of simple ones; in the paper, for simplicity, $SU(N)$, $N > 1$) states in the abstract "It is shown that no continuous choice of exactly one connection on each orbit can be made.", concluding "No gauge fixing is possible." Known . Precisely, it says there is no **continuous global** gauge fixing; local slices and discontinuous choices are not forbidden.

Homotopy groups. $\pi_k(X)$ is the set of classes of based continuous maps from the k-dimensional sphere into X, grouped by continuous deformation into one another. π_0 is the connected components, π_1 the classes of loops, and $\pi_3(SU(2))$ is the degree of 1-12, with $\pi_k(X \times Y) = \pi_k(X) \times \pi_k(Y)$. The proof has four steps (Singer 1978 § 2).

1. $\mathcal{A}$ is contractible (§ 02). So is $\mathcal{A}^*$, with the set of reducible connections C removed: all its homotopy groups vanish, $\pi_j(\mathcal{A}^*) = 0$ (Singer's Theorem 2). The removed set C leaves, at every point, infinitely many directions orthogonal to it, so the image of a sphere can be pushed off C.
2. $\hat{\mathcal{G}} = \mathcal{G}/\mathbb{Z}_N$ (the centre $\mathbb{Z}_N$ fixes every connection, so it is divided out) acts freely on $\mathcal{A}^*$, and $\mathcal{A}^* \to \mathcal{N}^* = \mathcal{A}^*/\hat{\mathcal{G}}$ is a principal bundle with fibre $\hat{\mathcal{G}}$ (Singer's Theorem 1). A **principal bundle** is a bundle in which over each point of the base (here $\mathcal{N}^*$) sits a copy of the group $\hat{\mathcal{G}}$ (the **fibre**; here one orbit), and $\hat{\mathcal{G}}$ acts freely and transitively within each copy (every element other than 1 moves every point, and any point can be moved to any point). If there is a section, the whole is the product of the base and the group.
3. If there were a continuous gauge, it would be a section of the principal bundle, the bundle would be the product $\mathcal{A}^* \cong \mathcal{N}^* \times \hat{\mathcal{G}}$, and from $\pi_j(\mathcal{A}^*) = 0$ and the product

formula, $\pi_j(\hat{\mathcal{G}}) = 0$ would follow for every j.

4. But some π_j of $\hat{\mathcal{G}}$ is not 0 (Theorem 3). Contradiction (Corollary 4).

$$\pi_j(\mathcal{G}_m) \cong \pi_{j+r}\big(SU(N)\big) \qquad (M = S^r)\ \text{(13)}$$

The tool of the fourth step is (13) (Theorem 5). Singer notes that $\mathcal{G}_m$ acts freely on the whole of $\mathcal{A}$, and by (13) some homotopy group of it is not 0, so $\mathcal{A} \to \mathcal{A}/\mathcal{G}_m$ has no section either.

Known **Table 4** $\pi_j(\mathcal{G}_m)$ $(M = S^r)$ read from (13). The non-abelian columns follow Singer's text (which cites Toda's tables).

M	$U(1)$	$SU(2)$	$SU(N)$, $N \geq 3$
S^3	All 0 (note)	$\pi_0(\mathcal{G}_m) = \pi_3(SU(2)) = \mathbb{Z}$	$\pi_0(\mathcal{G}_m) = \pi_3(SU(N)) = \mathbb{Z}$
S^4	All 0 (note)	$\pi_0(\mathcal{G}_m) = \pi_4(SU(2)) = \mathbb{Z}_2$, $\pi_2(\mathcal{G}_m) = \pi_6(SU(2)) = \mathbb{Z}_{12}$	$\pi_0(\mathcal{G}_m) = \pi_4(SU(N)) = 0$ (the fourth step uses $\pi_1(\hat{\mathcal{G}}) \neq 0$)

Note: the $U(1)$ column is read from the homotopy groups $\pi_j(\mathcal{G}_m) = \pi_{j+r}(U(1))$ of the space of based maps (the π_k of $U(1)$ is 0 for $k \geq 2$); Singer's Theorem 5 is a statement for $SU(N)$.

For S^3, the contradiction comes just from the fact that gauge transformations of different winding numbers cannot be connected (the circle of § 06 is the abelian version, by $\pi_0(\mathcal{G}_m) = \pi_1(U(1)) = \mathbb{Z}$). The higher homotopy groups of $U(1)$ are 0, so for S^3 and S^4 the fourth step does not work, and Singer too writes that in the abelian case "An integral submanifold provides the gauge for quantum electrodynamics." The obstruction comes from the topology made by non-commutativity and the condition at infinity Known .

08

How it has been avoided — the lattice, local gauges, the orbit space itself

The Faddeev–Popov procedure (4-08 § 03) is correct on a local slice but cannot be used globally as it is. Singer named the route of covering the orbit space by regions where local slices work and adding them up with a partition of unity (assigning to each piece of the cover a weight such that the weights add up to 1 everywhere), and wrote "This partition of unity argument would be useful if the covering used could be made explicit." Later work has avoided the obstruction by three routes.

- **The lattice.** Wilson's lattice integrates with Haar measure, so it is finite without fixing the gauge (4-02, 5-01) Known . The gauge that sets all edges of a tree joining all vertices (a maximal tree) to 1 is a global continuous slice for the based gauge group Known . Set $g = 1$ at the root vertex, and moving along the tree to neighbouring

vertices, fix g in turn so that each edge becomes 1 (if g_x is fixed on the edge $\ell : x \to y$, then $g_y = g_x U_\ell^{-1}$). g is determined uniquely by products and inverses of the edge elements, and depends continuously on U. Since the based gauge group acts freely and transitively on the configurations of the tree edges, each orbit has exactly one representative, and it can be chosen continuously. $SU(N)^E$ is not contractible ($\pi_3(SU(N)) = \mathbb{Z}$; for $SU(2)$, the degree of 1-12), so this does not contradict Singer's theorem. Instead the symmetry of the lattice is broken. Fixing a covariant gauge BRST-symmetrically (so that a formal symmetry remains even after fixing) produces 0/0 (7-11, 4-08 § 08; for BRST, 4-08 § 04).

- **Local gauges.** On Bałaban's renormalisation group on the three-dimensional periodic lattice, the problem description writes "One defines these gauges in phase cells: The choices vary locally in space-time, as well as on different length scales." (Jaffe–Witten § 6.5) Known . It is a method of changing slices by place and by scale (7-06 § 05, 7-05 § 05). Bałaban's four-dimensional work is said to use the same idea (7-07 § 07; the specific choice in the four-dimensional papers has not been checked in the text).
- **The orbit space itself.** Stochastic quantisation does not fix the gauge and takes the orbit space as the state space. A review writes "The space O is a non-linear space if G is non-Abelian, which makes non-trivial even the construction of the state space" (Chevyrev 2022), and Chandra–Chevyrev–Hairer–Shen, on the two-dimensional torus, took as state space the space of orbits of connections as distributions (1-11) and constructed a stochastic dynamics on it (abstract) Known (7-09).

7-11 § 05 treats this as one of the obstacles other than divergence (the obstacle of the curved state space), and separates it into the part settled on the lattice and the 0/0 that remains when weak-coupling analysis needs gauge fixing.

What this chapter can say

Item	What is said	What is not said
Connections and orbits	Derivation of (1)–(9) Known	How to topologise the function spaces; the proof of the local slice theorem
Lattice models	Tables 1–3, Figure 1 Computed	The number or arrangement of copies in the four-dimensional continuum theory
Singer's theorem	The statement and the four-step outline (checked in the original paper) Known	Reproducing the proofs of Singer's Theorems 2 and 5
Detour	Feynman's overall picture Physics	Not a theorem

Item	What is said	What is not said
Yang–Mills	The meaning of the global obstruction to gauge fixing	Nothing is shown about the mass gap or the construction of the theory

Sources and reproduction

Item	Kind	Source / tools
Tables 1–3, Figure 1	**Computed on this machine**	Python 3.12, NumPy (`ring_copies.py`, `ring_bruteforce_check.py`, `su2_classes.py`). The figure computes (10)(11) on the spot
Singer's statement, quotations, Table 4	Primary source (text checked)	I. M. Singer, "Some remarks on the Gribov ambiguity", Commun. Math. Phys. 60 (1978) 7–12
Gribov copies	Bibliographic data only	V. N. Gribov, "Quantization of non-Abelian gauge theories", Nucl. Phys. B 139 (1978) 1–23 (the content in the text follows Singer's account in § 1; Singer cites the 1977 SLAC Translation 176 version)
The function (10) and its second derivatives	Primary source (§ I–II checked)	D. Mehta, N. S. Daleo, J. D. Hauenstein, C. Seaton, Phys. Rev. D 90 (2014) 054504, arXiv:1406.6678
Detour	Text § 1 checked / Feynman bibliographic data only	D. Karabali, V. P. Nair, Nucl. Phys. B 464 (1996) 135–152, arXiv:hep-th/9510157; R. P. Feynman, Nucl. Phys. B 188 (1981) 479
Quotations in § 08	Text checked / abstract	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory" (Clay, 2000) § 6.5; I. Chevyrev, "Stochastic quantisation of Yang–Mills", J. Math. Phys. 63 (2022) 091101, arXiv:2202.13359 § 1; A. Chandra, I. Chevyrev, M. Hairer, H. Shen, Publ. Math. IHÉS 136 (2022) 1–147, arXiv:2006.04987
(7), values of homotopy groups	Known (bibliographic data only)	Weyl integration formula: T. Bröcker, T. tom Dieck, *Representations of Compact Lie Groups*, Graduate Texts in Math. 98, Springer (1985), Chapter IV; homotopy groups of spheres: H. Toda, *Composition Methods in Homotopy Groups of Spheres*, Ann. of Math. Studies 49, Princeton (1962) (following Singer's bibliography)

Next chapter to read: 3-05 Instantons and topology (the integer that separates S^4 bundles, and the orbits of least action). The quantum theory of gauge fixing is 4-08.

Instantons and topology — the integer in the lower bound on the action comes from the winding number at infinity

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That the integer Q in the lower bound on the action of 3-03 comes from $\operatorname{tr} F \wedge F$ being a total derivative and from the winding number at infinity. The BPST solution written with quaternions, and how one of two instantons shrinks to a point. How the θ term changes the vacuum in quantum theory (the model of a particle on a circle).

Prerequisite chapters — 3-03 (self-duality and the lower bound (12)). For differential forms, 1-06; for pure gauge, 1-07; for the quantum mechanics of § 06 and § 07, 2-05. § 02, § 03, § 05 and § 06 of the detour chapter 1-12 (quaternions, degree, self-dual 2-forms, the fraction of the action inside) are worth reading; the minimum this chapter needs is defined in the text (quaternions and degree in § 03, the fraction inside in § 04).

Said first — § 01– § 05 are known mathematics of classical fields. The quantum discussion of § 06– § 08 is argument at the level of physics, not theorems. This chapter shows nothing about the mass gap.

Order of this chapter

1. An integer that does not change under continuous change — the winding number and conventions
2. tr F∧F is a total derivative — the integral is fixed by the boundary
3. The integer is fixed at infinity — degree (Table 1), SO(3) bundles, Detour: the sign
4. The BPST solution — one line in quaternions (Table 2)
5. Two instantons — one shrinks (Figure 1, Table 3)
6. The θ term — a particle on a circle (Figure 2, Table 4)

7. Vacua are joined by tunnelling — θ vacua (Table 5)
8. The lattice and θ in the problem statement — sectors and θ = π

01

An integer that does not change under continuous change — from the winding number in the plane

Topology deals with properties that do not change under continuous deformation. The winding number of a closed curve $z(t)$ that does not pass through the origin,

$$n = \frac{1}{2\pi i}\oint \frac{dz}{z} \in \mathbb{Z} \quad (1)$$

is an integer, so it cannot change when the curve is moved continuously; it changes only at the moment the curve crosses the origin. Example. $z = e^{it} + c$ (c real) gives 1 for $|c| < 1$ and 0 for $|c| > 1$. Moving c along the real axis, the curve passes through the origin at the moment it crosses $|c| = 1$, and only there does the value jump. The first Chern number worked the same way (1-07 § 08). This chapter derives, from the side of the field, where the integer Q in (12) of 3-03 comes from.

Conventions. A is an anti-Hermitian matrix-valued 1-form, $D = d + A$, $F = dA + A \wedge A$, and g is absorbed into A ($A = -ig\,A_{\text{phys}}$; 3-03 § 01, the table of notation). In this notation the action, the second Chern number and Q take the following form. Q is the Q of 3-03 and the k of 1-12 (the sign is in the detour of § 03).

$$S_E = -\frac{1}{g^2}\int \operatorname{tr} F \wedge *F, \qquad c_2 = \frac{1}{8\pi^2}\int \operatorname{tr} F \wedge F, \qquad Q = -c_2 \quad (2)$$

In the physics notation (the $D_\mu = \partial_\mu - igA_\mu$ of 3-03, without absorbing g into A), with $\tilde{F}_{\mu\nu} = \frac{1}{2}\varepsilon_{\mu\nu\rho\sigma}F_{\rho\sigma}$ (the components of the $*F$ of 3-03),

$$Q = \frac{g^2}{32\pi^2}\int F^a_{\mu\nu}\tilde{F}^a_{\mu\nu}\,d^4x = \frac{g^2}{8\pi^2}\int \sum_a \mathbf{E}^a\cdot\mathbf{B}^a\,d^4x$$

which is exactly the Q of (12) of 3-03.

02

tr F∧F is a total derivative — the Chern–Simons form

For U(1), $F \wedge F = d(A \wedge dA)$, and the integral was fixed by the boundary alone (1-06). In the non-abelian case it holds too, once one correction is added Known .

$$\operatorname{tr} F \wedge F = d\,\mathrm{CS}(A), \qquad \mathrm{CS}(A) = \operatorname{tr}\Big(A \wedge dA + \tfrac{2}{3} A \wedge A \wedge A \Big) \quad (3)$$

Checking it needs only one rule. For a matrix-valued p-form α and q-form β, the trace of the product of the component matrices does not change when the order is swapped, and only the sign from exchanging the differential-form parts remains.

$$\operatorname{tr}(\alpha \wedge \beta) = (-1)^{pq} \operatorname{tr}(\beta \wedge \alpha)$$

Let A^k be the wedge product of A taken k times. Expand $d \operatorname{tr} A^3$ into three terms by the product rule $d(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^p \alpha \wedge d\beta$ (1-06), and move dA to the front in each term.

$$\begin{aligned} \operatorname{tr}(dA \wedge A \wedge A) &= \operatorname{tr}(dA \wedge A^2) \\ -\operatorname{tr}(A \wedge dA \wedge A) &= +\operatorname{tr}(dA \wedge A^2) \qquad (p=1,\ q=3) \\ +\operatorname{tr}(A \wedge A \wedge dA) &= +\operatorname{tr}(dA \wedge A^2) \qquad (p=q=2) \end{aligned}$$

The sum is $d \operatorname{tr} A^3 = 3 \operatorname{tr}(dA \wedge A^2)$. Together with $d \operatorname{tr}(A \wedge dA) = \operatorname{tr}(dA \wedge dA)$ ($d\,dA = 0$), $d\,\mathrm{CS} = \operatorname{tr}(dA \wedge dA) + 2 \operatorname{tr}(dA \wedge A^2)$. Expanding $\operatorname{tr} F \wedge F = \operatorname{tr}\big((dA + A^2) \wedge (dA + A^2)\big)$ gives the same two terms ($\operatorname{tr}(A^2 \wedge dA) = \operatorname{tr}(dA \wedge A^2)$) and $\operatorname{tr} A^4$, which is 0 since $\operatorname{tr} A^4 = \operatorname{tr}(A \wedge A^3) = -\operatorname{tr}(A^3 \wedge A) = -\operatorname{tr} A^4$. The name comes from Chern–Simons (1974).

There are two consequences. First, if A is smooth on the ball B_R, Stokes' theorem (1-06) gives

$$\frac{1}{8\pi^2} \int_{B_R} \operatorname{tr} F \wedge F = \frac{1}{8\pi^2} \int_{S^3_R} \mathrm{CS}(A) \quad (4)$$

Example. For the solution of § 04 the right-hand side is $R = \rho, 3\rho, 10\rho$ at -0.5, -0.972, -0.999708, agreeing with the fraction inside in Table 1 of 1-12 (Table 2) Computed . Second, for δA that is 0 far away, the Bianchi identity (3-03 § 04) gives $\delta \operatorname{tr} F \wedge F = 2\,d \operatorname{tr}(\delta A \wedge F)$, and the integral does not change. Adding c_2 to the action does not change the equations, and the θ term of § 06 starts here.

03

The integer is fixed at infinity — pure gauge and degree

If the action is finite, far away $A \to -(dg)g^{-1}$ ($g : S^3 \to SU(2)$, pure gauge). For solutions, the basis is the abstract of Uhlenbeck (1982), "every Yang-Mills field over R4 with bounded functional (L2 norm) may be obtained from a field on S4" Known ; for general fields we assume decay far away (1-07). For pure gauge $dA = -A^2$, so CS $= \frac{1}{3}$ $\operatorname{tr}(g^{-1}dg)^3$, and the $R \to \infty$ of (4) is the following formula.

$$c_2 = \frac{1}{24\pi^2}\int_{S^3} \operatorname{tr}\left(g^{-1}dg\right)^3 = -\deg g, \qquad Q = \deg g \quad (5)$$

Checking with the identity map and with numbers

We use quaternions (1-12 § 02; this chapter needs only the following rules). $x = x1 + x2i + x3j + x4k$ is multiplied with $i^2 = j^2 = k^2 = ijk = -1$, $\bar{x} = x_1 - x_2 i - x_3 j - x_4 k$, $|x|^2 = x\bar{x}$. The set of quaternions of length 1 is S^3, which under multiplication is the same group as SU(2). The degree $\deg g$ is the integer obtained by counting the preimages of a generic target point, with $+1$ for points that preserve orientation and -1 for points that reverse it (1-12 § 06).

The orientation of S^3 is the one that, as the boundary of the ball B^4, puts the outward normal first (at the point $x = 1$, $dx2 \wedge dx3 \wedge dx4$ is positive). For the identity map $g(x) = x$, at the point $x = 1$, $g^{-1}dg = i\,dx2 + j\,dx3 + k\,dx4$. The signed sum over permutations of i, j, k is -6, so $\operatorname{tr}(g^{-1}dg)^3 = -12\,dx2 \wedge dx3 \wedge dx4$ (the quaternion 1 is the 2×2 identity matrix, with trace 2). It is invariant under left multiplication, so it is the same everywhere, and multiplying by the area $2\pi^2$ and dividing by $24\pi^2$ gives $c_2 = -1$. For general g too, the integrand is -1 times the pullback of the normalised volume form, and the integral is -1 times the degree Known .

Table 1 Numerical integration of $-\frac{1}{24\pi^2}\int_{S^3} \operatorname{tr}(g^{-1}dg)^3$ ($x \in S^3$; 64×64×128 points) Computed . The value changes only when the map before normalisation passes through 0

g	Integral	Degree
x^{-2}, x^{-1}, 1, x, x^2, x^3	-2, -1, 0, 1, 2, 3 (errors below 10^{-5} here too)	n
$\frac{x+c}{\lvert x+c\rvert}$ ($c = 0.5, 0.9$ / $1.1, 1.5$), $\frac{x^2+ci}{\lvert x^2+ci\rvert}$ ($c =$ 0.7 / 1.3)	1, 1 / 0, 0; 2 / 0.000002	1, 2 at $\lvert c\rvert <$ 1

Fields of finite action split into classes (sectors) by Q, which cannot be connected to one another continuously. This is the statement that SU(2) bundles over S^4 are classified by c_2 (1-07).

A group divided by its centre — SO(3) bundles

Taking as group $SO(3) = SU(2)/\mathbb{Z}_2$, SU(2) divided by its centre $\{\pm 1\}$ (1-04 § 07), the Lie algebra is the same, so the local formulas do not change. The difference shows in how the bundle is glued. On S^4 (fields on $\mathbb{R}^4$ that approach pure gauge far away), the gluing is a map from S^3 to the group, and since S^3 is simply connected, a map to SO(3) lifts to a map to SU(2). So every SO(3) bundle over S^4 comes from an SU(2) bundle Known .

On a closed four-dimensional manifold X with $H^2(X;\mathbb{Z}_2) \neq 0$ (for example the four-dimensional torus T^4), there are SO(3) bundles that do not lift to SU(2) bundles. The obstruction is the second Stiefel–Whitney class $w_2 \in H^2(X;\mathbb{Z}_2)$, and SO(3) bundles are said to be classified by w_2 and the first Pontryagin class p_1 (Dold–Whitney 1959) Known . Measuring the number by $-p_1/4$, so that it agrees with c_2 for bundles coming from SU(2), it is not necessarily an integer for $w_2 \neq 0$, and half-integers also appear for T^4. In physics this is said to be known as the fractional topological charge of 't Hooft's twisted boundary conditions (1979, 1981) Known . The integer Q of this chapter is about SU(2) bundles (or S^4). Centre symmetry is in 18-03.

Detour: the Q of physics and the c₂ of mathematics have opposite signs

A subsection you may skip.

Physics writes c_2 with $Q = \frac{g^2}{8\pi^2}\int \mathrm{tr}\, F_{\mathrm{phys}} \wedge F_{\mathrm{phys}}$ and mathematics with $F = -ig\, F_{\mathrm{phys}}$, so the signs differ by $(-i)^2 = -1$. Physics instantons are self-dual with $Q = 1$, while the mathematical literature often treats the anti-self-dual case ($c_2 \geq 0$) (3-06); the two are mirror images (1-12).

For how this book uses this, see 25-01 § 01 (the lattice setting of the small twisted T^2).

04

The BPST solution — written in one line with quaternions

With the quaternions $x = x_1 + x_2 i + x_3 j + x_4 k$ of § 03, reading i, j, k as $-i\sigma_1, -i\sigma_2, -i\sigma_3$, the purely imaginary quaternions are elements of su(2). With $dx = \sum_\mu e_\mu dx_\mu$ ($e_\mu = 1, i, j, k$), the solution of size ρ has the following form Known .

$$A = \mathrm{Im}\,\frac{x\, d\bar{x}}{|x|^2 + \rho^2}, \qquad F = \frac{\rho^2\, dx \wedge d\bar{x}}{(|x|^2 + \rho^2)^2} \quad (6)$$

The i components of $dx \wedge d\bar{x}$ are $-2(dx_1 \wedge dx_2 + dx_3 \wedge dx_4)$, and the components of j, k are also self-dual (1-12 § 03). So $F = *F$, and it is a solution (3-03 § 07). Far away, with $g = x/|x|$, $A \to \mathrm{Im}(x\, d\bar{x})/|x|^2 = g\, d\bar{g} = -(dg)g^{-1}$; g is the identity map of S^3 (degree 1),

so $Q=1$, $c_2=-1$. The density is $48\rho^4/(|x|^2+\rho^2)^4$, the fraction of the action inside radius R is $1-3/(1+u)^2+2/(1+u)^3$ ($u=R^2/\rho^2$; (9) of 1-12), and $S_E=8\pi^2/g^2$ is equality in the lower bound (the same property as (13) of 3-03). The abstract of BPST (1975) reads “We find regular solutions of the four dimensional euclidean Yang-Mills equations” Known . The mirror image $\mathrm{Im}(\bar{x}\,dx)/(\cdots)$ is anti-self-dual, with $Q=-1$.

Table 2 Checks of (6) and its mirror image ($\rho=1$; differences and numerical integration) Computed

Item	(6)	Mirror image
Difference from the closed formula, $F \mp *F$, c_2, Q	1.6×10⁻¹⁰, 0, -1.00000000, +1	1.8×10⁻¹⁰, 0, +1.00000000, -1
Right-hand side of (4) ($R=0.5,1,3,10$)	-0.104, -0.500, -0.972, -0.999708	Opposite sign

The parameters are 8: centre 4, size 1, and constant gauge rotation 3; for $Q=k$ there are $8k$ (Tong's lecture notes: $4kN$ for SU(N)) Known . The construction that gives all self-dual fields of finite action by linear algebra is ADHM (1978).

05

Two instantons — an action independent of the distance, and one that shrinks

A self-dual field has the least action in each Q, so for $Q=2$, wherever two lumps are placed the action is $16\pi^2/g^2$, and no force acts between the lumps. There is a concrete family. Here only, in the component notation of 3-03 ($A_\mu=A^a_\mu T^a$, $D_\mu=\partial_\mu-igA_\mu$), we set $g=1$. $\bar{\eta}$ is 't Hooft's symbol η of 3-03 § 08 with the sign of η_{ai4} changed ($\bar{\eta}_{aij}=\varepsilon_{aij}$, $\bar{\eta}_{ai4}=-\bar{\eta}_{a4i}=-\delta_{ai}$).

$$\phi=1+\sum_{i=1}^{k}\frac{\rho_i^2}{|x-a_i|^2},\qquad A^a_\mu=-\bar{\eta}_{a\mu\nu}\,\partial_\nu\ln\phi \tag{7}$$

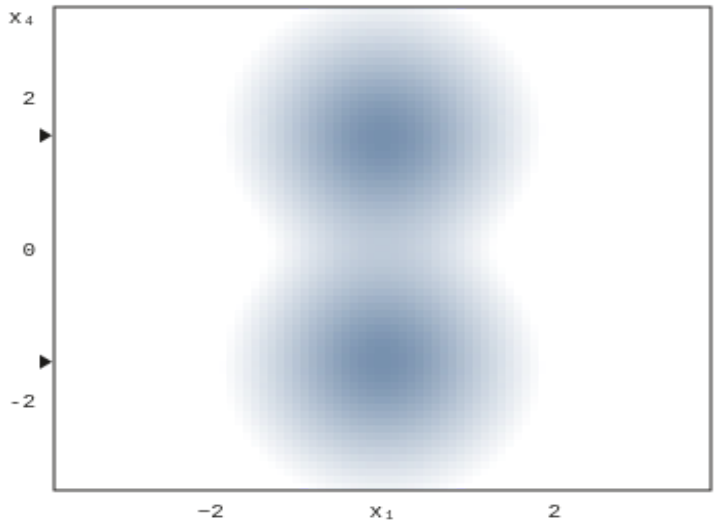
$1/|x|^2$ is a four-dimensional harmonic function, and this family is self-dual and $Q=k$ regardless of a_i,ρ_i. The A of (7) diverges at the poles a_i, but this singularity can be removed by a gauge transformation singular at that point. $k=1$ is thereby taken to the solution of § 04 translated to $a1$.

The density can be written with ϕ alone. The notation is $\Box=\sum_\mu\partial_\mu^2$ (the Euclidean four-dimensional Laplacian; not the d'Alembertian of 2-04), $\psi_\mu=\partial_\mu\phi$, $H_{\mu\nu}=\partial_\mu\partial_\nu\phi$, $|H|^2=\sum_{\mu\nu}H_{\mu\nu}^2$, $\psi\cdot H\psi=\sum_{\mu\nu}\psi_\mu H_{\mu\nu}\psi_\nu$, $|\psi|^2=\sum_\mu\psi_\mu^2$.

$$s=\tfrac{1}{4}F^a_{\mu\nu}F^a_{\mu\nu}=-\tfrac{1}{2}\,\Box\Box\ln\phi=\frac{|H|^2}{\phi^2}-\frac{4\,\psi\cdot H\psi}{\phi^3}+\frac{3|\psi|^4}{\phi^4}\qquad(x\ne a_i) \tag{8}$$

With two poles, the F from differences satisfies $F = *F$ and (8) with relative error 10^{-9} Computed . Corrigan–Fairlie (1977) treated solutions of this type as a relation between solutions of a scalar field theory of φ^4 type and solutions of the SU(2) gauge field. Jackiw–Nohl–Rebbi (1977) extended the known family of $5k$ parameters, representing the position and size of each lump, to solutions with $5k + 4$ parameters (both from the abstracts) Known .

Figure 1 Changing the distance d between the two poles with the knob changes the action density (log shading) on the cross-section $x_2 = x_3 = 0$ (horizontal x_1, vertical x_4, triangles are the poles). For small d a sharp peak rises at the origin



d = 3.00, s(0) = 2.100, 48/(d/2)⁴ (approximation for small d) = 9.481, total action = 16π² (Table 3)

Computed The numbers in this figure were computed on this machine. (8) is computed on the spot ($\rho_i = 1$, poles at $x_4 = \pm d/2$).

Table 3 The action of two poles ($\rho_1 = \rho_2 = 1$) (in units of $8\pi^2$; adaptive integration) Computed

d	4	1	0.5	0.2	0.1	0
Total action	2.000000	2.000000	2.000000	2.000000	2.000000	1.000000
Inside $\lvert x\rvert < d$	1.945544	1.140005	0.925267	0.896362	0.895881	–

The total action is $16\pi^2$ for every d. For small d the density at the origin approaches $48/(d/2)^4$ (7.661×10^6 against 7.68×10^6 at $d = 0.1$), and the 0.896 inside radius d is the fraction $1 - \frac{3}{25} + \frac{2}{125}$ for the BPST solution of size $d/2$ ((9) of 1-12). One lump shrinks while keeping its action, and at $d = 0$ it becomes the $Q = 1$ solution of $\phi = 1 + 2/\lvert x\rvert^2$, one lump's worth vanishing into a point. This is the concentration of 1-12 § 05 happening within a family of solutions (3-06).

06

The θ term — an angle invisible to the equations

By § 02, adding a constant multiple of Q to the action does not change the classical equations. In the quantum weight it stays imaginary even in imaginary time (4-02; the sign depends on conventions).

$$e^{-S_E[A]} \longrightarrow e^{-S_E[A]+i\theta Q[A]} \quad (9)$$

Known (i) In perturbation theory (the expansion around $A = 0$; 4-03) the integral of a total derivative is 0 at every order, and it cannot be seen. (ii) Since Q is an integer, θ is an angle of period 2π. (iii) In spacetime it is proportional to $\mathbf{E}^a \cdot \mathbf{B}^a$ (§ 01, 2-04); $\mathbf{E}$ changes direction under spatial reflection and $\mathbf{B}$ under time reversal, so $\theta \neq 0, \pi$ breaks spatial reflection and time reversal. The abstract of Jackiw–Rebbi (1976) describes θ vacua (§ 07) as "a family of quantum mechanical vacua, parametrized by a CP-nonconserving angle" (CP is the transformation combining the exchange of particles and antiparticles with spatial reflection). At θ = 0 and π it is invariant under time reversal (Gaiotto–Kapustin–Komargodski–Seiberg 2017, hereafter GKKS, abstract: "time reversal invariant at θ = 0 and θ = π").

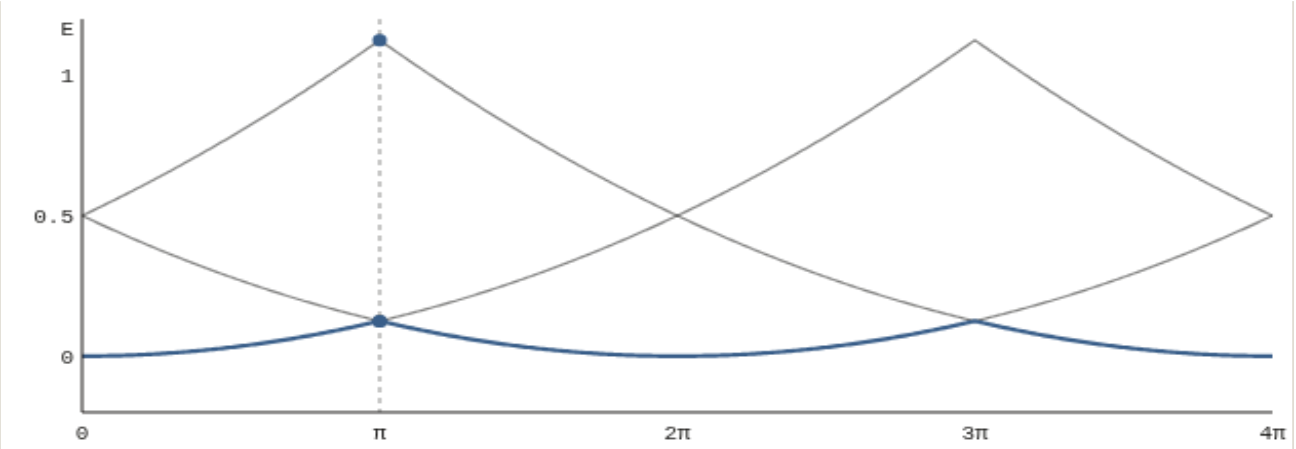
A particle on a circle — the smallest model of θ

There is quantum mechanics with the same structure (GKKS Appendix D.1). To a particle on the angle $q \sim q + 2\pi$ we add a term linear in the velocity.

$$L = \frac{1}{2}\dot{q}^2 + \frac{\theta}{2\pi}\dot{q}, \qquad E_n = \frac{1}{2}\left(n - \frac{\theta}{2\pi}\right)^2 \quad (n \in \mathbb{Z}) \quad (10)$$

The added term is $\frac{d}{dt}\frac{\theta q}{2\pi}$ and does not change the equations (2-01), but $\theta q/2\pi$ shifts by θ in one round, and the weight of periodic paths in imaginary time gets a factor $e^{i\theta \times \text{winding number}}$. This is the same form as (9). With periodic wave functions, the eigenvalues of the eigenfunctions e^{inq} of $H = \frac{1}{2}(-i\,d/dq - \theta/2\pi)^2$ are (10) (2-05). θ plays the role of a magnetic flux threading the circle (GKKS footnote 27; 2-03), and "for θ = π and only in this case the ground state is two-fold degenerate". Adding $-K\cos q$ opens the crossing, while for $-K\cos 2q$ the two-fold degeneracy remains exactly (same appendix: "still maintains some exact two-fold degeneracy").

Figure 2 Moving θ (horizontal axis 0–4π) with the knob moves the vertical line and the levels (dots). Switching the potential with the buttons between 0, $-K\cos q$ and $-K\cos 2q$ ($K = 0.4$) shows whether the crossing at $\theta = \pi$ opens or remains



θ = 1.00π, E_0 = 0.125, $E_1 - E_0$ = 0.000

Computed The numbers in this figure were computed on this machine. V = 0 is (10); the others are a table diagonalised with e^{inq} ($|n| \le 40$) (13 points of $0 \le \theta \le \pi$), extended by periodicity and symmetry.

Table 4 $E_1 - E_0$ ($K = 0.4$, $I = \hbar = 1$) Computed

Potential	θ = 0	π/2	π
0	0.5	0.25	0
$-K\cos q$	0.60436	0.46493	0.38481
$-K\cos 2q$	0.32981	0.21083	0

07

Vacua are joined by tunnelling — θ vacua

From here on the level is that of physics Physics . Setting $A_4 = 0$, the classical vacua are three-dimensional pure gauges $-(dh)h^{-1}$ (with $h : \mathbb{R}^3 \to SU(2)$ approaching 1 far away, regarded as $\mathbb{R}^3 \cup \{\infty\} = S^3$; a different sphere from the S^3 at infinity of § 03), numbered by the degree n, and the gauge transformation T of degree 1 shifts n by one. Applying (4) to a slab of imaginary time, Q is the difference of the vacuum numbers at $\tau \to \pm\infty$, and the BPST solution is a path in imaginary time from n to $n+1$. The weight is

$$e^{-8\pi^2/g^2}: \qquad g^2 = 1,\ 2,\ 4,\ 8 \text{ give } 5.1\times10^{-35},\ 7.2\times10^{-18},\ 2.7\times10^{-9},\ 5.2\times10^{-5}$$ (1

Computed , and since every derivative at $g^2 = 0$ is 0, it appears at no order of perturbation theory (the double well of 2-05, 4-07, 4-12). In 1976 't Hooft computed "the tunneling process described by the Belavin-Polyakov-Schwarz-Tyupkin field configuration", and Jackiw–Rebbi and Callan–Dashen–Gross described the family of vacua (abstracts) Known . The states that are invariant under small gauge transformations (4-01) and change only by a phase under T are the θ vacua.

$$|\theta\rangle = \sum_{n\in\mathbb{Z}} e^{in\theta}|n\rangle, \qquad T|\theta\rangle = e^{-i\theta}|\theta\rangle \tag{12}$$

For the particle on a circle, the valleys $q = 2\pi n$ of the line before it is wrapped onto the circle are $|n\rangle$, $q \mapsto q + 2\pi$ is T, and (12) has the form of a Bloch wave (GKKS also read q as the Chern–Simons quantity of Yang–Mills). When the valleys are deep the lowest level becomes a narrow band, fixed by the action $S_0 = \int_0^{2\pi} \sqrt{2K(1-\cos q)}\,dq = 8\sqrt{K}$ of the path from valley to valley. The band has the form $E_0(0) + \frac{W}{2}(1-\cos\theta)$, and W approaches the asymptotic form of the Mathieu equation (a rewriting of DLMF 28.8.2, $\frac{16}{\sqrt{\pi}} K^{3/4} e^{-S_0}(1 - \frac{7}{64\sqrt{K}})$) (Table 5).

Table 5 The width $W = E_0(\pi) - E_0(0)$ of the lowest band of $-K\cos q$ (bisection with 60-digit decimal arithmetic; the "—" at the right end means the deviation cannot be measured in double precision) Computed

K	W	W / asymptotic form	Deviation from $1-\cos\theta$ / W
1	2.66699×10⁻³	0.98887	2.8×10⁻³
4	2.71024×10⁻⁶	0.99782	1.7×10⁻⁶
16	8.89113×10⁻¹³	0.99951	—

In the dilute gas approximation of Yang–Mills (regarding instantons as scattered far apart), the vacuum energy density is said to be proportional to $-\cos\theta\, e^{-8\pi^2/g^2}$ Physics . It is an approximation that regards typical configurations as superpositions of smooth solutions, and it is mathematically unsettled (A dictionary of scales and limits). More precisely, in four dimensions the coupling runs with the scale (4-06), so the integral over the instanton size ρ diverges on the side of large ρ, and the approximation is said to be uncontrolled even at the level of physics Physics . The quantity that does not depend on the approximation is the topological susceptibility $\chi t = \langle Q^2\rangle/V$ (the second derivative of $E(\theta) = -\frac{1}{V}\ln\sum_Q Z_Q e^{i\theta Q}$ at $\theta = 0$).

For how this book uses this, see 25-02 § 03 (counting the string tension with a dilute gas of fractional instantons).

08

Integers on the lattice, and θ in the problem statement

The space of fields on the lattice (5-01) is connected, and “the concept of a topological sector has therefore no a priori well-defined meaning” (Lüscher 2010) Known . Lattice fields with small action density have an integer Q, which is said to return to the Chern number in the classical continuum limit (that they have an integer is from the abstract of Lüscher 1982; 7-11 § 09). For SU(3), the fraction of fields V smoothed by the Wilson flow that satisfy $\max_p \mathrm{Re}\,\mathrm{tr}\{1 - V(p)\} < 0.067$ is 0%, 8% and 70% at $a = 0.1,\ 0.07,\ 0.05$ fm, and the sector is “a dynamical property of the theory” (same, 2010) Physics . The

construction of Magnen–Rivasseau–Sénéor (1993) for SU(2) with a fixed infrared cutoff (a cutoff on the long-distance side) is restricted to the trivial sector (7-07) Known . In HMC the effective dynamical critical exponent of Q^2 is about 5 (Schaefer–Sommer–Virotta 2011; 5-07) Physics .

Neither the θ term nor instantons appear in the problem statement of Jaffe–Witten (full-text search). GKKS argue, from the 't Hooft anomaly of time reversal and centre symmetry (18-03) (an obstruction to making a trivial vacuum while keeping both symmetries), that “at θ = π the vacuum cannot be a trivial non-degenerate gapped state”, and state that the vacuum at θ = 0, by contrast, is trivial, gapped and non-degenerate (both at the level of physics) Physics . "A unique vacuum with a gap" is a claim that depends on θ, and $-K\cos 2q$ and $\theta = \pi$ in Figure 2 are its quantum-mechanical version.

What this chapter can say

Item	What is said	What is not said
Integrality of Q	Solutions of finite action (Uhlenbeck), and fields assumed to decay to pure gauge far away Known	General fields of finite action without the decay assumption
Classical theory and models	(3)–(10), Tables 1–5, Figures 1 and 2 Known Computed	Reproducing the proofs of Uhlenbeck and ADHM. The model's levels are not those of Yang–Mills
Quantum theory	θ vacua, the dilute gas, the lattice, the statements of GKKS Physics	Not theorems. Nothing is shown about the mass gap or the construction of the theory

Sources and reproduction

Item	Kind	Source / tools
BPST, Uhlenbeck, ADHM, the family (7), tunnelling, θ vacua and CP	Abstract	Belavin–Polyakov–Schwartz–Tyupkin, PLB 59 (1975) 85 / Uhlenbeck, CMP 83 (1982) 11 / Atiyah–Hitchin–Drinfeld–Manin, Phys. Lett. A 65 (1978) 185 / Corrigan–Fairlie, PLB 67 (1977) 69 / Jackiw–Nohl–Rebbi, PRD 15 (1977) 1642 / 't Hooft, PRD 14 (1976) 3432 / Jackiw–Rebbi, PRL 37 (1976) 172 / Callan–Dashen–Gross, PLB 63 (1976) 334
Chern–Simons form	Bibliographic data only	Chern–Simons, Ann. of Math. 99 (1974) 48

Item	Kind	Source / tools
Classification of SO(3) bundles, fractional topological charge (SO(3) bundles in § 03)	Bibliographic data only (content not checked against the text)	A. Dold, H. Whitney, Ann. of Math. 69 (1959) 667 / G. 't Hooft, Nucl. Phys. B 153 (1979) 141 / G. 't Hooft, Commun. Math. Phys. 81 (1981) 267
$-\cos\theta$ of the dilute gas, divergence of the size integral	Not checked	The text of the primary sources not checked
Number of parameters, $\theta = \pi$, the particle on a circle, the lattice	Text	Tong, arXiv:hep-th/0509216 § 1.1.2, § 1.2 (eq. (1.13)) / Gaiotto–Kapustin–Komargodski–Seiberg, JHEP 05 (2017) 091 (abstract, Appendix D.1) / Lüscher, JHEP 08 (2010) 071 § 4
Topological charge on the lattice, freezing, the trivial sector	Abstract	Lüscher, CMP 85 (1982) 39 / Magnen–Rivasseau–Sénéor, CMP 155 (1993) 325 / Schaefer–Sommer–Virotta, NPB 845 (2011) 93
The problem statement, the asymptotic form of Mathieu	Text	Jaffe–Witten, *Quantum Yang–Mills Theory* (Clay) / NIST DLMF (28.8.2)
Tables 1–5, (11), figures	Computed	Python 3.12, NumPy, SciPy, decimal (standard library; the 60-digit arithmetic of Table 5) (`winding.py`, `bpst.py`, `two_instantons.py`, `rotor.py`). The figures are plain JavaScript

Next chapter to read: 3-06 Detour: Donaldson and exotic R^4. The dictionary of topological charge is 18-04.

Detour: Donaldson and exotic R⁴ — the space of solutions of classical Yang–Mills tells smoothness in four dimensions apart

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — What an exotic structure is. Intersection forms and the map of topology (Freedman, Rokhlin). Donaldson's line of argument, which reads the space of solutions of self-dual connections as a cobordism and derives that definite intersection forms can be diagonalised, and why exotic structures on $\mathbb{R}^4$ follow from it.

Prerequisite chapters — 3-05 (the BPST solution, the limit of shrinking to a point). For smooth structures, 1-05; for an entry to intersection forms, 1-12 § 07; for the orbit space and reducible connections, 3-04 § 04. Homology, signature, cobordism, index and the group of conformal transformations are defined minimally where they appear.

Said first — This is a detour, and a prerequisite for no chapter. It follows only the line of known mathematics and does not reproduce the body of the analysis. It shows nothing about quantum Yang–Mills or the mass gap.

Order of this chapter

1. The same shape, yet different — homeomorphism and diffeomorphism
2. The intersection form — a matrix that counts intersection points
3. A map of the topological world — Freedman and Rokhlin, Detour: in five or more dimensions
4. Donaldson's theorem — the counting lemma (Table 1)
5. The moduli space — a five-dimensional space of solutions (Figure 1)
6. Reading the space of solutions as a cobordism — cones, the end and the signature (Figure 2)
7. Exotic R⁴ — a set that no sphere can enclose (Table 2)

8. Two faces of the same equation — classical and quantum

01

The same shape, yet different — homeomorphism and diffeomorphism

Being the same in the topological sense (homeomorphic) means there is a one-to-one correspondence that is continuous in both directions. For smooth manifolds (1-05) we ask for a correspondence that is smooth in both directions (a diffeomorphism).

$$X \underset{\text{diffeomorphic}}{\cong} Y \implies X \underset{\text{homeomorphic}}{\cong} Y, \qquad \text{the converse does not hold in general}$$ (1)

When two are homeomorphic but not diffeomorphic, one is called an "exotic" structure on the other. Structures that differ only in how the atlas is written (the $x \mapsto x^3$ of 1-05) are related by a diffeomorphism, so they do not count. The first example is said to be the seven-dimensional sphere (Milnor 1956) Known .

Example. The smooth structure on $\mathbb{R}^n$ is unique for $n \neq 4$. The introduction of Gompf (1983) writes “it is a standard fact that for n ≠ 4, exotic R^n's cannot exist”, and continues that applying the high-dimensional theory of smoothing, $\mathbb{R}^4$ should have one structure (“there should only be one structure on R4, as it has no cohomology”) Known . In fact there are uncountably many (§ 07, the article on four dimensions). Why is the subject of this chapter.

02

The intersection form — an integer matrix that counts intersection points of surfaces

The closed surfaces inside a closed (compact, without boundary) oriented four-dimensional manifold X fall into classes that do not change under continuous motion, and classes can be added. This collection is $H_2(X;\mathbb{Z})$, and for simply connected X it is $\mathbb{Z}^r$ ($r = b_2$; the Betti number of 1-06 § 08). Precisely, the elements of H_2 are homology classes, and here we use only the intuition (in four dimensions every class is said to be representable by an embedded surface). Two surfaces in general position meet at finitely many points (1-12 § 07), and the sum of ± 1, fixed by orientations, does not change when the surfaces are moved.

$$Q(\Sigma_1, \Sigma_2) = \sum_{p \in \Sigma_1 \cap \Sigma_2} (\pm 1), \qquad Q : \mathbb{Z}^r \times \mathbb{Z}^r \to \mathbb{Z} \ \text{ is symmetric with } \det Q = \pm 1$$ (2)

$\det Q = \pm 1$ (unimodular) is due to Poincaré duality Known . The numbers read from the form are the rank r, the signature $\sigma = b^+ - b^-$ (the number of positive eigenvalues minus the number of negative ones), and the parity (even if every $Q(a,a)$ is even). If $b^- = 0$, it is positive definite.

Example. In $\mathbb{CP}^2$ two lines $\mathbb{CP}^1$ meet at one point, so (1); $\mathbb{CP}^2$ with the orientation reversed gives (-1); and $S^2 \times S^2$ gives $H = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. The connected sum (remove a ball and glue) becomes a direct sum, and $\#r\,\mathbb{CP}^2$ is the identity matrix $\langle 1 \rangle^r$ (the standard form). The form of the Kummer surface (K3) is $E_8 \oplus E_8 \oplus H^{\oplus 3}$ (E_8 is the negative-definite sign; Gompf writes this $H^{\oplus 3}$ as H of rank 6) Known .

The smallest example of a positive-definite even form is E_8. The following Cartan matrix has determinant 1 and is even and positive definite Computed . Gompf states that the only even definite unimodular form of rank 8 is E_8 ("the unique symmetric form over Z which has rank 8, is even, and is negative definite").

$$E_8 = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 2 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 2 \end{pmatrix} \quad (3)$$

03

A map of the topological world — Freedman's theorem and Rokhlin's restriction

Seen topologically only, four dimensions resembles higher dimensions. Freedman (1982) showed that every symmetric form of determinant ± 1 can be realised as the form of a closed simply connected topological manifold, and that the homeomorphism class is fixed by the form and a $\mathbb{Z}/2$ quantity (the Kirby–Siebenmann obstruction; 0 if smooth) (as presented in Donaldson 1983 § I.1) Known . There is also a topological manifold $|E_8|$ whose form is E_8.

One restriction from smoothness was already known. As Donaldson presents it, "Rohlin's theorem asserts that any even form coming from a smooth simply connected 4-manifold has signature divisible by 16" Known .

$$X \text{ smooth and simply connected, and } Q \text{ even} \implies \sigma(X) \equiv 0 \pmod{16} \quad (4)$$

Example. E_8 ($\sigma = 8$) is excluded by (4), and $|E_8|$ has no smooth structure. $E_8 \oplus E_8$ ($\sigma = 16$) passes (4), and (4) says nothing about odd $E_8 \oplus \langle 1 \rangle$. What settled these is the theorem of § 04.

Detour: why smoothness does not matter in five or more dimensions

A subsection you may skip.

In higher dimensions, the key is said to be the operation of removing an extra pair of intersection points of opposite sign by sliding along a disc spanning a loop joining the two points (called the Whitney trick). In five or more dimensions the disc in general position meets nothing else, but in four dimensions, with $2 + 2 = 4$, the disc itself meets surfaces at points (1-12 § 07). Freedman showed that an infinite tower replacing the disc (a Casson handle) is homeomorphic to the open 2-handle $D^2 \times \mathbb{R}^2$. Precisely, a Casson handle is an open four-dimensional manifold made by thickening a disc with intersection points (a kinky handle) and, for each intersection point, gluing on another at the next stage, repeated infinitely. From Donaldson's theorem Gompf writes “they are not all diffeomorphic to it” Known .

04

Donaldson's theorem — definite forms can be diagonalised

Donaldson's diagonalisation theorem (Theorem 1 of Donaldson 1983) is “Let X be a compact smooth simply connected oriented 4-manifold with the property that the associated form Q is positive definite. Then that form is equivalent, over the integers, to the standard diagonal form” Known . That is, a change of integer basis turns it into the identity matrix. Reversing the orientation, the same holds for negative definite forms, and the simply connected assumption is said to have been removed later (Donaldson 1987). The proof uses one property of numbers, a lemma on the number $n(Q)$ of a in $Q(a, a) = 1$, counting $\pm a$ as one pair (Lemma 2 of the paper) Known .

$$n(Q) = \#\{\pm a \in \mathbb{Z}^r : \ Q(a,a) = 1\}, \qquad n(Q) \le r, \qquad \text{equality} \iff Q \cong \langle 1 \rangle^r \ (5)$$

The proof is by induction on the rank. If $Q(a, a) = 1$, every b splits with integer coefficients as $b = m\,a + c$ ($m = Q(b, a)$, $c \perp a$), and $Q(b, b) = m^2 + Q(c, c)$. Since the form is positive definite, $Q(b, b) = 1$ is $b = \pm a$ or "$m = 0$ and $Q(c, c) = 1$", so $n(Q) = n(Q|_{a^\perp}) + 1$, and the rank also drops by 1. When the a run out, the form is standard if the remaining rank is 0, and otherwise $n(Q) < r$.

Example. Table 1 gives $n(Q)$ counted by enumeration Computed . D_n^+ is the union of the set D_n of all integer vectors with even sum of components and that set shifted by $(\frac{1}{2},$

$\dots, \frac{1}{2}$) (an integer form of determinant 1 when n is a multiple of 4). Even forms that pass (4) are excluded by the theorem if $n(Q) < r$.

Table 1 $n(Q)$ of positive-definite unimodular forms (determinant 1; by enumeration) Computed

Form	r	Parity	$n(Q)$	Norm 2	(4)	Theorem
$\langle 1\rangle^8$ ($\#8\,\mathbb{CP}^2$)	8	Odd	8	112	—	Standard form
E_8	8	Even	0	240	excluded	excluded
$E_8 \oplus \langle 1\rangle$	9	Odd	1	240	—	excluded
D_{12}^+	12	Odd	0	264	—	excluded
$E_8 \oplus E_8$, D_{16}^+	16	Even	0	480	passes	excluded

05

The tool: the moduli space of self-dual connections

Conventions. We use the same geometric notation as 3-05 (anti-Hermitian A, $c_2 = \frac{1}{8\pi^2}\int \operatorname{tr} F \wedge F$). Donaldson (1983) also uses the same sign, with an SU(2) bundle of $c_2(P) = -1$ over a positive-definite X, and self-dual connections Known . Self-dual means $F^- = 0$, where F^- is the anti-self-dual part of F (1-12 § 03; the same condition as the $F = *F$ of 3-03). Many textbooks reverse the orientation and write in terms of negative definite forms and anti-self-duality ($c_2 = k > 0$) (the detour of 3-05, 1-12 § 07). Changing the orientation, it is the same story. The collection of solutions with those related by gauge transformations identified is the moduli space, the part of the orbit space of 3-04 made of solutions.

$$\mathcal{M}_k = \{A : F_A^- = 0\}/\mathcal{G} \subset \mathcal{A}/\mathcal{G}, \qquad k = -c_2(P), \qquad \dim \mathcal{M}_k = 8k - 3(1 - b_1 + b^-) \quad (6)$$

k is the Q of 3-05. b_1 is the number of one-dimensional holes (the Betti number of 1-06 § 08; 0 if simply connected). The dimension comes from the **index** of the linearised equation. The index is an estimate of the dimension of the solution space of a linear equation; for a finite system of linear equations it corresponds to "the number of unknowns − the number of equations" (precisely, the dimension of the kernel minus the dimension of the cokernel; here we use only the name). Donaldson (1983) § II writes this index as $8|c_2(P)| - \frac{3}{2}(\chi(X) - \tau(X))$ Known . τ is the signature σ and $\chi = b_0 - b_1 + b_2 - b_3 + b_4$ is the Euler number (a relative of the $V - E + F$ of 1-06); in closed four dimensions $b_0 = b_4 = 1$, $b_3 = b_1$, $b_2 = b^+ + b^-$, so $\frac{3}{2}(\chi - \tau) = 3(1 - b_1 + b^-)$, that is, (6). In the anti-self-dual convention b^- is replaced by b^+.

For S^4 ($b_1 = b^- = 0$), $8k - 3$. Removing from the $8k$ parameters of 3-05 § 04 the 3 constant gauge rotations (identified because they are related by gauge

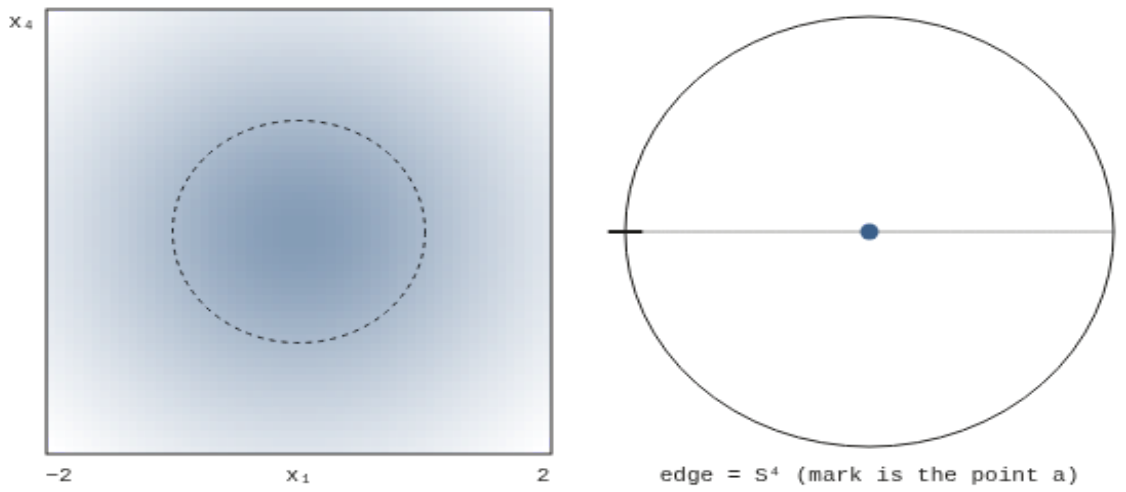
transformations) gives $8k-3$, which for $k=1$ is 5: centre 4 + size 1. For the X of the theorem ($b1 = b^- = 0$, $k=1$) it is also 5.

Example. For $k=1$ on S^4, Donaldson writes $\mathcal{M}(S^4) = SO(5,1)^+/SO(5)$ as the open five-dimensional ball B^5. $SO(5,1)^+$ is the group of conformal transformations (angle-preserving transformations) of S^4. Donaldson explains that a solution is fixed by a "centre" and a "size", and states for the closed ball $\bar{B}^5 = \mathcal{M}(S^4) \cup S^4$ that “a point x of S4 represents the limit point of a sequence of connections whose curvature becomes concentrated in diminishing balls about x” Known . The limit of "shrinking to a point" in 3-05 § 05 is the edge. Figure 1 reads the centre a (on the x_1 axis) and the size λ as a point $w = a + i\lambda$ of the upper half-plane, and is the slice mapped by the linear fractional transformation $z = (w-i)/(w+i)$ that maps the upper half-plane to the unit disc (coordinates for the figure).

The hyperbolic distance in Figure 1

The hyperbolic distance in the figure's readout is the shortest length joining the points (a,λ) and $(0,1)$ when the length of curves in the upper half-plane $\lambda > 0$ is measured as $\int \sqrt{da^2+d\lambda^2}/\lambda$. This metric $(da^2+d\lambda^2)/\lambda^2$ (a slice of the metric of upper half-space) is invariant under the group of conformal transformations $SO(5,1)^+$, but it is not the L^2 metric that enters naturally on the space of solutions (the one measuring the size of a change of connection δA by $\int |\delta A|^2$). The hyperbolic distance of $a=0$ is $|\ln\lambda|$ ($\ln(1/\lambda)$ if $\lambda \le 1$) Computed . In this metric the edge is infinitely far away. However small the size λ is made, it remains a solution, and the family does not break off before the edge because in four dimensions the action does not depend on the scale (1-12 § 04).

Figure 1 Moving the size λ (log) and the centre a with the knobs moves the action density on the left (cross-section $x_2 = x_3 = 0$, log shading, half the action inside the dotted circle) and the point in the disc on the right (a two-dimensional slice of B^5). As $\lambda \to 0$ the point goes to the position a on the edge (S^4)



λ = 1.00, a = 0.00, inside radius 0.1 = 0.0003, 1 − |z| = 1.00, hyperbolic distance = 0.000

Computed The numbers in this figure were computed on this machine. The density $48\lambda^4/(|x-a|^2+\lambda^2)^4$ ((9) of 1-12) and z are computed on the spot. Static values: shrinking as $\lambda = 1,\ 0.1,\ 0.01$ at $a = 0$ gives $1 - |z| = 1,\ 0.182,\ 0.0198$, hyperbolic distance $0,\ 2.30259,\ 4.60517$, fraction of the action inside radius 0.1 $0.0003,\ 0.5000,\ 0.9997$ (the closed formula and numerical integration agree).

06

Reading the space of solutions as a cobordism — cones, the end and the signature

The $\mathcal{M}$ of (6) consists of three parts (Donaldson 1983 § I.2, § II, § III) Known .

(i) Irreducible points. Connections whose holonomies (1-07) have no common invariant subspace; a small perturbation of the equation makes them an oriented smooth five-dimensional manifold.

(ii) Reducible points. The case where the bundle splits as $E = L \oplus L^{-1}$ and the connection preserves the splitting; from $c_2(E) = -c_1(L)^2$, $Q(\alpha,\alpha) = 1$ ($\alpha = c_1(L)$). Since $b^- = 0$, there is a self-dual connection for each $\pm\alpha$; since $b_1 = 0$, it is unique up to gauge equivalence, and their number is $n(Q)$. The group of gauge transformations that leaves the connection fixed (the group that fixes) is as large as U(1), so these are the "corners" of 3-04 § 04. Precisely, at points where the group that fixes is large, the orbit space stops being a smooth manifold and the neighbourhood takes the shape of a cone. The neighbourhood is $\mathbb{C}^3/S^1$: dividing the unit sphere S^5 of $\mathbb{C}^3$ by the circle S^1 gives $\mathbb{CP}^2$, and adding the radial direction gives a cone over $\mathbb{CP}^2$.

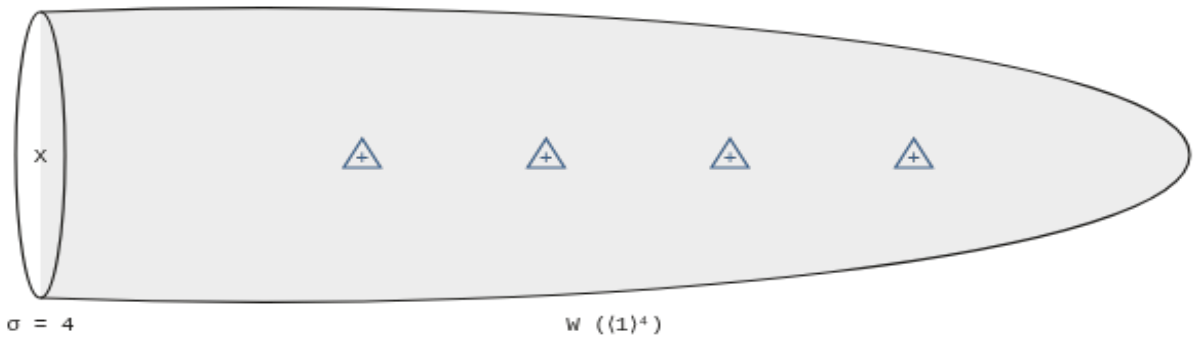
(iii) The end. Solutions of small size look like BPST solutions concentrated at one point of X. Taubes (1982) constructed them under $b^- = 0$. Donaldson § III uses the convergence results of Uhlenbeck (1982) to show that sequences that escape are only those that concentrate. Theorem 11 states that a neighbourhood of the end is diffeomorphic to $X \times (0, \lambda_0)$. The edge of Figure 1 is here a copy of X.

Cutting out the neighbourhoods of the cone vertices and gluing X onto the end gives an oriented compact five-dimensional manifold W whose boundary is X and $n(Q)$ copies of $\pm\mathbb{CP}^2$ (the relation of sharing a boundary is called cobordism). The signature is “an invariant of oriented cobordism”. That it is a cobordism invariant means that the signature of the whole boundary of an oriented compact five-dimensional manifold (here X and the sum of $n(Q)$ copies of $\pm\mathbb{CP}^2$, with orientations) is 0; from this follows $\sigma(X) = \sum \varepsilon_i\, \sigma(\mathbb{CP}^2)$ ($\varepsilon_i = \pm 1$ is how the orientations enter), and the following can be said.

$$r = \sigma(X) = \sum_{i=1}^{n(Q)} \varepsilon_i\, \sigma(\mathbb{CP}^2) \ \le\ n(Q)\ \le\ r \qquad (\varepsilon_i = \pm 1,\ \sigma(\mathbb{CP}^2) = 1) \tag{7}$$

The reasons for the four relations are as follows. The first equality: since the form is positive definite, $b^- = 0$, so $\sigma = b^+ = r$. The second equality: cobordism invariance (the paragraph above). The first inequality: each $\varepsilon_i \le 1$, so the sum is at most the number of terms $n(Q)$. The second inequality: (5). Both ends are r, so everything in between is equality, $n(Q) = r$, and the equality case of (5) gives the standard form. Example. For $E_8 \oplus E_8$ there are no cones and the boundary is only X, so $\sigma = 16$ cannot be 0, a contradiction. Figure 2 is the balance for each form. Most of the paper is the analysis of (i)–(iii).

Figure 2 Choosing a form with the buttons changes the schematic of W assuming there were a smooth X with that form (the left edge is X, triangles are cones) and the balance of (7). If there are too few cones, a contradiction (dashed line)



Q = ⟨1⟩⁴ r = σ(X) = 4, number of cones n(Q) = 4 (7) : 4 ≤ 4 holds → standard form (4) : odd, so it says nothing

Computed The numbers in this figure were computed on this machine. The rank and $n(Q)$ are the enumerated values of Table 1 ($\langle 1 \rangle^4$ is $n(Q) = 4$). Without JavaScript, putting r and $n(Q)$ from Table 1 into (7) gives the same verdict.

07

Exotic R⁴ — a set that no smooth sphere can enclose

The introduction of Taubes (1987) writes “Donaldson's theorem … has the surprising corollary that there exists an exotic smooth structure on R4. This corollary was deduced by M. Freedman” Known . The line has three steps (as presented by Taubes). (a) $|E_8 \oplus E_8|$ cannot be smoothed (§ 04). (b) With one point removed it can be smoothed (Freedman), and the neighbourhood of the removed point is diffeomorphic, outside a compact set, to the infinity of some smooth R homeomorphic to $\mathbb{R}^4$. (c) If R were standard, one could cut along a large smooth S^3 and cap it with a ball, and $|E_8 \oplus E_8|$ would become smooth. So the following can be said.

$$R \underset{\text{homeomorphic}}{\cong} \mathbb{R}^4, \qquad R \underset{\text{diffeomorphic}}{\not\cong} \mathbb{R}^4 \tag{8}$$

Gompf (1983) made a R^4_{DF} of the same kind from K3 and $\#3\, S^2 \times S^2$, and writes “there is a compact set in R4DF … which cannot be enclosed by any smooth S3. This

behavior is impossible in the standard R4." Known .

Example. Gompf § 2 uses $\mathbb{CP}^2\#9\mathbb{CP}^2$ (form $\langle 1\rangle \oplus 9\langle -1\rangle$, basis $e_0, \ldots, e_9$) instead of the Kummer surface to make another exotic $\mathbb{R}^4$. The key is $x = 3e_0 + e_1 + \cdots + e_8$. $x \cdot x = 9 - 8 = 1$, and the orthogonal complement $E_8 \oplus \langle -1\rangle$ (negative definite and not diagonalisable) contradicts the theorem, so cutting off by surgery (cutting out part of a manifold and gluing it back differently) the part of form $\langle 1\rangle$ spanned by x is possible topologically but cannot be done smoothly. Table 2 is the check Computed .

Table 2 The orthogonal complement of x in Gompf § 2 (inside $e_0, \ldots, e_8$) Computed

Basis	$e_i - e_{i+1}$ $(i = 1, \ldots, 7)$ and $e_0 + e_1 + e_2 + e_3$ (norm -2)
Gram matrix	-1 times (3) (agreeing after changing the sign of the last basis vector). Determinant 1, even, 8 negative eigenvalues
Number of vectors of norm -2, -4	240, 2160 (the same as E_8)

How many are there? Taubes showed, by gauge theory on manifolds with periodic ends, that the R_r obtained as images of balls of radius r under the homeomorphism of (b) are not diffeomorphic to one another for distinct r larger than some r_0. Theorem 1.1 is "There exists an uncountable family of diffeomorphism classes of oriented 4-manifolds which are homeomorphic to R4." Known . r runs over the interval $r > r_0$, and distinct r give distinct classes, so there are at least as many as there are real numbers (the classes of manifolds have at most the cardinality of the reals, so it is said to be exactly the cardinality of the continuum). On the other hand, whether S^4 has an exotic structure is said to be unknown.

08

Two faces of the same equation — a classical tool and a quantum problem

Although the equation is the same as in 3-05, Donaldson uses the shape of the collection of solutions as a tool, and uses neither probability nor time evolution. What the problem statement (0-03) asks about is the probability measure weighted by the same action (4-02) and the spectrum of its Hamiltonian.

Classical: $\mathcal{M} = \{F_A = *F_A\}/\mathcal{G}$ — study its shape, Quantum: $d\mu(A) = \frac{1}{Z} e^{-SE[A]} \mathcal{D}$

On the quantum measure, smooth classical solutions are expected not to be typical configurations (by analogy with the Cameron–Martin space of 1-09) Physics . The "four" that matters in the problem is the four at which the dimension of the coupling constant vanishes (0-04). Example. § 2 of the problem statement says "Classical properties of gauge theory are within the reach of established mathematical methods", and after describing the role in the study of four-dimensional manifolds, continues

“one does not yet have a mathematically complete example of a quantum gauge theory in four-dimensional space-time” Known . This was written in 2000; the committee's assessment is that the construction published in 2026-11 built a quantum Yang–Mills theory in four dimensions (6-00).

There are two bridges at the level of physics. The abstract of Witten (1994) writes, on Donaldson's invariants, “instead of defining four-manifold invariants by counting SU(2) instantons, one can define equivalent four-manifold invariants by counting solutions of a non-linear equation with an abelian gauge group” (background in 17-11) Physics . § 5 of the problem statement says that the clustering that follows from the mass gap “may make it possible to apply mathematical results established on R4 to any 4-manifold”, and concludes that the gap “may also be important in mathematical applications of four-dimensional quantum gauge theories to geometry” Physics .

What this chapter can say

Item	What is said	What is not said
Intersection forms	Definitions and examples, the proof of (5), Tables 1 and 2 Known Computed	Proofs of Poincaré duality and of the theorems of Rokhlin and Freedman
Donaldson's theorem	The statement and the line of (6)(7), Figures 1 and 2 Known Computed	The analysis of perturbation, orientation, convergence and the end. The hyperbolic distance of Figure 1 is that of a metric invariant under conformal transformations, not the L^2 metric on the space of solutions. Figure 2 is a schematic
Exotic $\mathbb{R}^4$	The line of (8), the statements of Gompf and Taubes Known	Reproducing the proofs of Freedman and Taubes
Relation to physics	The words of the problem statement and of Witten's abstract Physics	Nothing is shown about the construction of quantum Yang–Mills or the mass gap

Sources and reproduction

Item	Kind	Source / tools
Quotations and line in § 03– § 06, the dimension in (6) (the index formula of § II)	Text (the relevant parts of § I– § III)	S. K. Donaldson, “An application of gauge theory to four-dimensional topology”, J. Differential Geom. 18 (1983) 279–315

Item	Kind	Source / tools
Quotations and examples in § 01– § 03 and § 07	Text (§ 0– § 2)	R. E. Gompf, “Three exotic R^4's and other anomalies”, J. Differential Geom. 18 (1983) 317–328
(a)–(c), R_r, uncountably many	Text (§ 1)	C. H. Taubes, “Gauge theory on asymptotically periodic 4-manifolds”, J. Differential Geom. 25 (1987) 363–430
§ 08	Text (§ 2, § 5)	A. Jaffe, E. Witten, Quantum Yang–Mills Theory, Clay Mathematics Institute
§ 08	Abstract	E. Witten, Math. Res. Lett. 1 (1994) 769–796, arXiv:hep-th/9411102
Original papers (via the texts above)	Bibliographic data only	Freedman, JDG 17 (1982) 357 / Taubes, JDG 17 (1982) 139 / Uhlenbeck, CMP 83 (1982) 11, 31 / Atiyah–Hitchin–Singer, Proc. R. Soc. A 362 (1978) 425 / Donaldson, JDG 26 (1987) 397 / Milnor, Ann. of Math. 64 (1956) 399
Milnor's (1956) example, the question of S^4, the version without the simply connected assumption (1987)	Secondary sources	Wikipedia “Exotic sphere”“Exotic R4”“Donaldson's theorem”
The Whitney trick, the construction of Casson handles, representing classes by surfaces, proofs of duality and of cobordism invariance, the upper bound on the cardinality of classes of manifolds, the analogy of § 08	Not checked (common account)	Text not checked
Properties of (3), Table 1, Table 2, Figure 1, Figure 2	Computed	Python 3.12, NumPy 1.26, SciPy 1.11 (lattices.py, gompf_check.py, moduli_disc.py). The formulas of the figures are checked with node

Next chapter to read: 3-07 Yang–Mills inside the Standard Model. For the physics of Seiberg–Witten, 17-11; for the "four" that matters in the problem, 0-04.

Yang–Mills inside the Standard Model — a product of three groups, with matter and the Higgs added

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — The gauge part of the Standard Model is the sum of three Yang–Mills fields, for $SU(3)\times SU(2)\times U(1)$. The charges satisfy an arithmetic called anomaly cancellation, and the value of the Higgs field gives mass to the W and Z while leaving the photon massless. We also sort out why the problem statement takes out pure Yang–Mills alone.

Prerequisite chapters — 3-02 (also using 3-01 and 1-04). Spin-1/2 fields are defined minimally in § 02.

Said first — This is a classical chapter. § 03 checks only the arithmetic of conditions from quantum theory, and the numbers in § 06 are estimates at the lowest order of perturbation theory (tree level). This chapter shows nothing about the mass gap.

Order of this chapter

1. Three groups — twelve gauge fields
2. Matter fields — one generation and the charges (Table 1)
3. Counting gives zero — anomalies (Table 2). Detour: SU(5)
4. The SU(3) of colour — quarks and colour singlets
5. The Higgs mechanism — an abelian model (Figure 1)
6. The electroweak $SU(2)\times U(1)$ — W, Z and the photon (Figure 2, Table 3)
7. Where mass comes from — the proton and the pion
8. The difference from pure Yang–Mills (Table 4)

01

Three groups — twelve gauge fields

The conventions are the same as in 3-02 § 02 ($\hbar = c = 1$, the metric $(-,+,+,+)$, Hermitian generators, $D_\mu = \partial_\mu - igA_\mu$).

The Standard Model is a gauge theory with three groups side by side, and the PDG review defines it as "the renormalizable field theory with gauge group $G_{\rm SM} = \mathrm{SU}(3)_C \times \mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$" Known ($C$ is colour, L is left-handed, Y is hypercharge). Each factor carries a Yang–Mills field, and the number of fields is $8 + 3 + 1 = 12$: the gluons G^a_μ, W^i_μ and B_μ.

$$\mathcal{L}_{\rm gauge} = -\tfrac{1}{4} G^a_{\mu\nu} G^{a\,\mu\nu} - \tfrac{1}{4} W^i_{\mu\nu} W^{i\,\mu\nu} - \tfrac{1}{4} B_{\mu\nu} B^{\mu\nu} \quad (1)$$

The field strengths of G and W have commutator terms, in the form of (7) of 3-02. $B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$, like the electromagnetic field, has no self-interaction. The action on matter is collected into one covariant derivative.

$$D_\mu = \partial_\mu - i g_s \, G^a_\mu T^a - i g \, W^i_\mu \, \frac{\sigma^i}{2} - i g' \, Y \, B_\mu \quad (2)$$

T^a acts only on colour triplets and $\sigma^i/2$ (1-04) only on doublets; on singlets the term is dropped. Since the group is not simple, there is one coupling constant for each factor, three in all (1-03). Example. $g_s \approx 1.22$ is obtained from the strength of the strong coupling $\alpha_s = g_s^2/4\pi$ measured at the scale of the Z mass, $\alpha_s(m_Z) = 0.1180$ (the MS convention; the note to Table 3 in § 06). $g \approx 0.65$ and $g' \approx 0.35$ are tree-level estimates worked back from the W and Z masses in Table 3 of § 06 Computed .

02

Matter fields — one generation and $Q = T_3 + Y$

Quarks and leptons are spin-1/2 fields. We do not use what is inside the components (the Dirac matrices), only how they turn under the groups. What is needed is that such a field splits into a **left-handed** part ψ_L and a **right-handed** part ψ_R (for zero mass, the way the spin turns relative to the direction of motion), and that SU(2) acts only on the left-handed part Known . Spatial reflection $\mathbf{x} \mapsto -\mathbf{x}$ (parity) exchanges left-handed and right-handed, so a force acting only on the left-handed part is not invariant under this reflection. This is called **parity violation**.

Table 1 The fields of one generation ($(\mathbf{3},\mathbf{2})$ is a colour triplet and a doublet; for T_3, the upper component of the doublet is $+\frac{1}{2}$ and the lower is $-\frac{1}{2}$)

Field	Representation	Y	$Q = T_3 + Y$	Components

Field	Representation	Y	$Q = T_3 + Y$	Components
$Q_L = (u_L, d_L)$	$(\mathbf{3}, \mathbf{2})$	1/6	2/3, −1/3	6
u_R	$(\mathbf{3}, \mathbf{1})$	2/3	2/3	3
d_R	$(\mathbf{3}, \mathbf{1})$	−1/3	−1/3	3
$L_L = (\nu_L, e_L)$	$(\mathbf{1}, \mathbf{2})$	−1/2	0, −1	2
e_R	$(\mathbf{1}, \mathbf{1})$	−1	−1	1
Higgs Φ	$(\mathbf{1}, \mathbf{2})$	1/2	1, 0	2 (complex)

Computed Q was computed as exact fractions. Known (93.1) of the PDG review of grand unification writes the right-handed fields as their conjugates u^c, d^c, e^c (§03; colour becomes $\bar{\mathbf{3}}$ and the sign of Y flips), and attaches Y as a subscript, normalised to twice its value. For example, the u_R of Table 1 ($(\mathbf{3}, \mathbf{1})$, $Y = 2/3$) is the $(\bar{\mathbf{3}}, \mathbf{1})_{-4/3}$ of u^c. The Higgs is $(\mathbf{1}, \mathbf{2})_1$ in the PDG and corresponds to the $Y = 1/2$ of Table 1. Converted, they agree with Table 1.

$$u_L:\ \tfrac{1}{2} + \tfrac{1}{6} = \tfrac{2}{3}, \qquad e_L:\ -\tfrac{1}{2} - \tfrac{1}{2} = -1, \qquad \nu_L:\ \tfrac{1}{2} - \tfrac{1}{2} = 0 \ \text{(3)}$$

One generation has 15 components counting colour, and there are three generations (PDG) Known . The gauge principle does not decide which field goes into which representation (3-01), and the PDG too states that the “apparently rather complex group-theoretic data” of Table 1 are not explained.

03

Counting gives zero — anomaly cancellation

First we prepare one word. Taking the complex conjugate of each component of a representation $\psi \mapsto \Omega\psi$ gives $\psi^* \mapsto \Omega^*\psi^*$, which is again a representation; we call it the **conjugate representation**. The conjugate of the triplet is written $\bar{\mathbf{3}}$ (the antitriplet), and in SU(3) it is not equivalent to $\mathbf{3}$. In U(1), $e^{iY\alpha}$ becomes $e^{-iY\alpha}$, so the sign of Y flips Known . For the SU(2) doublet $\sigma^2\Omega^*\sigma^2 = \Omega$ (σ^2 is the second Pauli matrix, not a square), so the conjugate is a representation equivalent to the original and is not distinguished Computed .

The Y of Table 1 is strongly constrained. Even when the classical action is gauge invariant, invariance under transformations that turn left and right separately can break at the stage of quantisation; this is called a **gauge anomaly**. The abstract of Gross–Jackiw (1972) states that it “may destroy this gauge invariance in perturbation theory” Known . We do not treat the derivation, and use only the fact that the condition becomes arithmetic. Taking the conjugate of a spin-1/2 field turns right-handed into left-handed, so aligning the right-handed fields to left-handed by conjugation (the sign of Y flips, and u^c becomes $-2/3$), the condition is that all the following sums are 0. The brackets on the right show which combination of groups each anomaly condition belongs to.

$$\begin{aligned} \textstyle\sum_{\text{colour triplets and antitriplets}} Y &= 0 \quad (\mathrm{SU}(3)^2\,\mathrm{U}(1)) \\ \textstyle\sum_{\text{doublets}} Y &= 0 \quad (\mathrm{SU}(2)^2\,\mathrm{U}(1)) \\ \textstyle\sum Y^3 &= 0 \quad (\mathrm{U}(1)^3) \\ \textstyle\sum Y &= 0 \quad (\text{gravity}^2\,\mathrm{U}(1)) \\ \#\mathbf{3} - \#\bar{\mathbf{3}} &= 0 \quad (\mathrm{SU}(3)^3) \end{aligned} \tag{4}$$

The sums are weighted by the dimension of the representations of the groups not in the combination (in the sum over doublets Q_L is multiplied by the number of colours, 3; in the sum over triplets and antitriplets Q_L is multiplied by 2). $\sum Y^3$ and $\sum Y$ run over the 15 components counting colour and doublet components, and the latter is the condition for the combination with gravity, $\text{gravity}^2\,\mathrm{U}(1)$. The abstract of Bouchiat, Iliopoulos and Meyer (1972) states that in models with leptons and quarks "the troublesome anomalies cancel" Known . Example. Fixing the charges and changing only the number of colours N_c gives Table 2.

Table 2 The sums of (4) for one generation (charges fixed, only N_c changed)

N_c	$\mathrm{SU}(N_c)^2\mathrm{U}(1)$	$\mathrm{SU}(2)^2\mathrm{U}(1)$	$\mathrm{U}(1)^3$	$\text{gravity}^2\,\mathrm{U}(1)$
1	0	−1/3	1/2	0
2	0	−1/6	1/4	0
3	**0**	**0**	**0**	**0**

Computed Added as exact fractions. For general N_c, the $\mathrm{SU}(2)^2\mathrm{U}(1)$ column is $N_c/6 - 1/2$ and the $\mathrm{U}(1)^3$ column is $(3 - N_c)/4$. The $\mathrm{SU}(3)^3$ condition has no column in the table; it is $\#\mathbf{3} - \#\bar{\mathbf{3}} = 2 - 2 = 0$ for $N_c = 3$ (the two Q_L and the two u^c, d^c). Known SU(2) has another condition too: the abstract of Witten (1982) states that an SU(2) gauge theory with an odd number of left-handed doublets (and no other representations) is mathematically inconsistent. One generation has $3 + 1 = 4$ doublets, an even number.

For $\mathrm{U}(1)^3$, the quarks contribute $-3/4$ and the leptons $+3/4$; they cancel only when a whole generation is present, and only when there are 3 colours Computed . In pure Yang–Mills, with no spin-1/2 fields, this condition does not arise.

Detour: it fits into SU(5)

A subsection you may skip.

Georgi–Glashow (1974) "conjectured" that the three forces arise from a single interaction of SU(5) (one of the groups of the problem statement, 0-03). Placing SU(3) at the top left of 5×5 and SU(2) at the bottom right leaves a traceless $Y = \mathrm{diag}(-\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}, \frac{1}{2}, \frac{1}{2})$ that commutes with both. One generation splits exactly into $\bar{\mathbf{5}}$ (d^c, L) and $\mathbf{10}$ (Q, u^c, e^c). The $T\mathbf{3}$ of $Q = T\mathbf{3} + Y$ is also a traceless element of SU(5), so the trace of Q is 0 too. Written with $\bar{\mathbf{5}}$ it is $3Q(d^c) + Q(\nu) + Q(e) = 0$, and from $Q(\nu) = 0$ and $Q(e) = -1$ follows $Q(d^c) = \frac{1}{3}$, that is, $|Q(d)| = |Q(e)|/3$. That the

charge of the down quark is exactly a third of the electron's follows from the shape of the group Computed . Grand unification is a hypothesis.

04

The SU(3) of colour — quarks and colour singlets

The SU(3) part is quantum chromodynamics (QCD). The PDG calls QCD “the SU(3) component of the SU(3)×SU(2)×U(1) Standard Model”, and writes that quarks come in three colours in the fundamental representation and gluons are of 8 kinds in the adjoint representation Known .

$$\mathcal{L}_{\mathrm{QCD}} = \sum_q \bar{\psi}_q \left(i\gamma^\mu D_\mu - m_q \right) \psi_q - \tfrac{1}{4} G^a_{\mu\nu} G^{a\,\mu\nu}, \qquad D_\mu = \partial_\mu - i g_s G^a_\mu T^a \tag{5}$$

The sum runs over the 6 kinds of quark (u, d, s, c, b, t), γ^μ is a 4×4 matrix acting on spin, and $\bar{\psi} = \psi^\dagger \gamma^0$. It is the form of (9.1) of the PDG with the sign of the covariant derivative matched to 3-02, and the last term is pure Yang–Mills. The γ^μ term is left in the PDG's convention (metric $(+,-,-,-)$), and its content is not used in this chapter. The mass term $m_q \bar{\psi}\psi$ multiplies left and right; colour turns left and right the same way, so it can be written, but SU(2) turns only the left-handed part, so $\bar{u}_L u_R$ is not invariant, and in the Standard Model mass enters through the Higgs field (§ 07) Known . Colour singlets can be made from two properties of SU(3).

$$\bar{\psi}_a \psi^a \quad (\Omega^\dagger \Omega = 1), \qquad \varepsilon_{abc}\, \psi^a \psi^b \psi^c \quad (\det \Omega = 1) \tag{6}$$

The former is the form of mesons, the latter of baryons (proton, neutron). Example. Under $\Omega = \mathrm{diag}(e^{i\alpha}, e^{i\beta}, e^{-i(\alpha+\beta)})$ the phase of $\psi^1 \psi^2 \psi^3$ does not change. The elements of the centre of 1-04, $\mathbb{Z}_3$, also give 1 when taken three times. Jaffe–Witten state confinement as the physical particle states being “SU(3)-invariant” Known (for the mechanism, 17-02).

05

The Higgs mechanism — mass grows in an abelian model

A mass term breaks gauge invariance (3-02 § 08), yet the W and Z are heavy particles. What bridged this gap was the idea of placing the minimum of the potential of a charged scalar field away from $\phi = 0$. We look at it in a U(1) model.

$$\mathcal{L} = -\tfrac{1}{4} F_{\mu\nu} F^{\mu\nu} - (D_\mu \phi)^* (D^\mu \phi) - \mu^2 |\phi|^2 - \lambda |\phi|^4, \qquad D_\mu = \partial_\mu - i q A_\mu \tag{7}$$

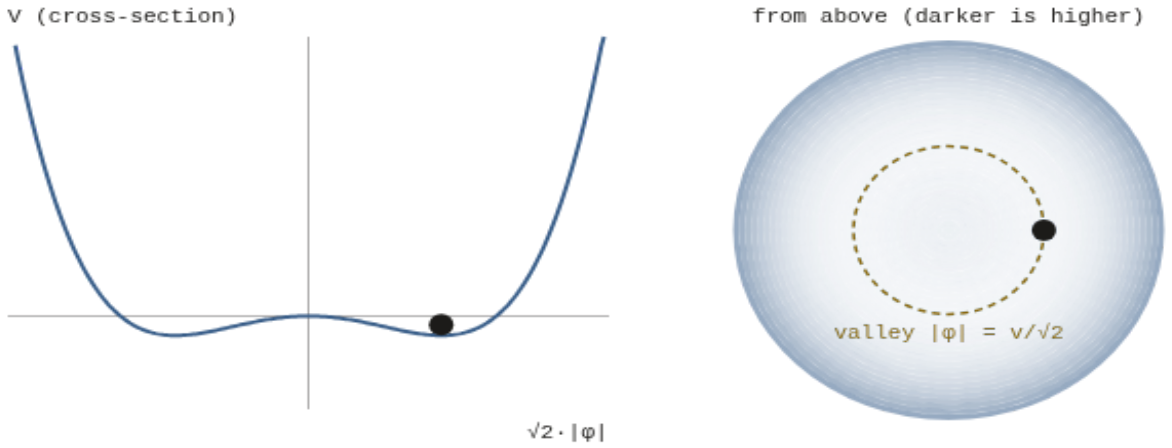
Let $\lambda > 0$. If $\mu^2 < 0$, the minima line up on the circle $|\phi| = v/\sqrt{2}$ ($v = \sqrt{-\mu^2/\lambda}$), and the gauge transformation $\phi \mapsto e^{i\alpha(x)}\phi$ moves along this circle. Turning the phase at each point and writing $\phi = (v + h)/\sqrt{2}$ (real),

$$-(D_\mu\phi)^*(D^\mu\phi) = -\tfrac{1}{2}\,\partial_\mu h\,\partial^\mu h - \tfrac{1}{2}\,q^2(v+h)^2 A_\mu A^\mu \quad (8)$$

follows. The $h = 0$ part is exactly the mass term forbidden in (14) of 3-02, and the masses are as follows Known . The action stays invariant, and the solution (the position of the minimum) brings in the mass.

$$m_A = q\,v, \qquad m_h = \sqrt{2\lambda}\,v = \sqrt{-2\mu^2} \quad (9)$$

Figure 1 Changing μ^2 with the knob changes the potential (left: cross-section; right: from above) and v, m_A, m_h. "Play" shows motion along the valley (the direction removable by a gauge transformation) and the radial oscillation h (on devices set to reduce motion it advances one frame at a time). $\lambda = 1$, $q = 0.6$.



The five points of Figure 1 ($\lambda = 1$, $q = 0.6$)

μ^2	v	$m_A = qv$	Scalar mass	Minimum of V
+0.5	0	0	0.707, 0.707	0
0	0	0	0, 0	0
−0.25	0.500	0.300	$m_h = 0.707$	−0.0156
−0.5	0.707	0.424	$m_h = 1.000$	−0.0625
−1	1.000	0.600	$m_h = 1.414$	−0.2500

Computed The numbers in this figure and table were computed on this machine (by (9)). In the $\mu^2 \geq 0$ rows the scalar mass is that of the two real components of ϕ (both $\sqrt{\mu^2}$); in the $\mu^2 < 0$ rows it is that of the radial h.

Degrees of freedom and the Goldstone boson

Example. We count the degrees of freedom. A massless vector field has 2 polarisations, because of its four components, one is removed by gauge transformations and one by the constraint (2-03 § 04). With a mass term the one removable by gauge

transformations is gone and only the constraint remains, so $4-1=3$ Known . Then before, there are the massless A (2) and ϕ (2 real components); after, the massive A (3) and h (1); both make 4. The phase that disappeared has become the longitudinal oscillation of A.

A massless excitation that appears when a continuous symmetry is spontaneously broken is called a **Goldstone boson** Known . The motion along the valley in Figure 1 corresponds to it, and when there is a gauge field, this direction can be removed by a gauge transformation. Higgs (1964) wrote "the longitudinal degrees of freedom of these particles (which would be absent if their mass were zero) go over into the Goldstone bosons when the coupling tends to zero", and called it the "relativistic analog" of Anderson's (1963) remark on the plasmon in superconductors. In the same year Englert–Brout and Guralnik–Hagen–Kibble also published papers Known . Higgs also cautioned that "nothing is proved about the quantized theory".

"Spontaneous symmetry breaking" is a way of speaking in a gauge-fixed description. Elitzur (1975) argued that in lattice gauge theory without gauge fixing, local symmetries are not spontaneously broken Known (3-01 § 07).

06

The electroweak SU(2)×U(1) — W, Z and the photon

In the Standard Model the same thing is done with SU(2)×U(1) and the Higgs doublet Φ of Table 1. What leaves one of the minima of $V=\lambda(\Phi^\dagger\Phi-v^2/2)^2$, $\Phi_0=(0,\,v/\sqrt{2})$, unchanged is the U(1) of the charge Q, which is $T_3+Y=-\tfrac12+\tfrac12=0$ on the lower component. In Jaffe–Witten's words, the structure group is reduced "to a U(1) subgroup (diagonally embedded in SU(2) × U(1))" Known . Letting (2) act on Φ_0,

$$-(D_\mu\Phi_0)^\dagger(D^\mu\Phi_0)=-\frac{v^2}{8}\Big[g^2\big(W^1_\mu W^{1\mu}+W^2_\mu W^{2\mu}\big)+\big(gW^3_\mu-g'B_\mu\big)\big(gW^{3\mu}-g'B^\mu\big)\Big] \quad (1$$

follows. W^3 and B appear only in the combination $gW^3-g'B$, so we rotate the plane (W^3,B) by the **weak mixing angle** θ_W ($\tan\theta_W=g'/g$).

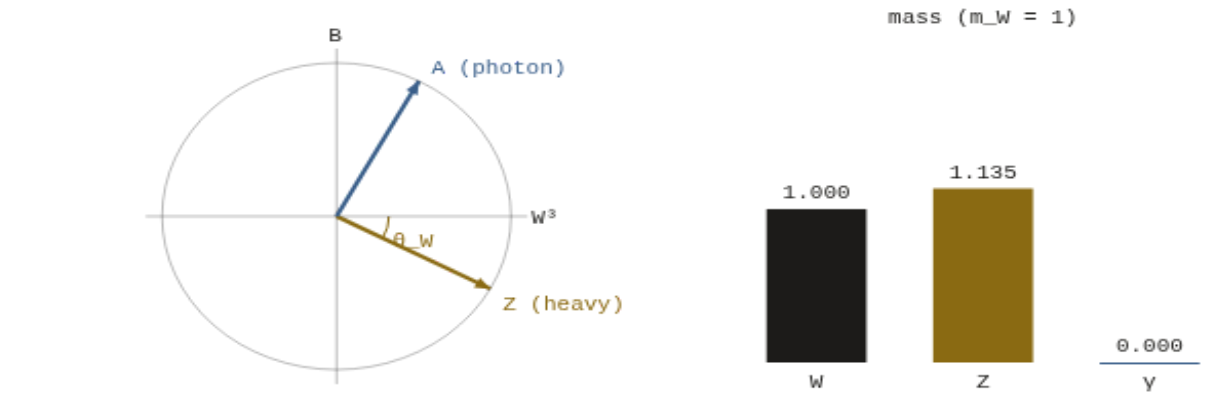
$$Z_\mu=\cos\theta_W\,W^3_\mu-\sin\theta_W\,B_\mu,\qquad A_\mu=\sin\theta_W\,W^3_\mu+\cos\theta_W\,B_\mu \quad (11)$$

$$m_W=\frac{gv}{2},\qquad m_Z=\frac{m_W}{\cos\theta_W},\qquad m_\gamma=0,\qquad e=g\sin\theta_W \quad (12)$$

$W^\pm=(W^1\mp iW^2)/\sqrt{2}$. This is the same as (10.3)(10.5) of the PDG, and the PDG cautions that these are masses "at tree level, i.e., to lowest order in perturbation theory" Known . Computed Building the mass matrix of 4×4 from Pauli matrices, the

eigenvalues agreed with (12) for three choices of (g, g', v) (relative differences below 10^{-15}, of the order of rounding error).

Figure 2 Changing $g'/g = \tan\theta_W$ with the knob rotates the directions of Z and the photon A in the (W^3, B) plane and changes $m_Z/m_W = 1/\cos\theta_W$. "Measured" returns it to 0.536.



g'/g = 0.536 θ_W = 28.19° sin²θ_W = 0.2232 m_Z/m_W = 1.1346 e/g = 0.4724

The five points of Figure 2

g'/g	θ_W	$\sin^2\theta_W$	m_Z/m_W	e/g
0	0.00°	0.0000	1.0000	0.0000
0.25	14.04°	0.0588	1.0308	0.2425
0.536	28.19°	0.2232	1.1346	0.4724
1	45.00°	0.5000	1.4142	0.7071
1.5	56.31°	0.6923	1.8028	0.8321

Computed The numbers in this figure and table were computed on this machine (by (12)). For $g'/g = 0$ the photon is B itself. 0.536 is the value from the measured values of Table 3.

Example. Working (12) back from the PDG 2024 values gives Table 3.

Table 3 Tree-level estimates (input: $G_F = 1.1663788 \times 10^{-5}\ \text{GeV}^{-2}$, $m_W = 80.3692$, $m_Z = 91.1880$, $m_H = 125.20$ GeV)

Quantity	Formula	Value
v	$(\sqrt{2}G_F)^{-1/2}$	246.22 GeV
$g,\ g'$	$2m_W/v,\ g\tan\theta_W$	0.6528, 0.3499
$\sin^2\theta_W$	$1 - m_W^2/m_Z^2$	0.22321
$1/\alpha$	$4\pi/e^2$	132.1
λ	$m_H^2/(2v^2)$	0.1293

Computed In the PDG, the on-shell $\sin^2\theta_W = 0.22348$ and $1/\alpha$ is 137.036 at low energy, and 127.930 at m_Z (MS). On-shell and MS are names for ways of including quantum corrections (renormalisation conventions), and quantities with the same name differ slightly in value (4-04). The λ of Table 3 is the value in the $V = \lambda(\Phi^\dagger\Phi - v^2/2)^2$ convention of the text. The PDG's λ becomes $m_H/v = 0.5085$ with the normalisation $m_H = \lambda v$ ((10.5a)), and the two are related by $\lambda = \lambda_{\text{PDG}}^2/2$. m_W is the average excluding CDF (2022).

The 132.1 for $1/\alpha$ is 3.6 % smaller than the low-energy 137.036. The v of Table 3 was taken from G_F (a constant fixed by the muon lifetime), so the quantum correction Δr linking G_F and α is not included. Squaring (10.22b) of the PDG electroweak review gives $m_W^2 \sin^2\theta_W\,(1-\Delta r) = \pi\alpha/(\sqrt{2}\,G_F)$ ($\sin^2\theta_W$ is on-shell), and the PDG gives $\Delta r = 0.03685$ Known . The tree-level $1/\alpha$ corresponds to $(1-\Delta r)/\alpha$, and $137.036 \times (1 - 0.03685) = 131.99$ agrees with the 132.1 of Table 3 to 0.09 % Computed (quantum corrections in 4-04). The 127.930 at MS is a quantity with a different convention, listed only for reference. The degrees of freedom are 12 both before and after (from the 8 of W^i, B and the 4 of the doublet, to the 9 of the heavy $W^\pm$, Z, the 2 of the photon and the 1 of H) Computed .

Weinberg (1967) wrote that the masses of the W and Z come “not from a mass term put in at the beginning”, and in 2012 ATLAS and CMS reported the observation of the Higgs particle Known .

07

Where mass comes from — the proton and the pion

What v gives directly is the masses of the W and Z, and the masses of quarks and leptons through the **Yukawa couplings** Known . A Yukawa coupling is a term $\bar\psi_L \Phi \psi_R$ that is a product of the Higgs field and spin-1/2 fields; for up-type quarks $\tilde\Phi = i\sigma^2\Phi^*$ is used (a doublet of $Y = -\frac{1}{2}$; the conjugate of § 03). In every term the sum of Y is 0 (for example $\bar Q_L \Phi d_R$ is $-\frac{1}{6} + \frac{1}{2} - \frac{1}{3} = 0$), so it is gauge invariant Computed . Setting Φ to Φ_0 gives a mass term proportional to $\bar\psi_L \psi_R$. But this hardly explains the mass of the proton. The sum of the masses of the core u, u, d is about 1% of the proton's.

$$\frac{2m_u + m_d}{m_p} = \frac{2 \times 2.16 + 4.70}{938.272} \approx 0.0096 \quad (13)$$

Computed Values from the PDG 2024 Summary Tables (m_u, m_d is MS, 2 GeV). Quark masses depend on the definition, so (13) is a guide to orders of magnitude. The rest is thought to come from the dynamics of QCD, whose scale is the Λ produced by quantisation (4-07) Physics .

The pion is light, about 15% of the proton ($m_{\pi^\pm} = 139.57$ MeV) Computed . According to the PDG, in the limit where the masses of the three light quarks are set to 0, the flavour symmetry that turns left and right separately (chiral symmetry) breaks, and “eight massless Goldstone bosons, the π’s, K’s, and η” appear Physics . That QCD is thought to have no gap. This is the “chiral symmetry breaking” that Jaffe–Witten list as the third property of QCD Known . Pure Yang–Mills has no quarks, and its lightest states are thought to be glueballs (5-08) Physics . The gap Δ of the problem statement is a scale

that should arise from the Yang–Mills field alone, relying on neither quarks nor the Higgs field (4-11).

08

The difference from pure Yang–Mills — what the problem statement leaves out

The problem statement concerns pure Yang–Mills theory for any compact simple group (0-03). Table 4 shows what to remove from the Standard Model to get there.

Table 4 The Standard Model, QCD and pure Yang–Mills

Item	Standard Model	QCD	Pure Yang–Mills
Group	$SU(3)\times SU(2)\times U(1)$	$SU(3)$	Compact simple group
Matter fields	Quarks, leptons, Higgs	6 kinds of quark	None
Parameters	19 (as the PDG counts)	g_s, 6 m_q, θ	g, θ (3-05)
Classical mass scale	μ^2 (v) and the Yukawa couplings	m_q	None (3-02 § 07)
Massless particles	Photon	π and others in the limit of massless light quarks Physics	Expected to be none Physics
Treatment in the problem statement	Outside the scope	Extension problem	The main problem

There are three differences. **The group**: it is not simple. The abelian $U(1)_Y$ part is not asymptotically free (the property that the coupling weakens at short distances) (Coleman–Gross, 4-06 § 04) Known , and at one loop the coupling grows at short distances Physics . **Massless particles**: the photon remains, so the Standard Model as a whole has no gap, and a gap can be asked about only for the strong-force part. **Where mass comes from**: in the Higgs mechanism the scale μ^2 is in the classical action and visible from the solution, but the classical action of pure Yang–Mills has no scale, and the gap is thought to arise from quantisation Physics . Jaffe–Witten too, in § 2, give the Higgs mechanism as an example of "classical properties", and the mass gap and confinement as examples of "quantum properties that do not have classical analogs" Known .

What was left out remains in extension problems. Jaffe–Witten list "other four-dimensional gauge theories (incorporating additional fields that preserve asymptotic freedom)" (0-03), and at the time of writing the problem description they call the two-dimensional abelian Higgs model of Brydges–Fröhlich–Seiler "the only complete example of an interacting gauge theory satisfying the axioms" Known . A special limit of the four-dimensional Higgs model is in 7-10, and a picture of where the groups come from is in 17-07.

What this chapter can say

Matter	What is said	What is not said
(1)–(6), Table 1, Table 2	The standard definitions, the arithmetic of charges and sums Known Computed	Why the representations are these, the derivation of the anomaly conditions, confinement
(7)–(12), figures, Table 3	How mass appears in classical fields, tree-level estimates, comparison of $1/\alpha$ using the PDG's Δr Known Computed	A theorem on the Higgs mechanism in quantum theory, values including quantum corrections
Table 4, the SU(5) detour	The PDG's counting, and the arithmetic of assigning representations Known Computed	That grand unification is right, the origin of the 19
(13), § 07	Computation of ratios and the PDG's account Physics	Proofs of these from QCD
The mass gap	—	That pure Yang–Mills has a gap (open; this site shows nothing) Physics

Sources and reproduction

Item	Kind	Source / tools
Definition of the Standard Model, representations, SU(5), the 19, electroweak and QCD formulas and values, quark masses, Goldstone bosons	Reviews and Summary Tables (text checked)	S. Navas et al. (Particle Data Group), Phys. Rev. D 110 (2024) 030001: § 9 QCD, § 10 Electroweak Model, § 60 Quark Masses, § 93 Grand Unified Theories, Summary Tables
Quotations of Jaffe–Witten	Primary source (text checked)	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory”, the official problem description of the Clay Mathematics Institute, § 1, § 2, § 5, § 6
Quotations of Higgs and Weinberg	Primary source (text checked)	P. W. Higgs, Phys. Rev. Lett. 13 (1964) 508 / S. Weinberg, Phys. Rev. Lett. 19 (1967) 1264
Other quotations and summaries	Primary source (abstract checked)	Englert, Brout, PRL 13 (1964) 321 / Bouchiat, Iliopoulos, Meyer, PLB 38 (1972) 519 / Gross, Jackiw, PRD 6 (1972) 477 / Georgi, Glashow, PRL 32 (1974) 438 / Elitzur, PRD 12 (1975) 3978 (same as 3-01) / Witten, “An SU(2) anomaly”, PLB 117 (1982) 324 / ATLAS, PLB 716 (2012) 1 / CMS, PLB 716 (2012) 30
Anderson	Primary source (abstract checked; the wording in the text follows the text of Higgs 1964)	P. W. Anderson, Phys. Rev. 130 (1963) 439

Item	Kind	Source / tools
Brydges–Fröhlich–Seiler	Bibliographic data only (via reference [7] of Jaffe–Witten § 6.4)	“On the construction of quantized gauge fields” I: Ann. Phys. 121 (1979) 227 / II: CMP 71 (1980) 159 / III: CMP 79 (1981) 353
Guralnik–Hagen–Kibble	Bibliographic data only	PRL 13 (1964) 585
Tables 1–3, values in the figures and tables, eigenvalues of the mass matrix, (13), the comparison of Δr, checks of conjugation and Yukawa terms	Computed on this machine	Python 3.12, NumPy 1.26, fractions (`sm_tables.py`, `sm_checks.py`). The figures do (9)(12) in plain JavaScript

Next chapter to read: 4-01 Canonical quantisation and the Hamiltonian.

Canonical quantisation and the Hamiltonian — Gauss's law becomes a condition on states

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — The procedure of imposing commutation relations and building the Hamiltonian (canonical quantisation).
That A_0 has no momentum, and Gauss's law becomes a condition on states: "unchanged by small gauge transformations".
That this condition removes the flat motion in gauge directions, and on a circle fixes the levels by the Casimir.

Prerequisite chapters — 2-05 (spectrum, gap), 3-03 (action, Gauss's law). For oscillators, 2-06.

Said first — In the four-dimensional continuum theory, the Hamiltonian of this chapter is a formula in form only and has not been constructed as an operator of this form (what is rigorous is the lattice, 5-03, and low-dimensional models). The construction published in 2026-11 built a Hamiltonian by reconstruction from the Euclidean side (the committee's assessment is that it is proved in Lean; 12-08). It shows nothing about the mass gap.

Order of this chapter

1. The procedure of canonical quantisation
2. A_0 has no momentum — a constraint
3. Quantising in the temporal gauge — wave functionals
4. Gauss's law is a condition on states — charges and large gauge transformations
5. Seen in the free field — removing the flat directions (Figure 1, Table 1)
6. Yang–Mills on a circle — class functions and the Casimir (Figure 2, Table 2, Table 3)
7. Detour: small boxes, and arguments from wave functionals
8. The route of solving Gauss's law outright — Gribov

9. What is formal

01

The procedure of canonical quantisation — making coordinates and momenta into operators

The problem statement defines the mass gap by the spectrum of the Hamiltonian H (the vacuum is $H\Omega = 0$): "A quantum field theory has a mass gap Δ if H has no spectrum in the interval $(0, \Delta)$ for some $\Delta > 0$." (Jaffe–Witten § 4) Known . One route to writing down that H from the classical field is canonical quantisation (the other is 4-02).

With finitely many degrees of freedom there are three steps (2-05). (i) Make the momenta $p_i = \partial L / \partial \dot{q}_i$. (ii) Write $H = \sum_i p_i \dot{q}_i - L$ in terms of q, p. (iii) Replace q, p by operators on a Hilbert space and impose the following.

$$[q_i, p_j] = i\,\delta_{ij} \qquad (\hbar = 1) \tag{1}$$

On wave functions, $p_j = -i\,\partial/\partial q_j$. For fields the index becomes a point $\mathbf{x}$, and δ_{ij} becomes $\delta^3(\mathbf{x} - \mathbf{y})$. **Example.** A scalar field of mass m splits into an oscillator for each wave number, and the gap was m (2-06 § 08). In Yang–Mills, at step (i) a variable appears for which no momentum can be made.

02

A_0 has no momentum — Gauss's law appears as a constraint

The conventions are the same as in 3-03 § 01 ($\hbar = c = 1$, $D_\mu = \partial_\mu - igA_\mu$, $(D_i X)^a = \partial_i X^a + g f^{abc} A_i^b X^c$ on adjoint fields). The colour electric field is

$$E_i^a = F_{i0}^a = -\partial_t A_i^a + (D_i A_0)^a \tag{2}$$

and $L = \frac{1}{2}\int \sum_a (|\mathbf{E}^a|^2 - |\mathbf{B}^a|^2)\, d^3x$ (the magnetic field contains no time derivative). The momentum of A_i^a is $\Pi_i^a = -E_i^a$, but L does not contain $\partial_t A_0$, so the momentum of A_0 is identically 0.

We write step (ii). What corresponds to $\sum_i p_i \dot{q}_i$ is $\int \Pi_i^a\, \partial_t A_i^a\, d^3x$. Reading (2) as $\partial_t A_i = -E_i + D_i A_0$ and substituting,

$$\int \Pi_i^a\, \partial_t A_i^a\, d^3x = \int \left(|\mathbf{E}|^2 - \mathbf{E}\cdot\mathbf{D}A_0\right) d^3x$$

follows. Subtracting L from this leaves $\tfrac{1}{2}(|\mathbf{E}|^2+|\mathbf{B}|^2)$ and $-\int \mathbf{E}\cdot\mathbf{D}A_0$. In the last term, integration by parts moves D onto $\mathbf{E}$ (we take the fields to be 0 far away). The derivative part ∂_i changes sign by ordinary integration by parts. For the $gf^{abc}A_i^b$ part, f^{abc} changes sign under exchange of indices (it is antisymmetric), so relabelling the indices likewise changes the sign once. So $-\int \mathbf{E}\cdot\mathbf{D}A_0\,d^3x=\int A_0^a(\mathbf{D}\cdot\mathbf{E})^a\,d^3x$, and

$$H=\int\Big[\tfrac{1}{2}\big(|\mathbf{E}|^2+|\mathbf{B}|^2\big)+A_0^a\,(\mathbf{D}\cdot\mathbf{E})^a\Big]d^3x \quad (3)$$

follows ($|\mathbf{E}|^2=\sum_a|\mathbf{E}^a|^2$) Known . A_0 is an undetermined multiplier that appears only linearly (the λ of the toy in 2-01 § 08), and its equation $(\mathbf{D}\cdot\mathbf{E})^a=0$ is Gauss's law ((3) of 3-03). Such a condition is called a **constraint**, and its general treatment is said to begin with Dirac (1950).

03

Quantising in the temporal gauge — wave functionals and the electric field operator

We set $A_0=0$ by a gauge transformation (the **temporal gauge**; it is enough to solve $\partial_t\,\Omega=-ig\,\Omega A_0$ at each point). The remaining freedom is time-independent gauge transformations. The coordinates are $A_i^a(\mathbf{x})$ and the momenta are the $\Pi_i^a=-E_i^a$ of § 02. Since $\Pi=-E$, the field version of (1), $[A,\Pi]=i\delta$, is

$$[A_i^a(\mathbf{x}),E_j^b(\mathbf{y})]=-i\,\delta^{ab}\delta_{ij}\,\delta^3(\mathbf{x}-\mathbf{y}) \quad (4)$$

(the sign is negative because $E=-\Pi$). We represent states by functions of configurations (**wave functionals**) $\psi[A]$. For small ϵ and a function $h_i^a(\mathbf{x})$,

$$\psi[A+\epsilon h]-\psi[A]\approx\epsilon\int h_i^a(\mathbf{x})\,\frac{\delta\psi}{\delta A_i^a(\mathbf{x})}\,d^3x$$

the function $\delta\psi/\delta A_i^a(\mathbf{x})$ for which this holds is called the **functional derivative**. It is the finite-dimensional $\psi(q+\epsilon h)-\psi(q)\approx\epsilon\sum_i h_i\,\partial\psi/\partial q_i$ with the index i replaced by the point $\mathbf{x}$ and the sum by an integral. Following the $p=-i\,\partial/\partial q$ of (1), $\Pi_i^a(\mathbf{x})=-i\,\delta/\delta A_i^a(\mathbf{x})$, so it acts as $E_i^a(\mathbf{x})=i\,\delta/\delta A_i^a(\mathbf{x})$, and (3) is formally

$$H=\int\Big[-\tfrac{1}{2}\frac{\delta^2}{\delta A_i^a(\mathbf{x})^2}+\tfrac{1}{2}B_i^aB_i^a\Big]d^3x \quad (5)$$

($\delta^2/\delta A_i^a(\mathbf{x})^2$ is summed over a and i) Known . It can be read as a particle with infinitely many coordinates moving in the potential energy $\tfrac{1}{2}\int|\mathbf{B}|^2$. Strictly, (5) is form only (§ 09). Rescaling as $\hat{A}=gA$,

$$H = \int \Big[-\frac{g^2}{2}\,\frac{\delta^2}{\delta \hat{A}^2} + \frac{1}{2g^2}\,|\hat{\mathbf{B}}|^2 \Big]\, d^3x \tag{6}$$

and when g is large the electric term wins, when small the magnetic term. The coefficients $g^2/2a$ and $2N/g^2 a$ in (16) of 5-03 are the lattice version of these two terms. **Example.** Of the $3(N^2-1)$ components A_i^a at each point, the condition of the next section removes N^2-1 gauge directions, leaving $2(N^2-1)$ (the same number as in 3-03 § 03). The one component that the classical theory removed with two things, the constraint and the remaining gauge transformations, is removed at once by (10) in the quantum theory.

04

Gauss's law is a condition on states — unchanged by small gauge transformations

Since A_0 has been removed, Gauss's law is no longer among the equations of motion. The Gauss operator

$$G^a(\mathbf{x}) = (\mathbf{D}\cdot\mathbf{E})^a = \partial_i E_i^a + g f^{abc} A_i^b E_i^c \tag{7}$$

has three properties Known . (a) **It generates gauge transformations.** Setting $G(\alpha) = \int \alpha^a G^a \, d^3x$, integration by parts and (4) give

$$G(\alpha)\,\psi[A] = -i\,\frac{d}{d\epsilon}\,\psi[A + \epsilon\, D\alpha]\Big|_{\epsilon=0} \tag{8}$$

($A \mapsto A + \epsilon D\alpha$ is (4) of 3-02 with α taken as $g\epsilon\alpha$, that is, an infinitesimal gauge transformation). (b) **It commutes with H.** The two terms of (5) are gauge invariant, so $[H, G^a(\mathbf{x})] = 0$. (c) **It forms a closed algebra.**

$$[G^a(\mathbf{x}), G^b(\mathbf{y})] = -ig\, f^{abc}\, G^c(\mathbf{x})\, \delta^3(\mathbf{x}-\mathbf{y}) \tag{9}$$

$G^a = 0$ cannot be made an operator identity, because the commutator of $G(\alpha)$ and A_j is $-i\,D_j\alpha \;/= 0$. Instead it is imposed on states.

$$G^a(\mathbf{x})\,\psi = 0 \iff \psi[A^\Omega] = \psi[A] \tag{10}$$

The Ω on the right are those continuously connected to the identity transformation. States satisfying (10) are the **physical states**. By (b) time evolution preserves this subspace, and by (c) the conditions do not contradict one another. Observables are restricted to gauge-invariant ones, and the mass gap is the gap of H on the physical states (4-11 § 09).

Example. For a uniform field of $SU(2)$ (a real matrix A_i^a of 3×3; uniform fields are in 3-02 § 07), $G^a = ig\,\varepsilon^{abc}A_i^b\,\partial/\partial A_i^c = -gL^a$. $L^a = -i\varepsilon^{abc}A_i^b\,\partial/\partial A_i^c$ is the angular momentum that turns the colour index, and satisfies $[L^a,L^b] = i\varepsilon^{abc}L^c$. This is the same commutation relation as that of the generators $[T^a,T^b] = if^{abc}T^c$ (3-02 § 02), so $G = -gL$ gives the $-ig$ of (9). Gauge-invariant functions are functions of $A^{\mathsf{T}}A$ and of the sign of $\det A$. With differences we checked that G^a annihilates $\frac{1}{2}|\mathbf{B}|^2$ and $\det A$ to within 2×10^{-10} but does not annihilate $A\frac{1}{1}$, and that (9) holds to relative 2×10^{-11} Computed .

Charges, and large gauge transformations

If there are static colour charges, the condition is $G^a\psi = \rho^a\psi$. The abstract of Kogut–Susskind (1975) describes the gauge-invariant configurations on the lattice as "a collection of strings with quarks at their ends. The strings are lines of non-Abelian electric flux." Known (Table 3). Also, what (8) generates is only transformations connected to the identity. A transformation T that approaches 1 far away and has winding number 1 from $\mathbb{R}^3\cup\{\infty\} = S^3$ to $SU(2)$ (3-05 § 07) is not connected to the identity. So $\psi[A^T] = e^{-i\theta}\psi[A]$ does not contradict (10) (with the convention $(T\psi)[A] = \psi[A^T]$ it has the same direction as (12) of 3-05; with the opposite convention, $e^{+i\theta}$). This is the θ of θ vacua Physics .

05 Seen in the free field — Gauss's law removes the flat directions

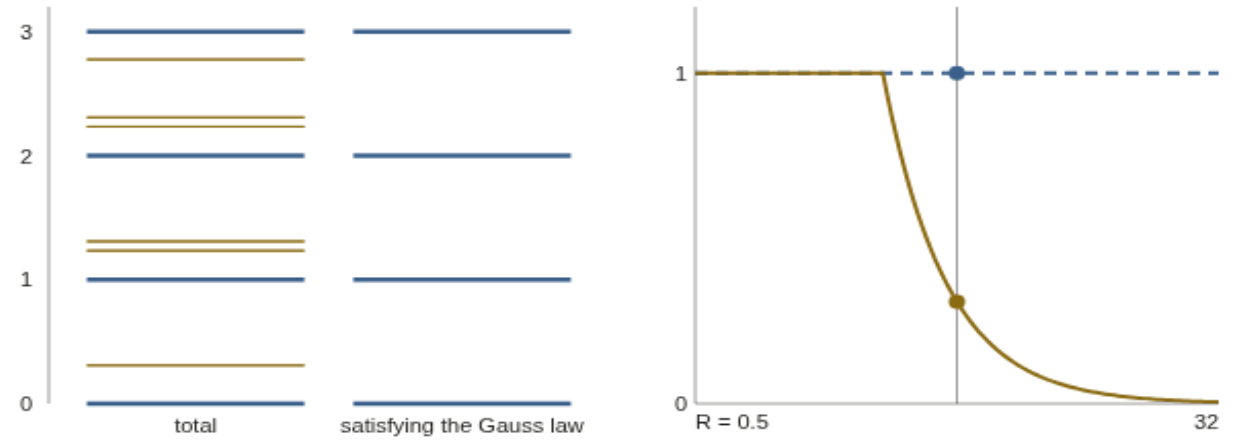
With $g = 0$, split the component of wave number $\mathbf{k}$ into the longitudinal part a_L parallel to $\mathbf{k}$ and the transverse part a_T perpendicular to it. The magnetic field contains only the transverse part, so writing just one transverse direction,

$$H_{\mathbf{k}} = \tfrac{1}{2}\left(p_T^2 + p_L^2\right) + \tfrac{1}{2}|\mathbf{k}|^2 a_T^2 \quad (11)$$

and $G_{\mathbf{k}} \propto \mathbf{k}\cdot\mathbf{E}_{\mathbf{k}} \propto p_L$. The longitudinal direction is the direction of the gauge transformation $\mathbf{A}\mapsto\mathbf{A}+\nabla\chi$ and has no restoring force, so in the whole H the energy of longitudinal motion piles up continuously on top of each level. Gauss's law $p_L\psi = 0$ removes this continuum, and the Fock space of the transverse oscillators (photons; 2-06 § 07) remains Known . Figure 1 closes the longitudinal direction into a circle of length $2R$ ($|\mathbf{k}| = 1$) and shows how, of the levels $E(n,m) = n + \frac{1}{2}(\pi m/R)^2$ ($m\in\mathbb{Z}$), $m = 0$ remain.

Figure 1 Changing the length R of the longitudinal direction with the knob changes the levels on the left and the gap on the right. Thick lines are states constant along the longitudinal direction ($m = 0$), thin lines states that

move along it. On the right the solid line is the gap of the whole space, the dashed line the gap of the states satisfying Gauss's law




```
R = 4.00: total gap 0.3084, physical 1, levels with E ≤ 3: 16, physical 4
```

Computed The numbers in this figure were computed on this machine (closed formulas computed on the spot; agreeing with Table 1).

Computed Table 1 The longitudinal length R and the gap ($|\mathbf{k}| = 1$). The number of levels is $E \le 3$ (always 4 for physical ones)

R	Total gap	Physical gap	Number of levels (total)
2	1.00000	1	8
4	0.30843	1	16
32	0.00482	1	120

The total gap $\min(1, \pi^2/2R^2)$ is 0 for $R \to \infty$: the gap of H itself in the temporal gauge is filled by motion in gauge directions. The physical gap $|\mathbf{k}|$ also tends to 0 as the box is enlarged, and the free field has no mass gap (for a field of mass m the gap was m; 2-06 § 08).

06

Yang–Mills on a circle — states are class functions of the holonomy

When space is a circle of length L ($1+1$ dimension) there is no magnetic field, and $H = \frac{1}{2}\int_0^L E^a E^a \, dx$. Gauss's law says that E is carried by parallel transport (1-07), and parallel transport in the adjoint preserves length, so $H = \frac{L}{2} E^a E^a(0)$. All that remains for gauge-invariant functions is the conjugacy class of the holonomy round the circle, $U = P\exp(ig\oint A_1 dx)$. A function unchanged by conjugation $U \mapsto gUg^{-1}$ is called a **class function**, and the trace of a representation matrix a **character** (1-02). Witten (1992) § 4 writes "Ψ must be a class function of U, invariant under conjugation, and so must have an expansion in characters", and expresses H by the Casimir operator Known . $E^a(0)$ acts as g times the derivative that turns U in the direction T^a, and

$$H\,\chi_r(U) = \frac{g^2 L}{2}\,C_2(r)\,\chi_r(U) \tag{12}$$

follows (χ_r is the character of the irreducible representation r, and C_2 the value of $\sum_a T^a T^a$; 1-04). The eigenvalues were found exactly by Rajeev (1988), and the route from the transfer matrix is 7-02 § 07. **Example.** For $SU(2)$, writing the eigenvalues of U as $e^{\pm i\theta}$, $\chi_j = \sin((2j+1)\theta)/\sin\theta$, and the levels are $g^2L/2$ times $j(j+1)$. The gap $\frac{3}{8}g^2L$ can be read as the energy of a loop of electric flux running once round the circle (tension $\frac{3}{8}g^2$).

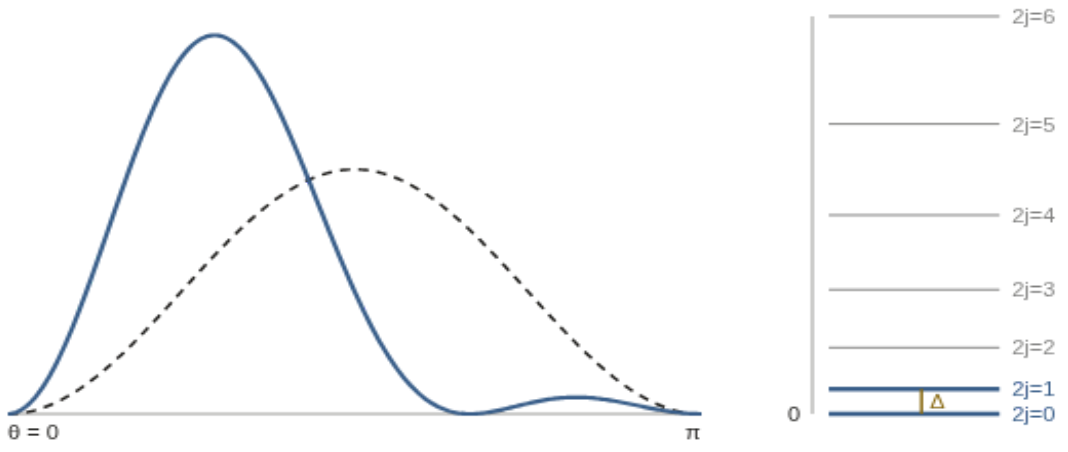
The volume of the orbit fixes the levels

The inner product of class functions is taken with Haar measure $\frac{2}{\pi}\sin^2\theta\,d\theta$. Taking the density $\frac{2}{\pi}\sqrt{1-a_0^2}\,da_0$ (a semicircle) of a_0 in 1-08 § 03 and setting $a_0 = \cos\theta$ gives this form. $\sin^2\theta$ is the area of the conjugacy class (a sphere in S^3), that is, the volume of the gauge orbit (4-02 § 07). Writing (12) with θ,

$$H = \frac{g^2 L}{2}\Big(-\frac{1}{4\sin^2\theta}\frac{d}{d\theta}\sin^2\theta\frac{d}{d\theta}\Big) \tag{13}$$

and on $u = \sin\theta\, f$, $\frac{g^2L}{8}(-u'' - u)$, that is, a particle in a box on the interval $[0,\pi]$ ($u = \sin n\theta$, $n = 2j+1$) minus 1. If the measure is forgotten and the flat Laplacian is used, the levels change to $k^2/4$ (Table 2). In the three-dimensional Coulomb gauge this volume factor appears as the Faddeev–Popov determinant (§ 08).

Figure 2 A circle of $SU(2)$ ($g = 1$). Choose $2j$ and L and press "Play" to advance time (1 second on screen is time 1); the density (solid line) of θ for the state that superposes χ_0 and χ_j half and half oscillates with angular frequency $\Delta = E_j - E_0$. Making $2j$ or L larger makes the oscillation faster. The dashed line is the Haar weight; on the right, the levels (12)




```
2j = 1, L = 2.0: Δ = 0.750, period 2π/Δ = 8.38, phase Δt = 0.00π
```

Computed The numbers in this figure were computed on this machine (characters and the closed formula of (12) computed on the spot; agreeing with Table 2).

Computed Table 2 Eigenvalues of (13) (in units of $g^2L/2$; differences with 4000 Haar-weighted points), and the $k=2j$-th from the bottom of $-\frac{1}{4}d^2/d\theta^2$ with the measure forgotten (slope 0 at both ends of $[0,\pi]$; a periodic function symmetric under $\theta\mapsto-\theta$)

2j	j(j+1)	Differences (Haar)	k²/4 with the measure forgotten
0	0	0.000000	0
1	0.75	0.750000	0.25
2	2	2.000000	1
3	3.75	3.749998	2.25

Counting on the lattice — states that violate Gauss's law

Cut the circle into N links ($a=L/N$) and set $H=\frac{g^2a}{2}\sum_\ell E_\ell^2$ (the two-dimensional version of 5-03 § 07; there are no plaquettes). As a metaphor, one link is a top spinning on $SU(2)$ (5-03 § 06), with $(2j+1)\times(2j+1)$ states of angular momentum j for left and right rotations. Precisely, this is the Peter–Weyl theorem, which splits the state space $L^2(SU(2))$ of a link by representation Known . So the energy is $\frac{g^2a}{2}\sum_\ell j\ell(j\ell+1)$ and the multiplicity $\prod_\ell(2j\ell+1)^2$. At a vertex the ends of two links meet, and Gauss's law requires that the combination of their two representations be unchanged by rotation (be a scalar). 0 appears in the coupling of angular momenta $j\otimes j'=|j-j'|\oplus\cdots\oplus(j+j')$ only when $j=j'$, and then exactly once (Schur's lemma) Known . So neighbouring j are equal, there is one physical state for each j, and the levels are (12) regardless of N.

Computed Table 3 A lattice circle of $SU(2)$ (in units of $g^2L/2$). The lowest physical excitation is always 0.75, and there are 3 physical states with $E\le 2$

N	Lowest excitation (total)	Number of states with $E\le 2$ (total)
1	0.75	14
4	0.1875	115,504
16	0.046875	2.44×10^{22}

The lowest excitation of the whole space is a single link with $j=\frac{1}{2}$, that is, a state with a pair of charges at its two ends (§ 04), and its energy $\frac{3}{8}g^2a$ is the tension times the distance a. It tends to 0 as $a\to0$, so $\frac{3}{8}g^2L$ remains only once we restrict to physical states.

07

Detour: small boxes, and arguments from wave functionals

A section you may skip.

In a three-dimensional periodic box (side ℓ) the uniform components $A_i^a = c_i^a$ are special. Restricting (5) to these components gives $H_0 = -\frac{1}{2\ell^3}\partial_c^2 + \frac{\ell^3}{2}|\mathbf{B}(c)|^2$ (the magnetic field is only the commutator term $g\,c^2$; 3-02 § 07), and rescaling as $c = g^{-1/3}\ell^{-1}y$,

$$H_0 = \frac{g^{2/3}}{\ell}\Big[-\tfrac{1}{2}\frac{\partial^2}{\partial y^2} + \tfrac{1}{2}|\mathbf{B}_1(y)|^2\Big] \quad (14)$$

($\mathbf{B}_1$ is the magnetic field of $g = 1$). The levels are proportional to $g^{2/3}/\ell$, and if g is small they lie below the $2\pi/\ell$ of the non-uniform components. The abstract of Lüscher (1983) writes "When non-abelian gauge fields are enclosed in a box with periodic boundary conditions, the spectrum of the hamiltonian becomes discrete", and expands the levels in powers of $g^{2/3}$ Known . Keeping only two components inside the bracket gives the x^2y^2 well of 2-05. The small box (the side where g is small, by asymptotic freedom) is the opposite end from the four-dimensional question, which asks about infinite volume. Lüscher's g is the renormalised coupling measured at the size of the box, not the classical g of (14) itself.

Feynman (1981), in $2+1$ dimensions, after cautioning in the abstract that "Only qualitative or suggestive discussions are made.", argues that the gauge invariance of the wave functional leads to "a finite energy for any excitation ("glueball") above the ground state" Physics .

08

The route of solving Gauss's law outright — the Coulomb gauge and Gribov

There is also the route of fixing the gauge further and solving the constraint first. Splitting as $\mathbf{E} = \mathbf{E}_T - \nabla\phi$ ($\nabla\cdot\mathbf{E}_T = 0$) in the Coulomb gauge $\nabla\cdot\mathbf{A} = 0$, Gauss's law is a linear equation for ϕ.

$$M\phi = -\rho, \qquad M = -\nabla\cdot\mathbf{D} \quad (15)$$

$\rho^a = g f^{abc} A_i^b E_{T,i}^c$ is the colour charge carried by the transverse field itself. If M can be inverted, then with $K = M^{-1}(-\nabla^2)M^{-1}$,

$$H = \tfrac{1}{2}\int(|\mathbf{E}_T|^2 + |\mathbf{B}|^2) + \tfrac{1}{2}\int \rho\, K\rho \quad (16)$$

(the integral is d^3x) Known . The abstract of Christ–Lee (1980) writes that the change of variables from the temporal gauge produces "an unconventional operator ordering", and that "new nonlocal interactions, even in the familiar Coulomb gauge" appear Known . The volume factor of the orbit corresponding to the $\sin^2\theta$ of § 06 is $\det M$ Known . This is the Jacobian of changing variables to the gauge condition (coordinates on the slice $\nabla\cdot\mathbf{A} = 0$), and is called the Faddeev–Popov determinant (details in 4-08). The difficulty is that when A is large, M has zero eigenvalues. The abstract of Gribov (1978) writes "the fixing of the divergence of the potential in non-Abelian theories does

not fix its gauge” Known . That a slice cannot be chosen globally is Singer's theorem in 3-04. The (10) of the temporal gauge avoids this difficulty.

09

What is formal — the problem of defining the Hamiltonian

In four dimensions (5) lacks meaning in three respects. (i) $E_i^a(\mathbf{x})E_i^a(\mathbf{x})$ is a product of distributions at the same point (1-11 § 09, 1-09 § 09). (ii) The sum of zero-point energies $\frac{1}{2}|\mathbf{k}|$ diverges (2-06 § 09), and with interactions this is paired with renormalisation (4-04). (iii) In the inner product $\int |\psi|^2 \mathcal{D}A$, the absence of Lebesgue measure in infinite dimensions (1-09 § 02) overlaps with the divergence of the orbit volume (the fact that an interacting H cannot be placed on the Fock space of the free field, that is, Haag's theorem, is also involved; 4-09).

On the lattice the state space is fixed, and Gauss's law can be imposed exactly at each vertex (Table 3). What remains is whether the gap stays positive in the limit where the lattice spacing goes to 0 and the box to infinity. The problem statement writes “Establishing a quantum mechanical Hilbert space is part of the solution to this Millennium problem.” (§ 6.5) Known . § 6.1 states that the Osterwalder–Schrader construction links the route of Hilbert spaces and the route of functional integrals. § 6.5 writes that much of the mathematical progress so far has come from understanding functional integration (“Much of the mathematical progress reviewed above results from understanding functional integration”). On this route H is reconstructed from the measure in imaginary time (4-02 § 08, 4-10 § 09). This chapter shows nothing about four-dimensional Yang–Mills.

What this chapter can say

Matter	What is said	What is not said
(1)–(10)	Formal derivation Known . Check on uniform fields Computed	Meaning as operators for fields in four dimensions. θ is at the level of physics
(11)–(13), figures, tables	The free-field model and the levels on a circle Known , differences and counting Computed	The case of space of two or more dimensions
(14)–(16), detour	Counting of scales and classical rewriting Known	Terms from quantum ordering (text not checked). Feynman's argument is qualitative Physics

Matter	What is said	What is not said
The mass gap	—	In four dimensions, H was built by reconstruction in the construction published in 2026-11 (at the level of the committee's assessment; 12-08). Its gap has not been shown (6-04). This chapter shows nothing

Sources and reproduction

Item	Kind	Source / tools
The problem statement (§ 4, § 6.1, § 6.5)	Text	A. Jaffe, E. Witten, *Quantum Yang–Mills Theory*, Clay Math. Inst. (2000)
Class functions and the Casimir on a circle	Text § 4, § 4.2	E. Witten, J. Geom. Phys. 9 (1992) 303–368, arXiv:hep-th/9204083
The ordering terms of § 08 and $\det M$	Abstract (text bibliographic data only; the $\det M$ term not checked)	N. H. Christ, T. D. Lee, Phys. Rev. D 22 (1980) 939
Quotations in § 04 and § 06–§ 08	Abstract	J. Kogut, L. Susskind, Phys. Rev. D 11 (1975) 395 / S. G. Rajeev, Phys. Lett. B 212 (1988) 203 / M. Lüscher, Nucl. Phys. B 219 (1983) 233 / R. P. Feynman, Nucl. Phys. B 188 (1981) 479 / V. N. Gribov, Nucl. Phys. B 139 (1978) 1
General theory of constraints	Bibliographic data only	P. A. M. Dirac, Can. J. Math. 2 (1950) 129–148 (attribution not checked)
(2)–(10), (15)(16)	Known	Standard derivation. Signs recomputed in the conventions of 3-03
The check in § 04, Tables 1–3, figures	Computed	Python 3.12, NumPy, SciPy (`gauss_uniform.py`, `flat_mode.py`, `holonomy.py`, `ring_counts.py`). The figures are plain JavaScript

Next chapter to read: 4-02 The path integral and Euclideanisation. The lattice Hamiltonian is in 5-03, gauge fixing in 4-08.

The path integral and Euclideanisation — make time imaginary, and field theory becomes a problem of probability in four dimensions

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That the path integral of a field is a sum over all field configurations.
That making time imaginary turns the weight e^{iS} into e^{-S_E}, and for Yang–Mills, since $S_E \geq 0$, it can be read as a probability weight.
That Lorentz transformations become four-dimensional rotations, and time becomes one of four equal directions.

Prerequisite chapters — 2-07 (the path integral, imaginary time), 2-04 (Lorentz transformations, the action of a field). § 04 uses the non-abelian field strength of 3-02 and 3-03. For the Hamiltonian, 4-01; for Gaussian measures, 1-09.

Said first — Many of the manipulations in this chapter are formal. Wick rotation is a heuristic, and the rigorous correspondence is the Osterwalder–Schrader reconstruction (4-10). The measure of four-dimensional Yang–Mills has not been constructed, and this chapter shows nothing about the mass gap.

Order of this chapter

1. From paths to field configurations — on the lattice, an integral tens of millions of times multiple
2. Making time imaginary — iS becomes $-S_E$
3. The rotation seen in the propagator — turning while avoiding the poles (Figure 1, Table 1)
4. Rotating the Yang–Mills field — $S_E \geq 0$, and the θ term stays imaginary
5. Spacetime becomes space — boosts become rotations (Table 2)
6. Field theory as a probability measure — on the lattice, a genuine measure (Figure 2)

7. What is missing for gauge fields — the volume of the orbits (Table 3)
8. Correlations in imaginary time read the energy — the link to the gap
9. Detour: writing the vacuum wave function on half of spacetime — checked with the oscillator

---

01

From paths to field configurations — what is summed becomes a function

In 2-07 we summed e^{iS} over the paths $x(t)$ of one particle. In field theory, instead of paths, we sum over all functions on spacetime $\varphi(t,\vec{x})$, all histories of the field. The correlation in the vacuum Ω (T is the convention of putting later times to the left) is formally written

$$\langle\Omega|\,\mathrm{T}\,\varphi(x_1)\cdots\varphi(x_n)\,|\Omega\rangle=\frac{\int\mathcal{D}\varphi\;\varphi(x_1)\cdots\varphi(x_n)\,e^{iS[\varphi]}}{\int\mathcal{D}\varphi\;e^{iS[\varphi]}}\qquad(1)$$

Even for one particle $e^{iS}\mathcal{D}x$ was not a measure (2-07 § 03). For fields there are also infinitely many variables at each time (1-09).

Example. Cut into a lattice, (1) is a finite multiple integral. A lattice of space 32^3 and time 64 has 2,097,152 points, and the Yang–Mills field of $SU(3)$ has 4 links per point and 8 real numbers per link, so it is a 67,108,864-fold integral Computed . The difficulties are two: the oscillation of e^{iS}, and the limit of finer steps and a larger box. This chapter removes the first (the second is Parts 5 and 7).

02

Making time imaginary — iS becomes $-S_E$

The substitution $t=-i\tau$ has appeared twice. In 2-04 § 09 we only substituted it into the wave equation and saw oscillation turn into decay. In 2-07 § 05 we rotated the weight of the one-particle path integral. Here we rotate the action and correlations of fields. Putting $t=-i\tau$ ($\hbar=c=1$, signs $(-,+,+,+)$, $x_4=\tau$) into the scalar field action ((7) of 2-04 with the potential energy V added), from $dt=-i\,d\tau$ and $(\partial_t\varphi)^2=-(\partial_\tau\varphi)^2$,

$$iS_M=-S_E,\qquad S_E[\varphi]=\int d^4x\Big[\tfrac12\sum_{\mu=1}^{4}(\partial_\mu\varphi)^2+\tfrac12 m^2\varphi^2+V(\varphi)\Big]\qquad(2)$$

follows. The sign of the time derivative has come into line with the space derivatives, and the four directions are equal. S_E is called the Euclidean action, and if V is

bounded below, the weight e^{-S_E} is positive and bounded. (1) is replaced by the **Schwinger functions** (the name is said to come from Schwinger 1958).

$$S_n(x_1,\dots,x_n)=\frac{1}{Z}\int \mathcal{D}\varphi\ \varphi(x_1)\cdots\varphi(x_n)\,e^{-S_E[\varphi]},\qquad Z=\int \mathcal{D}\varphi\ e^{-S_E[\varphi]} \tag{3}$$

Example. If $V=0$, (3) is a Gaussian integral with covariance $(-\Delta+m^2)^{-1}$ (Δ is the Laplacian $\sum_\mu \partial_\mu^2$; not the Δ of the mass gap), and $S_2(x,y)=G(x-y)$ is the propagator of (10) of 1-11. Note. (1) is not a convergent integral, so putting in $t=-i\tau$ is not a change of variables but a rewriting rule. The precise content is the analytic continuation of correlation functions; seen in a single integral, it takes the form of the next section.

03

The rotation seen in the propagator — turning the contour while avoiding the poles

The two-point function of a free field of mass m, written in momentum, is

$$\langle\Omega|\,\mathrm{T}\,\varphi(x)\varphi(0)\,|\Omega\rangle=\int\frac{d^4p}{(2\pi)^4}\,e^{ip\cdot x}\,\frac{-i}{p^2+m^2-i\varepsilon},\qquad p^2=-(p^0)^2+|\boldsymbol{p}|^2 \tag{4}$$

$\varepsilon\to+0$ fixes on which side the poles are avoided (Feynman's prescription). We first look at the component with time difference $t=0$. Then in $e^{ip\cdot x}$ the factor e^{-ip^0t} from p^0 is 1. Setting $\omega=\sqrt{|\boldsymbol{p}|^2+m^2}$, the integrand in p^0 is $i/((p^0)^2-\omega^2+i\varepsilon)$, and the poles $\pm\sqrt{\omega^2-i\varepsilon}$ are in the fourth and second quadrants.

Here we use just one fact about complex integrals. A **pole** is a point where the denominator is 0. In the complex plane, integrating along a closed path that does not enclose a pole gives 0 (**Cauchy's integral theorem**; precisely, the condition is that the function be complex-differentiable on and inside the path, which a rational function whose denominator does not vanish satisfies) Known . So moving the contour continuously without crossing a pole does not change the value.

Turn the contour anticlockwise as $p^0=e^{i\theta}s$ (s real). While θ is between 0 and 90°, the moving line sweeps only the first and third quadrants and does not cross a pole. Joining the two lines before and after turning by distant arcs of radius R makes a closed path that encloses no pole. The length of the arcs is proportional to R and the integrand decreases as $1/R^2$, so the contribution of the arcs vanishes as $1/R$. So the integral along the real axis equals the integral along the turned line, $\theta=90^\circ$, that is, at $p^0=ip_4$,

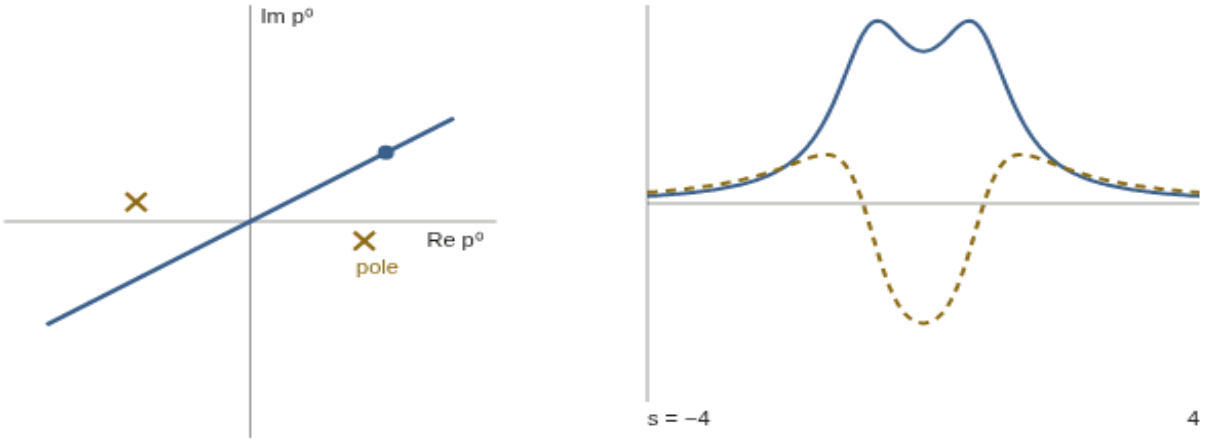
$$\int_{-\infty}^{\infty} \frac{dp^0}{2\pi}\, \frac{i}{(p^0)^2 - \omega^2 + i\varepsilon} = \int_{-\infty}^{\infty} \frac{dp4}{2\pi}\, \frac{1}{p_4^2 + \omega^2} = \frac{1}{2\omega} \qquad (\varepsilon \to 0) \quad (5)$$

The integrand on the right-hand side is positive and smooth. In four dimensions $-i\,d^4p/(p^2 + m^2 - i\varepsilon)$ is replaced by $d^4p_E/(p_E^2 + m^2)$, and this is the Euclidean propagator Known . $1/(2\omega)$ is the $\langle x^2 \rangle$ of the ground state of an oscillator of frequency ω (2-06), and each Fourier component of a free field was an oscillator (2-06 § 08).

The $t \;/= 0$ component carries the factor $e^{-ip^0 t}$. If $p^0 = e^{i\theta} s$ is turned into the first quadrant, on the side of $st > 0$, $|e^{-ip^0 t}| = e^{st \sin\theta}$ grows far away, so the contribution of the arcs does not vanish. So t is also turned by the same angle in the opposite direction, as $t = e^{-i\theta}\tau$. Then at every angle along the way $p^0 t = s\tau$ is real, the factor stays an oscillation of modulus 1, and the argument with arcs goes through. At $\theta = 90°$, $t = -i\tau$, $p^0 = ip4$, $p^0 t = p4\,\tau$. This is why time turns clockwise and energy anticlockwise.

Turning the other way crosses a pole and changes the sign (Figure 1, Table 1), so the direction is fixed by $i\varepsilon$. The name is said to come from Wick's (1954) treatment of the Bethe–Salpeter equation.

Figure 1 Changing the angle θ changes the contour in the p^0 plane (thick line in the contour plot) and the integrand on it (in the integrand plot, solid line real part, dashed line imaginary part). As long as no pole (×) is crossed, the integral $I(\theta)$ does not change ($\omega = 1$)



θ = 30°, ε = 0.40: I(θ) = 0.4731 + 0.0911i, 1/(2√(ω²−iε)) = 0.4731 + 0.0911i, (above the pole angle -10.9°)

Computed The numbers in this figure were computed on this machine (trapezoidal rule with $s = \tan u$).

Computed Table 1 $I(\theta)$ of Figure 1 (python3). The angle of the pole for $\varepsilon = 0.4$ is $-10.90°$.

θ	ε	I(θ)
90°, 30°, 0°, −10°	0.4	+0.473094 + 0.091110i
−12°, −45°, −90°	0.4	−0.473094 − 0.091110i
0°	0.001	0.500000 + 0.000250i

The value is the same whatever the angle; turning the other way past the pole only reverses the sign.

04

Rotating the Yang–Mills field — $E^2 - B^2$ becomes $E^2 + B^2$

The notation follows the $\hat{A} = gA$ of 3-03 § 01, written without hats. That is, g is absorbed into A, A_μ are Hermitian matrices, $D_\mu = \partial_\mu - iA_\mu$, $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - i[A_\mu, A_\nu]$ (the same as (9) of 2-04), $\operatorname{tr}(T^a T^b) = \frac{1}{2}\delta^{ab}$.

The gauge field is a 1-form (1-07), so it is rotated as $A_0\, dt = A_4\, d\tau$, that is, $A_0 = iA_4$. Together with $\partial_0 = i\partial_4$,

$$F_{0j} = i\partial_4 A_j - i\partial_j A_4 - i[iA_4, A_j] = i\,F_{4j}, \qquad F_{\mu\nu}F^{\mu\nu} = -2F_{0j}F_{0j} + F_{jk}F_{jk} = \sum_{\mu,\nu=1}^{4} F$$

In words, only the components with one time index (the components of the electric field) get one i, which squares to -1. In the Minkowski metric the square of the electric field carried -1, so the two cancel, and all four indices line up with the same sign. This is the same thing as the sign of $(\partial_t \varphi)^2$ coming into line with the space derivatives in (2) for the scalar field. Putting this into $S_M = -\frac{1}{2g^2}\int dt\, d^3x \operatorname{tr} F_{\mu\nu}F^{\mu\nu}$ ((1) of 3-03 written in hatted notation) leaves $i \cdot (-i) = 1$, and with $iS_M = -S_E$,

$$S_E[A] = \frac{1}{2g^2}\int d^4x \operatorname{tr} F_{\mu\nu}F_{\mu\nu} = \frac{1}{g^2}\int d^4x \operatorname{tr}\left(E_E^2 + B^2\right) \;\geq\; 0, \qquad E_{E,j} = F_{4j},\; B_j = \frac{1}{2}\varepsilon$$

For Hermitian X, $\operatorname{tr} X^2 = \sum |X_{ij}|^2 \geq 0$, so $S_E \geq 0$, with equality only when $F = 0$. For configurations of topological charge Q, $S_E \geq 8\pi^2|Q|/g^2$ ((12) of 3-03, 3-05) Known .

Example. For configurations with $A_4 = 0$ and A_j independent of τ, $E_E = 0$, so over imaginary time of length T, S_E is $T\mathcal{E}$ ($\mathcal{E} = \frac{1}{g^2}\int \operatorname{tr} B^2 d^3x$ is the magnetic energy), and the weight is $e^{-T\mathcal{E}}$.

The θ term stays imaginary

The topological term θQ is different. Each component contains exactly one F with a time index, so the i from the rotation and the $-i$ of dt cancel, and the weight is $e^{-S_E} e^{i\theta Q}$ (the sign depends on the orientation convention), which for $\theta \neq 0$ is complex. Vicari–Panagopoulos (2009) write 'the complex nature of the θ term in the Euclidean QCD Lagrangian prohibits a direct MC simulation at θ ≠ 0' Known . From here on $\theta = 0$.

05

Spacetime becomes space — boosts become rotations

Putting $t = -i\tau$ and an imaginary rapidity $\phi = i\alpha$ into the boost $t' = t\cosh\phi - x\sinh\phi$, $x' = x\cosh\phi - t\sinh\phi$ of 2-04, from $\cosh i\alpha = \cos\alpha$ and $\sinh i\alpha = i\sin\alpha$,

$$\tau' = \tau\cos\alpha + x\sin\alpha, \qquad x' = x\cos\alpha - \tau\sin\alpha \quad (8)$$

which is a rotation of the (τ, x) plane. The Lorentz group $SO(3,1)$ is replaced by the rotation group $SO(4)$ Known . The interval becomes $\tau^2 + |\underline{x}|^2 \geq 0$, and the light cone and the distinction between past and future disappear. Which direction to call time is chosen afterwards. If reflection positivity (4-10) also holds, then by rotation invariance the same theory comes back whichever direction is chosen as time. The method of carrying each into the other is an achievement of axiomatic field theory Known (§ 3 of the exposition by Jaffe–Witten [2000, 0-03], the official problem statement of the Clay prize problem; hereafter "the exposition").

The lattice reduces rotations to 90°

The rotations left on the lattice are only steps of 90° and reflections, and that $SO(4)$ comes back for $a \to 0$ also has to be shown. For the free field on a four-dimensional lattice ($m = 1$, $a = 1/2n$), we compare the correlations at distance-1 points $(2n, 0, 0, 0)$ and (n, n, n, n). The continuum correlation is $G(r) = \dfrac{m\,K_1(mr)}{4\pi^2 r}$. K_1 is a special function called the modified Bessel function, which decays as e^{-mr} far away (near the origin it behaves like $1/(mr)$, $G \approx 1/(4\pi^2 r^2)$).

Computed Table 2 Ratio of the lattice two-point function to the continuum $G(1) = K_1(1)/4\pi^2$. The lattice side is computed from a single-integral representation using the family of modified Bessel functions I_ν.

a	1/2	1/4	1/8	1/16	1/32
Axis	1.4329	1.1297	1.0247	1.0058	1.0014
Diagonal	0.8126	0.9475	0.9869	0.9967	0.9992

On a coarse lattice the correlation differs by a factor of about 1.8 depending on direction, and the difference vanishes in proportion to a^2. Wilson's lattice (5-01) keeps reflection positivity (4-10) exactly at the price of breaking rotations. The exposition writes that if a regularisation loses reflection positivity, there are few ways to recover it (Jaffe–Witten § 6.5).

```
06
```

Field theory as a probability measure — reading e^{-S_E} as a probability

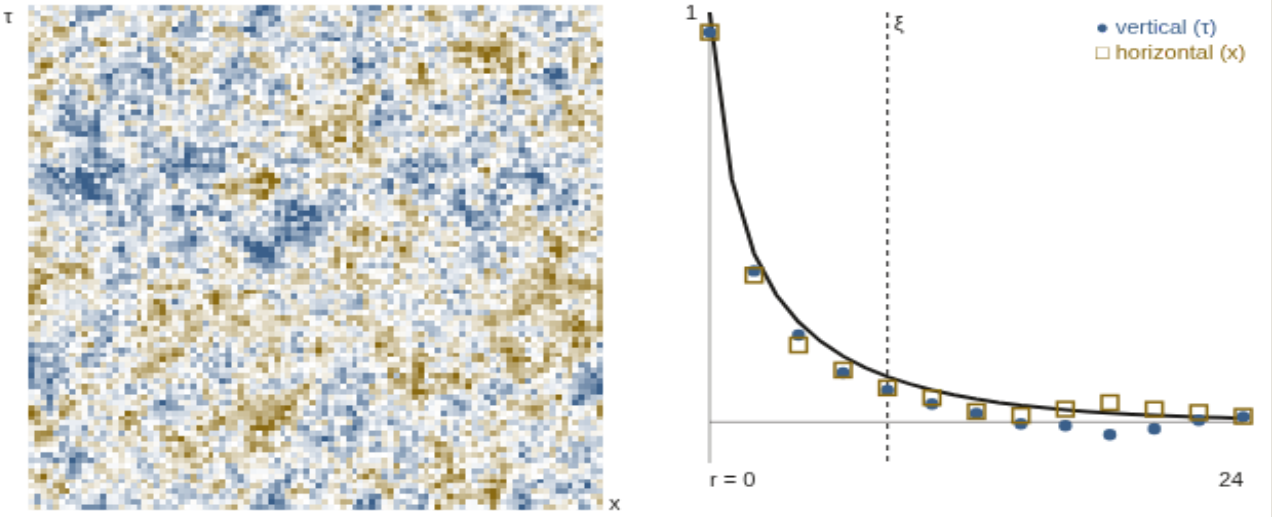
If S_E is bounded below, e^{-S_E}/Z can be read like a probability density on field configurations, and (3) are its moments. Field theory becomes a theory of random functions on four-dimensional space. The exposition attributes this view to Symanzik (Jaffe–Witten § 3) Known .

Precisely, there is no Lebesgue measure $\mathcal{D}\varphi$ in infinite dimensions (1-09), so it is correct as it stands only in finite dimensions. On a periodic lattice $\Lambda_{a,L}$ with spacing a and side L,

$$d\mu_{a,L}(\varphi) = \frac{1}{Z_{a,L}}\, e^{-S_E^{(a)}(\varphi)} \prod_{x\in\Lambda_{a,L}} d\varphi(x) \qquad (9)$$

is a genuine probability measure ($S_E^{(a)}$ is (2) with derivatives replaced by differences and integrals by sums). Making a field theory means showing the limits of the moments as $a \to 0$, $L \to \infty$ and their properties (7-01). The limit of the free field is a Gaussian measure on a space of distributions, and the exposition writes that nonlinear fields need non-Gaussian measures ('one requires non-Gaussian dμ to obtain nonlinear fields') (Jaffe–Witten § 6.2, 1-09).

Figure 2 Changing the correlation length ξ changes a sample of the free field on a two-dimensional periodic lattice (96×96) (left; colour is $\varphi/\sqrt{G(0)}$) and its vertical and horizontal correlations (right, ● and □; the solid line is the exact $G(r)/G(0)$). It is a single sample, so the correlations at points fluctuate ("Redraw" gives another sample; the average over 200 samples is in the text)



ξ = 8 : G(0) = 0.6056、G(8)/G(0) = 0.111 (exact), this sample: vertical 0.078, horizontal 0.083 (a single sample fluctuates; the 200-sample average is in the text)

Computed The numbers in this figure were computed on this machine (random numbers with a fixed seed).

Looking at the samples, one cannot tell which axis is time. In 200 samples made the same way (python3), with $\xi = 8$ (the mass coefficient of the action $M = 2\sinh\frac{1}{2\xi}$,

chosen so that the correlation decays exactly as $e^{-r/\xi}$), the correlation at distance 8 is 0.1120 ± 0.0032 vertically and 0.1112 ± 0.0030 horizontally, against 0.1107 exactly, and for $\xi = 2, 4, 16$ too they agree within twice the standard error Computed . That $G(0)$ grows logarithmically with ξ is a manifestation of the divergence of point values in two dimensions (2-07 § 09).

07

What is missing for gauge fields — the volume of the orbits and compact groups

$S_E[A]$ does not change under gauge transformations (3-02), so e^{-SE} is constant on gauge orbits (the collections of connections carried into one another by gauge transformations, 3-04). An orbit is infinite-dimensional, with one group attached at each point, and the integral in its direction is infinite even formally.

(A) The weight $e^{-(x-y)^2/2}$ does not change under the shift $(x, y) \to (x + c, y + c)$, and the orbits are infinitely long lines. The integral Z_R over the square $[-R, R]^2$ grows in proportion to R, but the average of $(x - y)^2$, which is unchanged by the shift, tends to 1. (B) The orbits of the weight $e^{\beta \cos(\theta_1 - \theta_2)}$ are circles of length 2π, and integrating with the uniform measure on the circle (Haar measure), $Z = (2\pi)^2 I_0(\beta)$ is finite for every β (checked by double integrals and Bessel functions) Computed . Here I_0 is the $\nu = 0$ of the I_ν of Table 2, and

$$I_0(\beta) = \frac{1}{2\pi} \int_0^{2\pi} e^{\beta \cos\theta}\, d\theta$$

Setting $\theta = \theta_1 - \theta_2$, the remaining integral over θ_2 just produces one factor of the orbit length 2π.

Computed Table 3 Toy (A). Z_R agrees with the closed formula to 10 digits.

R	$Z_R/2R$	$\langle (x-y)^2 \rangle_R$
1	1.528	0.4341
10	2.407	0.9584
100	2.497	0.9960
1000	2.506	0.9996

In (A) the denominator and numerator diverge at the same rate, and only the ratio remains ($Z_R/2R \to \sqrt{2\pi} = 2.5066$). Wilson's lattice (5-01) has the form of (B). On a finite lattice the link variables $U \in G$ are integrated with the Haar measure of a compact group, so the integral is finite without fixing the gauge ('the lattice approximation reduces the functional integration to a finite-dimensional integral', Jaffe–Witten § 6.5)

Known . Treating it in the continuum needs gauge fixing (4-08), and there is the obstruction that representatives cannot be chosen globally and continuously (3-04). Either way, only gauge-invariant observables (Wilson loops, $\operatorname{tr} F^2$ and so on) have meaning.

08

Correlations in imaginary time read the energy — the link to the mass gap

If there are a Hamiltonian H (with vacuum energy 0) and a vacuum Ω, the correlation of observables O separated by imaginary time τ, expanded in eigenstates $|n\rangle$, is

$$\langle O(\tau)O(0)\rangle = \langle\Omega|\, O\, e^{-\tau H}\, O\,|\Omega\rangle = \sum_n |\langle n|O|\Omega\rangle|^2\, e^{-E_n\tau} \quad (10)$$

(an integral if the spectrum is continuous). If the vacuum is unique (the only state of energy 0 is Ω), $\langle\Omega|O|\Omega\rangle = 0$, and the spectrum of H has nothing in $(0,\Delta)$, the correlation decays at least as $e^{-\Delta\tau}$. **Example.** Integrating the two-point function of the free field over space leaves only one particle with momentum 0, and the decay rate is exactly the mass Known .

$$\int d^3x\; S_2\big((\tau,\vec{x}),0\big) = \int \frac{dp_4}{2\pi}\,\frac{e^{ip_4\tau}}{p_4^2+m^2} = \frac{e^{-m|\tau|}}{2m} \quad (11)$$

This is how the gap shows on the Euclidean side; the definition is set out in 4-11. On the lattice the weight joining layers at neighbouring times is the transfer matrix $T = e^{-aH}$ (2-07 § 08), and in gauge theory, in the limit where the step in the time direction goes to 0, the Kogut–Susskind Hamiltonian comes out of the gauge transfer matrix of $A_0 = 0$ (abstract of Creutz 1977) Known . The lattice Hamiltonian is treated in 5-03.

Mind the direction

(10) is a formula for when H and the Hilbert space are there first. Building H from the Euclidean measure alone needs positivity with respect to the reflection $\tau \to -\tau$ (reflection positivity) (4-10). The exposition writes 'Establishing a quantum mechanical Hilbert space is part of the solution to this Millennium problem' (Jaffe–Witten § 6.5). "Gap $\Longleftrightarrow$ exponential decay of correlations" is a correspondence after reconstruction. For four-dimensional Yang–Mills, that Hilbert space was built by the construction published in 2026-11 (the committee's assessment is that it is proved in Lean; 12-08). That correlations decay exponentially there has not been shown (6-04).

09

Detour: writing the vacuum wave function as a path integral over half of spacetime

A section you may skip.

The $e^{-\tau H}$ of (10) has another use. Expanding a state in eigenstates of H, $e^{-\tau H}$ multiplies the E_n component by $e^{-E_n \tau}$. Over a long imaginary time, every component other than the lowest-energy one dies away relatively faster. That is, $e^{-\tau H}$ is a filter that extracts only the vacuum from any state that has a vacuum component (2-07 § 05). Precisely, the conditions are that the initial state has a vacuum component and that the vacuum is unique.

Writing this as a path integral, the vacuum wave function itself becomes a path integral. We look at one particle. $\langle x | e^{-TH} | x_i \rangle$ is the sum of e^{-S_E} over all paths that leave x_i at imaginary time $\tau = -T$ and arrive at x at $\tau = 0$. For $T \to \infty$, $\psi_0(x)\,\psi_0(x_i)\,e^{-E_0 T}$ dominates, and as a function of x only the vacuum wave function $\psi_0(x)$ remains.

$$\psi_0(x) \;\propto\; \int_{x(0)=x} \mathcal{D}x\; e^{-S_E[x]}, \qquad S_E[x] = \int_{-\infty}^{0} \left[\tfrac{1}{2}\dot{x}^2 + \tfrac{1}{2}\omega^2 x^2\right] d\tau \quad \text{(12)}$$

What is summed is the paths on the half-line of imaginary time $\tau \le 0$ with only the value at the end $\tau = 0$ fixed.

Example. For the oscillator, the exponent in (12) is quadratic in $x(\cdot)$, so the dependence of the answer on x is fixed by the path of least action (the contribution of fluctuations around it is a constant independent of x). The path that satisfies $\ddot{x} = \omega^2 x$, is x at $\tau = 0$ and is 0 at $\tau \to -\infty$ is $x(\tau) = x\, e^{\omega\tau}$. Its action is

$$S_E = \int_{-\infty}^{0} \omega^2 x^2 e^{2\omega\tau}\, d\tau = \frac{\omega x^2}{2}, \qquad \psi_0(x) \propto e^{-\omega x^2/2}$$

which is exactly the ground state of 2-06 Known .

It can also be checked on a lattice. On an imaginary-time lattice with spacing a, fix $x_0 = x$ and integrate $x_{-1}, x_{-2}, \ldots$ in turn. Each integration preserves the form $e^{-c\,x^2/2}$, and the coefficient c moves by a recurrence and settles to one value. Counting the potential energy at the end point x_0 by the convention of splitting it half and half with the neighbouring half of spacetime, the value it settles to is $c(a) = \omega\sqrt{1 + a^2\omega^2/4}$. At $\omega = 1$, for $a = 1/2$ it is 1.0308 and for $a = 1/10$ 1.00125, and the difference from the continuum $\omega = 1$ vanishes in proportion to a^2 Computed .

For fields

For fields the form is the same. The vacuum wave functional $\Psi_0[\varphi_0]$ (the wave functional of 4-01 § 03) is the sum over all field configurations on the half of spacetime $\tau < 0$ with $\varphi = \varphi_0$ at $\tau = 0$. For the free field each Fourier component is an oscillator of frequency $\omega_p = \sqrt{|p|^2 + m^2}$ (2-06 § 08), so multiplying the oscillator answers component by component,

$$\Psi_0[\varphi_0] \;\propto\; \exp\Big(-\frac{1}{2}\int \frac{d^3p}{(2\pi)^3}\,\omega_p\,|\tilde{\varphi}_0(p)|^2\Big), \qquad \tilde{\varphi}_0(p) = \int d^3x\, e^{-ip\cdot x}\varphi_0(x)$$

Known . The Osterwalder–Schrader reconstruction (4-10) follows this view in reverse. It builds states from quantities fixed by the fields on the half of spacetime $\tau > 0$ alone, and defines the inner product with the reflection $\tau \to -\tau$. What guarantees that the inner product is positive is reflection positivity. Precisely, (12) is a formula for when the Hilbert space and H are there first. For four-dimensional Yang–Mills, the Hilbert space and the Hamiltonian were built by reconstruction in the construction published in 2026-11 (at the level of the committee's assessment; § 08), but a measure on field configurations has not been constructed, and (12) as a path integral over half of spacetime remains formal.

What this chapter can say

	Content
What is said	The manipulations of (2), (5)–(8) and (11), the rotation of the contour in (5), that the lattice (9) is a probability measure, that for compact groups the lattice integral is finite Known . Tables 1–3, Figures 1–2, the numbers in § 01, § 06 and § 09 Computed
Conditional	(10) and (12) are formulas for when the Hilbert space and the Hamiltonian are there first and the vacuum is unique (the conditions for going back are in 4-10)
What is not said	The $\mathcal{D}\varphi$ of (1) and (3) is formal and not a measure. The existence of the measure of four-dimensional Yang–Mills has not been shown (6-04). The restoration of rotational symmetry and a continuum limit preserving reflection positivity were shown at the level of correlation functions by the construction published in 2026-11 (at the level of the committee's assessment; 12-01, 12-06). Nothing is shown about the mass gap

Sources and reproduction

Item	Kind	Source / tools
Quotations from the exposition (§ 3, § 6.2, § 6.5)	Text	A. Jaffe, E. Witten, *Quantum Yang–Mills Theory*, Clay Math. Inst. (2000)
That the θ term is complex	Text § 1	E. Vicari, H. Panagopoulos, Phys. Rept. 470 (2009) 93, arXiv:0803.1593
The transfer matrix and $A_0 = 0$	Abstract	M. Creutz, Phys. Rev. D 15 (1977) 1128–1136
The names of Wick and Schwinger	Bibliographic data only	G. C. Wick, Phys. Rev. 96 (1954) 1124–1134 / J. Schwinger, PNAS 44 (1958) 956–965 (origin of the names not checked)

Item	Kind	Source / tools
The correspondence between Euclidean and Minkowski	Bibliographic data only	K. Osterwalder, R. Schrader, CMP 31 (1973) 83–112, 42 (1975) 281–305 / K. Osterwalder, E. Seiler, Ann. Phys. 110 (1978) 440–471
The definition of the gap and the Hilbert space	Text	Jaffe–Witten § 4, § 6.5
Table 1, Figure 1	Computed	python3 (SciPy; agreeing with the closed form and 3×10^{-13}). The figure is JavaScript and agrees with the table
Table 2	Computed	python3 (SciPy). Checked against the value at the origin 0.154933 (the same as 1-09) and the FFT of a periodic lattice of 40^4 (agreeing to 7 digits at $a = 1/2$)
$c(a)$ of § 09	Computed	python3 (the recurrence and direct computation of the 400-point quadratic form agree to 10 digits; numerical integration of the action of the classical solution)
Figure 2 and § 06, Table 3 and § 01	Computed	python3 (NumPy FFT, 200 samples with a fixed seed / SciPy / counting). The figure is one sample made the same way

Next chapter to read: 4-03 Perturbation theory and Feynman diagrams. The Euclidean propagator of (5) becomes the lines. The conditions for going back to physical spacetime are in 4-10.

Perturbation theory and Feynman diagrams — expand around a Gaussian integral, and the terms become diagrams

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That each term of the expansion becomes a sum over the ways of pairing in a Gaussian integral, which can be counted with diagrams of lines and vertices.
Symmetry factors, connected diagrams, loops, and that the number of loops counts powers of $\hbar$.
That Yang–Mills needs gauge fixing, and three-point and four-point vertices and ghosts appear.

Prerequisite chapters — 4-02. For counting pairings, 1-08; for the propagator, 1-11; for the cubic and quartic terms, 3-02. $\hbar = c = 1$ (restored only in § 06).

Said first — Perturbation theory is a method of building a power series term by term, and gives neither its sum nor the existence of the theory. The integrals for the coefficients diverge (4-04), and the series is also thought to diverge (4-12). The mass gap is thought not to show at any order (4-07), and this chapter shows nothing about it.

Order of this chapter

1. The idea of perturbation theory — expanding around a Gaussian
2. Drawing pairings as lines — symmetry factors (Figure 1, Table 1) / Detour: 1949
3. Connected diagrams — the logarithm is connected diagrams (Table 2)
4. To fields — lines are propagators
5. Writing in momentum — conservation laws and loops
6. The number of loops is the power of ħ — trees are classical
7. The Feynman rules of Yang–Mills — gauge fixing is needed
8. The perturbation series — an asymptotic series (Table 3)

9. What perturbation theory says, and what it does not

01

The idea of perturbation theory — expanding around a Gaussian integral

What can be computed exactly in the Euclidean path integral ((3) of 4-02) is the case where the action is quadratic in the field, that is, a Gaussian integral (2-07 § 02). So we split the action into the quadratic part S_0 and the interaction S_{int}, and expand $e^{-S_{\text{int}}}$ ($\langle\cdot\rangle_0$ is the expectation with S_0 alone).

$$\langle F\rangle = \frac{\langle F\, e^{-S_{\text{int}}}\rangle_0}{\langle e^{-S_{\text{int}}}\rangle_0}, \qquad \langle F\, e^{-S_{\text{int}}}\rangle_0 \overset{\text{formal}}{=} \sum_{n\geq 0} \frac{(-1)^n}{n!}\,\langle F\, S_{\text{int}}^n\rangle_0 \quad (1)$$

Exchanging the sum and the integral is formal (§ 08). **Example.** In place of a field take one real number x, and let $S = \frac{1}{2}x^2 + \frac{u}{4!}x^4$ ($u \geq 0$; the same convention as 4-04). It is a "zero-dimensional field theory" whose spacetime is a single point. From $\langle x^{2k}\rangle_0 = (2k-1)!!$ (1-08 § 06),

$$Z(u) = \int \frac{dx}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2 - \frac{u}{4!}x^4} \;\sim\; \sum_{n\geq 0} \frac{(4n-1)!!}{n!}\left(-\frac{u}{4!}\right)^n = 1 - \frac{u}{8} + \frac{35}{384}u^2 - \frac{385}{3072}u^3 + \cdots$$

(the meaning of $\sim$ is in § 08; the coefficients computed as fractions Computed). All the values of the diagrams are numbers, so the bookkeeping can be checked by hand, and in four dimensions the shapes of the diagrams and the bookkeeping of weights are the same. Each term of the expansion can be written only with averages of polynomials under the Gaussian distribution, that is, moments of the normal distribution. Not one hard integral remains; all that remains is counting.

02

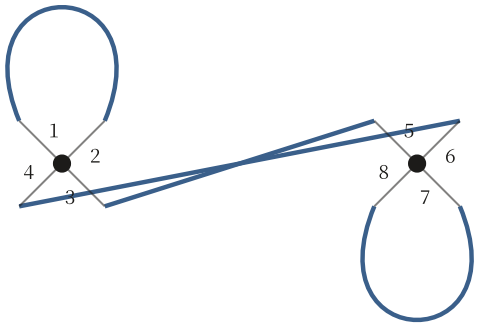
Drawing pairings as lines — from Wick's theorem to diagrams

The term of order n in (2) was the number of ways of pairing up $4n$ copies of x two by two (Isserlis' formula, in physics Wick's theorem; (9) of 1-08). Draw each $\frac{u}{4!}x^4$ as a point (vertex) with 4 legs, and join paired legs by lines. Pairings that give the same shape after joining have the same value, so they are grouped by shape. These are Feynman diagrams.

$$\frac{\#\{\text{pairings giving the shape } \Gamma\}}{n!\,(4!)^n} = \frac{1}{S(\Gamma)} \quad (3)$$

The denominator is the number of permutations of vertices and legs, and usually exactly this many pairings make one shape. If the shape has a symmetry, some of the permutations give back the same pairing and the number drops; that ratio is the symmetry factor $S(\Gamma)$, precisely the number of automorphisms of the diagram (exchanges of vertices and reattachments of legs that do not change how the lines are connected) Known . **Example.** With one vertex there are 3 pairings, all of them a "figure eight". $3/4! = 1/8$ is the coefficient of u in (2), and with the exchange of the loops and the reversal of each loop, $S = 2 \cdot 2 \cdot 2$.

Figure 1 Choose a case and move the pairing number with the knob, and the pairing of legs (blue lines) and the shape of the diagram it makes change. "Count all" counts by shape



`Pairing 4 / 105 : 「bubble with a loop on each side", pairings of this shape: 72, 1/S = 1/16`

Computed The counting in this figure is done in the browser and has been checked to agree with Table 1. The numbers are the legs of the vertices; x, y and 1–4 are external points (the x put in from outside).

Computed Table 1 All pairings generated and sorted into shapes (isomorphism classes) (python3). In brackets: (number of pairings, $1/S$)

Case	Pairings	Shapes
Vacuum, 1 vertex	3	figure eight (3, 1/8)
Vacuum, 2 vertices	105	bubble with a loop on each side (72, 1/16), joined by four lines (24, 1/48), two figure eights (9, 1/128)
Two-point, 1 vertex	15	line with a loop (12, 1/2), line and a figure eight (3, 1/8)
Two-point, 2 vertices	945	loop on a loop (288, 1/4), two loops in a row (288, 1/4), joined by three lines (192, 1/6), and those containing bubbles: line and bubble with a loop on each side (72, 1/16), line with a loop and a figure eight (72, 1/16), line and joined by four lines (24, 1/48), line and two figure eights (9, 1/128)
Four-point, 1 vertex	105	$\times$ (24, 1), line and line with a loop (6 ways $\times$ 12, each 1/2), two lines and a figure eight (3 ways $\times$ 3, each 1/8)

Detour: the 1949 papers and a theorem the following year

A subsection you may skip.

The abstract of Feynman (1949) says it presents a viewpoint from which matrix elements can be 'written down directly for any specific problem', and in the same year Dyson (1949) showed the equivalence of the theories of Feynman and Schwinger. The next year Wick (1950) proved an algebraic theorem that simplifies the reduction to 'a sum of "graph" terms'. The formula itself is found in Isserlis (1918), in statistics Known .

03

Connected diagrams — bubbles vanish by division, and the logarithm counts connected diagrams

Table 1 also includes diagrams containing parts not connected to the external points (vacuum bubbles). The value of a diagram made of separate parts is the product of the values of the parts, and the denominator Z of (1) is the sum of diagrams made only of bubbles, so bubbles vanish by division. Moreover, the sum of all diagrams is the exponential of the sum of connected diagrams. When k copies of the same connected diagram C appear side by side, the exchanges $k!$ are added to S, giving $\sum_k C^k/k! = e^C$ Known . In zero dimensions the value of a line is 1, so the value of a diagram is $(-u)^n/S$.

$$\log Z = \sum_{\text{connected vacuum diagrams}} \frac{(-u)^n}{S}, \qquad \langle x^2 \rangle = \sum_{\text{connected two-point diagrams}} \frac{(-u)}{S}$$

Example. For u^2, the $\frac{1}{16} + \frac{1}{48} = \frac{1}{12}$ of the connected vacuum diagrams agrees with the coefficient $\frac{35}{384} - \frac{1}{2}\cdot\frac{1}{64} = \frac{1}{12}$ of $\log Z$, and the $\frac{1}{4} + \frac{1}{4} + \frac{1}{6} = \frac{2}{3}$ of the connected two-point diagrams is the coefficient of $\langle x^2 \rangle$. Read the other way, a single integral is a generating function that counts diagrams, which leads to the counting of diagrams on surfaces (Bessis–Itzykson–Zuber 1980, 17-04).

Computed Table 2 Sums of $(-1)^n/S$ for connected shapes counted from all pairings (python3). All agree with the coefficients of the series. In brackets, the coefficient from the series (the two-point $n = 4$ was not counted)

n	Vacuum pairings	Shapes (connected)	$\log Z$	Two-point pairings	$\langle x^2 \rangle$
1	3	1 (1)	-1/8	15	-1/2
2	105	3 (2)	1/12	945	2/3
3	10,395	7 (4)	-11/96	135,135	-11/8
4	2,027,025	20 (10)	17/72	—	(34/9)

04

To fields — lines are propagators, vertices are integrals over position

For a four-dimensional field the procedure is the same. The quadratic part of the d-dimensional action

$$S_E[\phi] = \int d^d x \left[\tfrac{1}{2} (\partial_\mu \phi)^2 + \tfrac{1}{2} m^2 \phi^2 + \tfrac{u}{4!} \phi^4 \right] \quad (5)$$

is a Gaussian measure with covariance $G = (-\Delta + m^2)^{-1}$ (Δ is the Laplacian $\sum_\mu \partial_\mu^2$; not the Δ of the mass gap) (4-02 § 06), and Wick's theorem can be used as it is. Two things change. The value of a line becomes the propagator $G(x - y)$ (the degree to which the fluctuation of the field at the point x reaches the point y; it becomes smaller with distance), and vertices can be placed anywhere in spacetime, so their positions z are integrated over.

$$\langle \phi(x) \phi(y) \rangle = G(x - y) - \frac{u}{2} \int d^d z \, G(x - z) \, G(0) \, G(z - y) + O(u^2) \quad (6)$$

The second term is the "line with a loop"; the number 1 for $\langle x^2 \rangle = 1 - \frac{u}{2} + \cdots$ in zero dimensions has simply been replaced by a propagator. The lowest-order four-point "$\times$" is $-u \int d^d z \prod_i G(x_i - z)$, which can be read as four particles meeting at one point. Precisely, it means integrating the position of that point over all space. **Note.** $G(0)$ is infinite at $d \geq 2$ (1-11), and it diverges from first order. Its treatment is 4-04, and the $T(\Lambda)$ there is this loop.

05

Writing in momentum — conservation laws, and the loop integrals that remain

Putting in $G(x) = \int \frac{d^d p}{(2\pi)^d} \frac{e^{ip \cdot x}}{p^2 + m^2}$ (the inverse Fourier transform of the $\hat{G}$ of (10) in 1-11; the rotation is (5) of 4-02) and integrating over the positions of the vertices gives $\int d^d z \, e^{i(p_1 + \cdots + p_4) \cdot z} = (2\pi)^d \delta^{(d)}(p_1 + \cdots + p_4)$: momentum is conserved at each vertex. The δ here is the relation of 1-11 that integrating a plane wave over all space concentrates the value only where the wave number is 0. The integral over positions turns into a condition on adding momenta. There are V conditions for I internal lines, and one of them is overall conservation, involving only the external lines, so the number of momenta left free in a connected diagram is

$$L = I - V + 1 \quad (7)$$

which is the number of independent cycles (loops) of the diagram Known . The rules are $1/(p^2 + m^2)$ for each line, $-u$ for each vertex, $\int d^d k / (2\pi)^d$ for each loop, and $1/S$

overall. **Example.** The one-loop four-point "bubble" has $I = V = 2$, $L = 1$, and with momentum P flowing in,

$$\frac{u^2}{2}\int \frac{d^dk}{(2\pi)^d}\,\frac{1}{(k^2+m^2)\big((k+P)^2+m^2\big)} \tag{8}$$

The two-point "joined by three lines" diagram has $I = 3$, $V = 2$, $L = 2$, and is a double integral. (8) behaves like $k^{d-1}dk/k^4$ for large $|k|$, and diverges logarithmically at $d = 4$. Adding the three bubbles and changing the sign gives the one-loop term of (4) in 4-04.

06

The number of loops is the power of $\hbar$ — trees are classical, loops are quantum corrections

Restoring the weight to $e^{-SE/\hbar}$, propagators are multiplied by $\hbar$ and vertices by $1/\hbar$. For a connected diagram with I' lines including those going to external points, V vertices and E external points, counting external points as points too, $L = I' - V - E + 1$, so

$$\hbar^{I'-V} = \hbar^{L+E-1} \tag{9}$$

Fixing E, the power of $\hbar$ is fixed by the number of loops (Srednicki § 21) Known . The lowest order is diagrams without loops (trees), and the abstract of Boulware–Brown (1968) writes that the classical field made by a source is 'the generating functional of the tree-graph approximation' Known .

Example. The classical solution of $\frac{1}{2}x^2 + \frac{u}{4!}x^4 - Jx$, zero dimensions with a source added, is $x_{\mathrm{cl}} = J - \frac{u}{6}J^3 + \frac{u^2}{12}J^5 - \frac{u^3}{18}J^7 + \cdots$. In general it can be written $x_{\mathrm{cl}} = \sum_k (-1)^k T_k\,(u/6)^k J^{2k+1}$, where $T_k = 1, 1, 3, 12, 55, \ldots$ is the number of ternary trees. Integrating numerically at $u = 1$, $J = 0.8$, $(\langle x\rangle - x_{\mathrm{cl}})/\hbar$ approaches -0.2167, -0.2266, -0.2277 at $\hbar = 0.1,\ 0.01,\ 0.001$, the value -0.2278 of the one-loop diagram (a line with a loop around x_{cl}) Computed . In Yang–Mills, rescaling as $\hat{A} = gA$ makes the action $\frac{1}{2g^2}\int \mathrm{tr}\,\hat{F}^2$ ((7) of 4-02), which enters the weight only in the combination $g^2\hbar$, so the loop expansion is an expansion in g^2 (3-03).

07

The Feynman rules of Yang–Mills — a quadratic form that cannot be inverted, and three-point and four-point vertices

The conventions are the same as in 3-02: g is written split, $F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g f^{abc} A^b_\mu A^c_\nu$, $S_E = \frac{1}{4}\int F^a_{\mu\nu} F^a_{\mu\nu}\, d^4x$. Split into quadratic, cubic and quartic parts ((11) of 3-02 read in Euclidean form as in 4-02 § 04; all indices are lower, and repeated indices are summed over 1–4),

$$S_E = \frac{1}{2}\int A^a_\mu\left(-\partial^2\delta_{\mu\nu} + \partial_\mu\partial_\nu\right)A^a_\nu + g f^{abc}\int (\partial_\mu A^a_\nu) A^b_\mu A^c_\nu + \frac{g^2}{4} f^{abe} f^{cde}\int A^a_\mu A^b_\nu A^c_\mu A^d_\nu$$

The quadratic kernel is $K_{\mu\nu} = p^2\delta_{\mu\nu} - p_\mu p_\nu$ in momentum, and sends p_μ to 0. This is because $A_\mu = \partial_\mu\lambda$ is the direction of infinitesimal gauge transformations, which do not change the action (4-02 § 07). There is no inverse matrix, so no propagator can be made. Adding the gauge-fixing term $\frac{1}{2\alpha}\int(\partial_\mu A^a_\mu)^2$ (justified in 4-08),

$$K^{(\alpha)}_{\mu\nu} = p^2\delta_{\mu\nu} - \left(1 - \frac{1}{\alpha}\right)p_\mu p_\nu, \qquad D^{ab}_{\mu\nu}(p) = \frac{\delta^{ab}}{p^2}\left(\delta_{\mu\nu} - (1-\alpha)\frac{p_\mu p_\nu}{p^2}\right) \quad (11)$$

and for $\alpha = 1$ (the Feynman gauge), $\delta^{ab}\delta_{\mu\nu}/p^2$ (the same letter as in 4-08). The passage to momentum is $A^a_\mu(x) = \int \frac{d^4p}{(2\pi)^4} e^{ip\cdot x}\tilde{A}^a_\mu(p)$ (derivatives become ip_μ). A vertex is the term of order n of the action differentiated n times with respect to $\tilde{A}$, with the sign changed (excluding $(2\pi)^4\delta(\sum p)$); the sign of e^{-S} enters here (all momenta flowing in). Taking the opposite direction of Fourier transform reverses the sign of i in (12).

$$V^{abc}_{\mu\nu\lambda}(p,q,r) = ig\, f^{abc}\left[\delta_{\mu\nu}(p-q)_\lambda + \delta_{\nu\lambda}(q-r)_\mu + \delta_{\lambda\mu}(r-p)_\nu\right] \quad (12)$$

$$V^{abcd}_{\mu\nu\lambda\rho} = -g^2\left[f^{abe}f^{cde}(\delta_{\mu\lambda}\delta_{\nu\rho} - \delta_{\mu\rho}\delta_{\nu\lambda}) + f^{ace}f^{bde}(\delta_{\mu\nu}\delta_{\lambda\rho} - \delta_{\mu\rho}\delta_{\nu\lambda}) + f^{ade}f^{bce}(\delta_{\mu}\right.$$

The three-point vertex is proportional to momentum, the four-point vertex does not depend on momentum, and both contain f^{abc}. **Example.** Putting plane waves into (10), extracting the coefficients directly, and comparing with (11)–(13) for 20 random sets of momenta, polarisations and colours in SU(2) and SU(3), the relative error was at most 8×10^{-13} Computed .

Ghosts. The abstract of Faddeev–Popov (1967) writes 'Feynman and De Witt showed, that the rules must be changed for ... diagrams with closed loops', and proposed a method of obtaining the rules for arbitrary diagrams from the path integral. The result is the rule of adding ghosts with propagator δ^{ab}/p^2 and a vertex with two ghost lines and one gauge field line, and multiplying by -1 for each closed ghost loop (Srednicki § 71) Known . -1 comes from ghosts being anticommuting numbers (4-08). The derivation, and the limits of the local gauge fixing it presupposes (3-04), are in 4-08.

08

The perturbation series — the coefficients grow factorially, and it must be cut off somewhere

The ratio of the coefficients a_n of (2) is $|a_{n+1}/a_n| = \frac{(4n+1)(4n+3)}{24(n+1)} \approx \frac{2n}{3}$, so a_n grows at a rate of roughly $n!\,(2/3)^n$, and the series does not converge for any $u > 0$. That the integral (2) diverges for $u < 0$ shows the same circumstance. The argument of this form in quantum electrodynamics was made by Dyson (1952), whose abstract qualifies the conclusion as 'tentatively' (4-12). Even so, putting $s = \frac{u}{4!}x^4$ into the Taylor remainder $\left|e^{-s} - \sum_{n<N}(-s)^n/n!\right| \le s^N/N!$ $(s \ge 0)$ and integrating gives

$$\left| Z(u) - \sum_{n<N} a_n u^n \right| \le |a_N|\, u^N \qquad (u \ge 0) \tag{14}$$

Known . Fixing N and letting $u \to 0$, the error vanishes at the rate u^N. This is an asymptotic series, and it is the meaning of the $\sim$ in (2). Fixing u and increasing N, the error first falls and then rises, and the best is around $N \approx 3/(2u)$, where the terms are smallest (the valley of term sizes can be moved in Figure 1 of 4-12).

Computed Table 3 Truncation of $\langle x^2\rangle = 1 - \frac{1}{2}u + \frac{2}{3}u^2 - \frac{11}{8}u^3 + \frac{34}{9}u^4 - \cdots$ (python3, $N \le 16$)

u	Exact $\langle x^2\rangle$	Best N	Least error	Error at N=10
0.1	0.955577	14	1.3e−6	2.3e−6
0.2	0.919366	6	1.6e−3	3.2e−3
0.5	0.838595	2	7.8e−2	41
1.0	0.750511	0	0.25	4.7e4

For $u = 1$, adding nothing, $N = 0$, is best. (14) also held numerically for $u \le 1$, $N \le 16$ Computed . For field theory, the problem statement writes that in many known examples ('In most known examples') perturbation series are 'divergent expansions' (Jaffe–Witten § 6.1), and that in the two-dimensional models that have been constructed the scattering matrix can be expanded as 'an asymptotic series in the coupling constants' (§ 6.2). For four-dimensional Yang–Mills, the construction published in 2026-11 built the theory (at the level of the committee's assessment; 6-00), but agreement with perturbation theory is claimed only within a range such as the one-loop running of the coupling, and it has not been shown that the series is an asymptotic expansion of that theory (6-04).

09

What perturbation theory says, and what it does not

The coefficient at each order is a finite sum of diagram integrals, and that each order is finite after renormalisation is a theorem, gauge theories included (4-04 § 08, 4-05) Known . **The sum of the series** has no meaning in general. There are also **things that do not show in perturbation theory**. The propagator (11) has a pole at $p = 0$ and represents a massless particle. If there is a mass gap, it is thought to come in a form that looks 0 at every order in g^2, like $\Lambda \propto e^{-1/(2b_0 g^2)}$ (4-07) Physics . § 6.6 of the problem statement also writes 'the mass gap is not classically visible'.

Even so, perturbation theory is part of the conditions of the problem statement. § 4 requires that the correlation functions of the constructed theory agree at short distances with asymptotic freedom and the predictions of perturbative renormalisation theory (quoted in 0-03 § 07). Perturbation theory is not a method of making the theory but a mark that the theory made must meet. Asymptotic freedom is 4-06, and large N is 17-04.

What this chapter can say

	Content
What is said	The expansion of (1)(2)(4) and how to count diagrams, the symmetry factor of (3), the loop counting of (7)(9), the form of the Feynman rules (11)–(13), and (14) in zero dimensions Known . Tables 1–3, Figure 1, the numbers in § 06 and § 07 Computed
Conditional	The values of field diagrams such as (6)(8) have meaning only once a cutoff is put in (4-04). The ghost rules are rules under local gauge fixing (4-08). The view that the mass gap comes in the form e^{-1/g^2} is an expectation of physics Physics
What is not said	The sum of the perturbation series, an estimate corresponding to (14) in field theory, the existence of four-dimensional Yang–Mills, the mass gap (this site shows none of them)

Sources and reproduction

Item	Kind	Source / tools
Quotations from the problem statement (§ 4, § 6.1, § 6.2, § 6.6)	Text	A. Jaffe, E. Witten, *Quantum Yang–Mills Theory*, Clay Math. Inst. (2000)
Counting of $\hbar$, the -1 of ghosts	Text § 21, § 71	M. Srednicki, *Quantum Field Theory*, Cambridge UP (2007) (the author's public draft)
The beginnings of the diagram method	Abstract	R. P. Feynman, Phys. Rev. 76 (1949) 769 / F. J. Dyson, Phys. Rev. 75 (1949) 486 / G. C. Wick, Phys. Rev. 80 (1950) 268

Item	Kind	Source / tools
The formula for Gaussian moments	Bibliographic data only	L. Isserlis, Biometrika 12 (1918) 134–139
The ghost rules	Abstract	L. D. Faddeev, V. N. Popov, Phys. Lett. B 25 (1967) 29–30. The earlier Feynman, Acta Phys. Polon. 24 (1963) 697 and DeWitt, Phys. Rev. 162 (1967) 1195, 1239 bibliographic data only
Tree diagrams and the classical / counting diagrams / divergence of the series	Abstract	D. G. Boulware, L. S. Brown, Phys. Rev. 172 (1968) 1628 / D. Bessis, C. Itzykson, J.-B. Zuber, Adv. Appl. Math. 1 (1980) 109 / F. J. Dyson, Phys. Rev. 85 (1952) 631
Tables 1 and 2, Figure 1	Computed	python3 (all pairings generated and sorted into isomorphism classes by permutations; vacuum up to the 2,027,025 pairings at order 4). The figure's JS uses the same classification, checked to agree with node
The comparison in § 07	Computed	python3 (NumPy; coefficients extracted by putting in plane waves)
Table 3, the numbers in § 06	Computed	python3 (coefficients as fractions; exact values by SciPy numerical integration, Z also agreeing with the closed formula with the Bessel function $K_{1/4}$)

Next chapter to read: 4-04 Divergences and renormalisation. The $G(0)$ of (6) and the logarithmic divergence of (8) are rewritten in terms of the measured coupling.

Divergences and renormalisation — rewriting infinities in terms of measured quantities

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — That loop integrals diverge because they add up too many short-distance fluctuations. That the types of divergence can be found by counting dimensions, and are finitely many for ϕ^4 in four dimensions and for Yang–Mills. That absorbing the divergences into measured couplings makes predictions independent of the cutoff, and that those couplings run with the scale.

Prerequisite chapters — 4-03 (propagators, vertices, loops). Helpful: 1-11, 2-08. $\hbar = c = 1$, Euclidean momenta. β is the β function (not the inverse coupling of the lattice).

Said first — This is a chapter of perturbative calculations and how to read them. Renormalisability is a theorem at the level of each order of perturbation theory (formal power series), and gives neither the sum of the series nor the existence of the four-dimensional theory. This site shows neither the construction of the continuum limit nor the mass gap.

Order of this chapter

1. Where the infinities come from — adding up too many short-distance fluctuations
2. Putting in a cutoff — regularisation, and its price (Table 1)
3. Counting dimensions — the superficial degree of divergence (Table 2)
4. Counterterms — rewriting the bare coupling in terms of the measured coupling (Figure 1)
5. The running coupling and the renormalisation group equation
6. Wilson's view — relevant, irrelevant, marginal (Figure 2) / Detour: the Fermi interaction
7. In gauge theories — subtracting without breaking the symmetry

8. Renormalisation seen from mathematics — where the theorems are, and where they are not yet (Table 3)

01

Where the infinities come from — adding up too many short-distance fluctuations

When a perturbative diagram (4-03) has a loop, an integral over a momentum k not fixed from outside remains. Large $|k|$ are short-distance fluctuations, and in four dimensions there are too many of them for convergence. We cut at $|k| < \Lambda$ (Λ is the **cutoff**) and look at how it grows.

$$T(\Lambda) = \int_{|k|<\Lambda} \frac{d^4k}{(2\pi)^4} \frac{1}{k^2+m^2} = \frac{1}{16\pi^2}\left[\Lambda^2 - m^2 \ln\left(1+\frac{\Lambda^2}{m^2}\right)\right]$$
$$I(\Lambda) = \int_{|k|<\Lambda} \frac{d^4k}{(2\pi)^4} \frac{1}{(k^2+m^2)^2} = \frac{1}{16\pi^2}\left[\ln\left(1+\frac{\Lambda^2}{m^2}\right) - \frac{\Lambda^2}{\Lambda^2+m^2}\right] \quad (1)$$

T is the diagram closing one propagator, and I the diagram closed with two (a bubble) ($d^4k = 2\pi^2 k^3 dk$). Multiplying the volume $k^3 dk$ by k^{-2} for each propagator, T is $k\,dk$ (quadratic divergence) and I is $k^{-1}dk$ (logarithmic divergence). On the position side this corresponds to $G(x)^2 \sim |x|^{-4}$ not being integrable at the origin (1-11).

Example. With $m = 1$ and $\Lambda = 10, 10^3, 10^6$, $16\pi^2 T$ is 95.4, 1.0×10^6, 1.0×10^{12}, and $16\pi^2 I$ is 3.625, 12.816, 26.631. Each time Λ is multiplied by 10, T grows about 100-fold and $16\pi^2 I$ grows by $2\ln 10 = 4.605$ Computed .

02

Putting in a cutoff — regularisation, and its price

Putting in an artificial scale to make things finite is called **regularisation**. Regularisation is scaffolding, and the goal is that as $\Lambda \to \infty$ (on a lattice, $a \to 0$) the answer does not depend on the scaffolding. Each of the four in Table 1 loses something.

Table 1 Typical regularisations

Regularisation	Keeps	Loses
Momentum cutoff	Rotations	Gauge invariance (§ 07), reflection positivity
Lattice (5-01, $\Lambda \sim \pi/a$)	Gauge invariance, reflection positivity (5-03)	Continuous rotations (4-02)

Regularisation	Keeps	Loses
Pauli–Villars (subtracting heavy fields)	Rotations	Reflection positivity (4-10) (in the non-abelian case also gauge invariance; § 07)
Dimensional regularisation ($d = 4 - \epsilon$)	Rotations, gauge invariance	Reflection positivity (4-05)

Dimensional regularisation (Bollini–Giambiagi 1972, 't Hooft–Veltman 1972) is the standard tool of perturbation theory, Yang–Mills included. On the other hand, in Table 1 only the lattice row keeps reflection positivity. Reflection positivity is the property needed to assemble the Hilbert space of quantum mechanics (the space of states) from Euclidean correlation functions (4-10). The problem statement writes that there are few ways to recover reflection positivity once lost by regularisation, giving dimensional regularisation and Pauli–Villars as examples (§ 6.5; quoted in 0-03 § 03). Being able to keep from the start a property that is hard to recover once lost is one of the reasons the lattice is chosen on the side of construction.

03

Counting dimensions — the superficial degree of divergence

With $\hbar = c = 1$, the dimension $[\cdot]$ of a quantity is a power of mass (length is -1). The action is dimensionless, $d^d x$ is $-d$, and derivatives are $+1$, so the dimensions of fields and couplings are fixed.

$$\int d^d x \Big[\tfrac{1}{2}(\partial\phi)^2 + \tfrac{u}{4!}\phi^4\Big] : \ [\phi] = \tfrac{d-2}{2}, \ [u] = 4 - d; \qquad \frac{1}{2g^2}\int d^d x \, \mathrm{tr}\, F_{\mu\nu}F_{\mu\nu} : \ [A] = 1, \ [g^2]$$

In four dimensions both u and g^2 are dimensionless. This is what "four dimensions is marginal" means (0-03, 1-12; how to derive the counting of dimensions is (10) of 2-04 § 08). The power of momentum of a diagram with L loops and I internal lines is $\omega = dL - 2I$ (adding the number of vertices that carry derivatives), and if $\omega \geq 0$ it superficially diverges. For n vertices of ϕ^4 and E external lines, $4n = 2I + E$, and from $L = I - n + 1$,

$$\omega_{\phi^4} = d - \frac{d-2}{2}E - [u]\,n, \qquad \omega_{\mathrm{YM}} = d - \frac{d-2}{2}E - \frac{[g^2]}{2}(V_3 + 2V_4) \quad (3)$$

(V_3, V_4 is the number of three-point and four-point vertices of Yang–Mills, and the external lines are gluons). In four dimensions $\omega_{\mathrm{YM}} = 4 - E$. The dimension of the coupling appears in the coefficient of the order, so there are three cases Known . If positive, ω falls with the order, and only finitely many diagrams diverge (**super-renormalisable**). If 0, the types of divergence (values of E) are finitely many but appear at all orders (**renormalisable**). If negative, the types increase without bound (**non-renormalisable**).

Computed Table 2 Types of loop diagrams with $\omega \geq 0$ (excluding $E = 0$, up to order 30; enumerated from combinations of vertices and external lines, agreeing with (3); for Yang–Mills $E = 1$ is 0 by Lorentz symmetry and colour symmetry)

Theory	Dimension of the coupling	E (ω)	Orders	Classification
ϕ^4, $d = 2$	2	2 (0)	1 only	Super-renormalisable
ϕ^4, $d = 3$	1	2 (1, 0)	1, 2	Super-renormalisable
ϕ^4, $d = 4$	0	2 (2), 4 (0)	All	Renormalisable
ϕ^4, $d = 6$	−2	2, 4, 6, 8, ⋯	All	Non-renormalisable
Yang–Mills, $d = 4$	0	1 (3), 2 (2), 3 (1), 4 (0)	All	Renormalisable

The two-dimensional type has the same form as T and can be removed by Wick ordering (7-03; for three-dimensional Yang–Mills, 7-06). We say "superficial" because even with $\omega < 0$ a diagram diverges if it has a $\omega \geq 0$ subdiagram inside. The theorem that states convergence by a power condition including subdiagrams is Weinberg's 1960 theorem, and the method of subtracting subdiagrams in turn is treated in 4-05.

04

Counterterms — rewriting the bare coupling in terms of the measured coupling

We follow the procedure at one loop for ϕ^4 in four dimensions. The coefficients m_0, u_0 of the action are called the **bare** mass and coupling (coefficients under the cutoff, not directly measurable). The four-point strength Γ^4 (not the diagram symbol Γ of 4-03; the four-point function with the external propagators removed, summing only diagrams that do not split in two when one line is cut; u_0 at tree level) shifts by the amount of one inserted bubble.

$$\Gamma^4 = u_0 - \frac{u_0^2}{2} \sum_{j=1}^{3} B(P_j; \Lambda) + O(u_0^3), \qquad P_1 = p_1 + p_2,\ P_2 = p_1 + p_3,\ P_3 = p_1 + p_4 \tag{4}$$

The three terms are the three ways of dividing the external lines into pairs, and $\frac{1}{2}$ is the symmetry factor (counted the same way as 7-08). The bubble with momentum P flowing in reduces to the I of (1) by $\frac{1}{AB} = \int_0^1 \frac{dx}{[xA + (1-x)B]^2}$ (a Feynman parameter) and a shift of variable.

$$B(P; \Lambda) = \int \frac{d^4k}{(2\pi)^4} \frac{1}{(k^2 + m^2)\big((k+P)^2 + m^2\big)} = \int_0^1 I\big(\Lambda;\, m^2 \to m^2 + x(1-x)P^2\big)\, dx \tag{5}$$

The difference between the value cut in the shifted variable and the value cut in the original k is 0 as $\Lambda \to \infty$ ($P = m$, $\Lambda = 10^3 m$, multiplied by $16\pi^2$, 3×10^{-7}) Computed . So we redefine the coupling by a measured quantity. The Γ^4 measured in the configuration where all three $|P_j|$ are μ is taken as the **renormalised coupling** $u_R(\mu)$ at the scale μ.

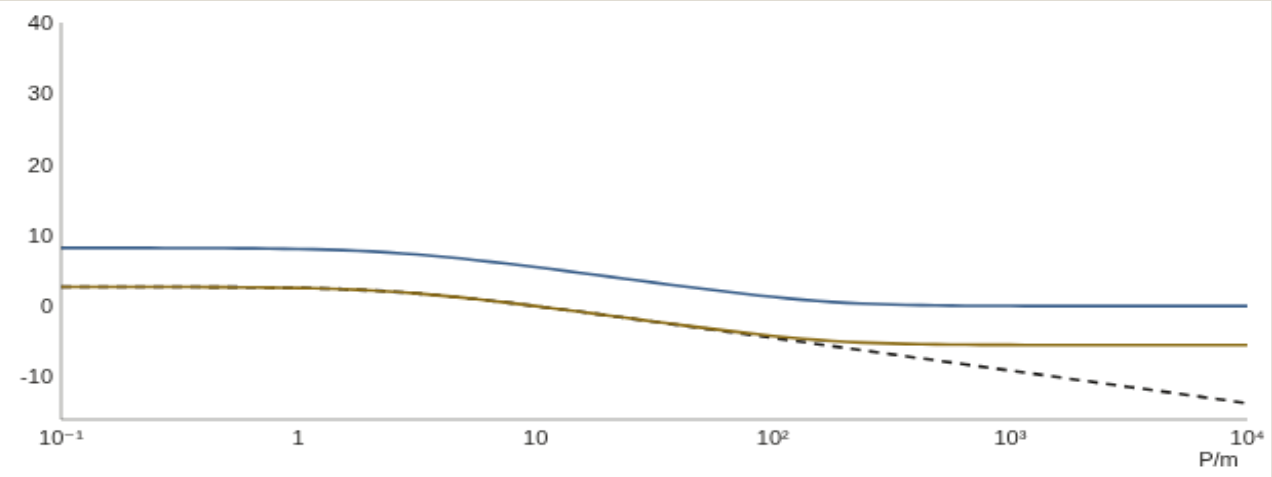
$$u_R(\mu) = u_0 - \tfrac{3}{2} u_0^2 B(\mu;\Lambda) + O(u_0^3), \qquad u_0 = u_R + \underbrace{\tfrac{3}{2} u_R^2 B(\mu;\Lambda)}_{\text{counterterm } \delta u} + O(u_R^3) \tag{6}$$

$\delta u \approx \frac{3u_R^2}{32\pi^2} \ln \frac{\Lambda^2}{\mu^2}$ diverges. Putting this back into (4), the divergence is shut up in the form of a difference.

$$\Gamma^4 = u_R - \frac{u_R^2}{2} \sum_j \big[B(P_j;\Lambda) - B(\mu;\Lambda) \big] + O(u_R^3) \xrightarrow{\Lambda\to\infty} u_R - \frac{u_R^2}{32\pi^2} \sum_j \int_0^1 \ln \frac{m^2 + x}{m^2 + x}$$

No Λ remains on the right-hand side. The content of renormalisation is that **predictions written in terms of the measured coupling are finite**, and the infinity is pushed into the unmeasurable u_0. The divergence of the mass (T) is pushed into m_0, and from two loops on also into the field normalisation Z; for ϕ^4 in four dimensions these three suffice at all orders. This is the practical meaning of "renormalisable" Known .

Figure 1 Raising the cutoff Λ with the knob, the bare bubble $16\pi^2 B(P;\Lambda)$ (solid line, upper) keeps rising, while the difference with the value at $\mu = 10m$ subtracted (solid line, lower) settles to the limit shown dashed. The horizontal axis is P (log). "Play" moves Λ back and forth



Λ = $10^{2.0}$ m: at P = 100m, bare 1.289, difference −4.208 (limit −4.496)

Computed The numbers in this figure were computed on this machine ((5)(7) integrated numerically in python3; the figure integrates the same formulas on the spot by Simpson's rule, agreeing to 4 digits). Static values ($16\pi^2$ times, $\mu = 10m$): the difference at $P = 1000m$ is −5.477 at $\Lambda = 10^2 m$, −9.096 at $10^4 m$, −9.099 at $10^8 m$ (limit −9.099). At the same Λ the bare $B(100m)$ keeps growing: 1.289, 10.208, 28.629.

05

The running coupling and the renormalisation group equation

The reference scale μ can be chosen freely. How u_R changes when u_0, Λ is held fixed and μ is moved is the **β function**.

$$\beta(u_R) = \mu \frac{du_R}{d\mu} = -\frac{3}{2} u_R^2 \, \mu \frac{\partial B(\mu;\Lambda)}{\partial \mu} + O(u_R^3) = \frac{3u_R^2}{16\pi^2} + O(u_R^3) \qquad (\mu \gg m) \quad (8)$$

Λ vanishes under differentiation. Differentiating (5) numerically, the coefficient approaches $3/16\pi^2 = 0.018998$: 0.01728 at $\mu = 10m$, 0.018997 at $10^3 m$ Computed . Integrating at one loop,

$$\frac{1}{u_R(\mu)} = \frac{1}{u_R(\mu_0)} - \frac{3}{16\pi^2} \ln \frac{\mu}{\mu_0} \quad (9)$$

and the coupling becomes stronger at shorter distances. For $u_R(\mu_0) = 1$, it is 1.78 at $\mu/\mu_0 = 10^{10}$, and the denominator is 0 at $10^{22.86}$ (the Landau pole; 7-08) Computed Physics . Yang–Mills has the opposite sign (4-06).

Example. At one loop in quantum electrodynamics, each Dirac particle of charge 1 (charged lepton) gives $d(1/\alpha)/d\ln\mu = -2/3\pi$. Extending from $1/\alpha = 137.036$ to $M_Z = 91.188$ GeV with the electron, μ and τ gives 132.20 (only the leading logarithm at one loop) Computed . The PDG 2024 value is $\hat{\alpha}^{(5)}(M_Z)^{-1} = 127.930$. MS is the scheme that, instead of a cutoff, subtracts only the poles of dimensional regularisation (4-05). The superscript (5) indicates the value with the five kinds of quark other than top included. The difference from 132.20 is due to the quarks and the constants of the scheme (3-07).

It can also be said for whole functions. The n-point function written in bare quantities knows nothing of μ, and the renormalised $\Gamma_R^{(n)}$ differs by the field normalisation Z, so

$$\left[\mu \frac{\partial}{\partial \mu} + \beta(u) \frac{\partial}{\partial u} - n\gamma(u)\right] \Gamma_R^{(n)}(P; u, \mu) = 0, \qquad \gamma = \tfrac{1}{2} \mu \, \partial_\mu \ln Z \quad (10)$$

(γ is the **anomalous dimension**; for ϕ^4 from two loops on). Where masses do not matter, multiplying momenta by s while moving the coupling along with them brings the left-hand side of (10) into a single ordinary derivative in s (the method of characteristics). Ignoring γ, $\Gamma^4(sP; u, \mu) = \Gamma^4(P; \bar{u}(s), \mu)$ ($s\,d\bar{u}/ds = \beta(\bar{u})$, $\bar{u}(1) = u$). Looking at momenta multiplied by s is replacing the coupling by its value at the scale $s\mu$ Known . Gell-Mann–Low 1954 defined the effective charge of the electron $e(\mu)$ as a function of scale (lecture notes: D. Tong, *Statistical Field Theory* § 3.5.3). Stueckelberg–Petermann 1953 are said to have noticed that the transformations of renormalisation form a group. The form of the equation that moves masses is due to Callan 1970 and Symanzik 1970 (the Callan–Symanzik equation), and in textbooks the

form (10) is also called by this name. All are statements at each order of perturbation theory.

06

Wilson's view — integrating out short scales

Regarding Λ as the genuine smallest scale, integrating only the fluctuations in $\Lambda/b < |k| < \Lambda$ and remeasuring lengths b times, all the coefficients of the action move a little. Repeating this is Wilson's 1971 renormalisation group, which implements Kadanoff's idea of blocking (2-08) as integration shell by shell in momentum (a rigorous version on the lattice is 7-05). As a metaphor, it is the operation of lowering the resolution of a photograph one step at a time. Precisely, it is a dynamical system in which the sequence of coefficients of the effective action is carried by a map at each step.

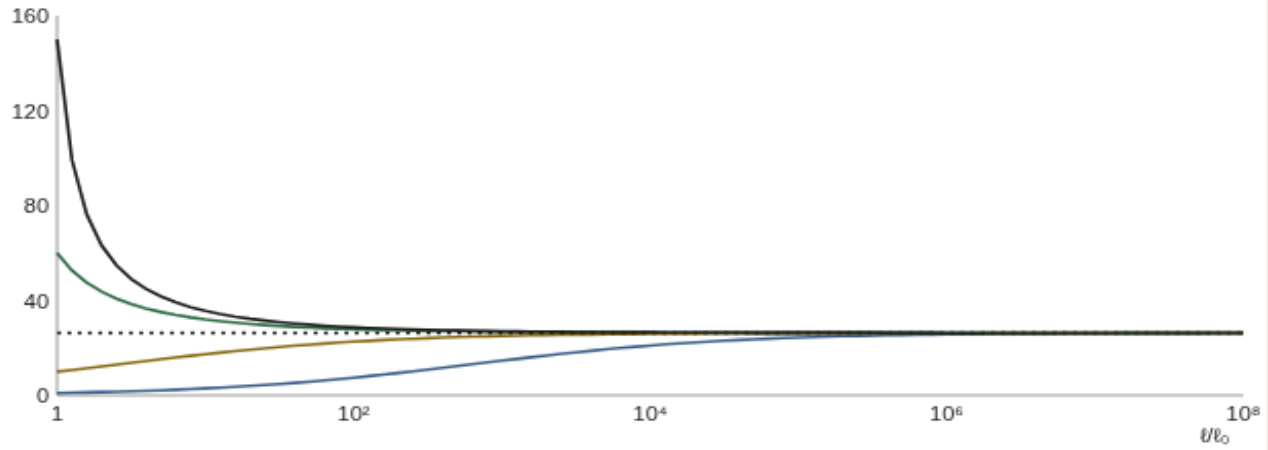
The coefficient of a local term of dimension $\Delta_{\mathcal{O}}$ has dimension $d - \Delta_{\mathcal{O}}$, and the dimensionless coefficient is multiplied by $b^{\,d-\Delta_{\mathcal{O}}}$ at each step towards long distances. The lecture notes (§ 3.2.2) divide them as 'Relevant if Δ_O < d', 'Irrelevant if Δ_O > d', 'Marginal if Δ_O = d' Known . Irrelevant terms are diluted to about $(E/\Lambda)^{\Delta_{\mathcal{O}}-d}$ at energy E, and at long distances only relevant and marginal terms remain. In four dimensions $\mathrm{tr}F^2$ ($\Delta = 4$) is marginal and $\mathrm{tr}F^4$ ($\Delta = 8$) is irrelevant (the question of 3-01). The expectation that the same continuum limit is reached regardless of the details of the lattice action (universality; 18-04) also comes from here Physics .

The direction of a marginal coupling is decided by quantum corrections. The lecture notes write 'most couplings which naively appear marginal will end up becoming either marginally relevant or marginally irrelevant due to such corrections', and give 'the Kondo effect and non-Abelian gauge theories' as examples that turn relevant (§ 3.4). ϕ^4_4 is on the side that weakens at long distances, Yang–Mills on the side that strengthens. With $t = \ln(\ell/\ell_0)$ towards long distances and the dimensionless $\hat{u} = u\,\ell^{\,4-d}$, extending d to real values,

$$\frac{d\hat{u}}{dt} = (4-d)\,\hat{u} - \frac{3}{16\pi^2}\,\hat{u}^2 \qquad (11)$$

(the formula of the lecture notes § 3.5 converted to the normalisation of $u/4!$; opposite in direction and in sign to (8)). For $d < 4$ there is a fixed point at $\hat{u}_* = 16\pi^2(4-d)/3$, and every positive initial value flows into it. This is the Wilson–Fisher 1972 fixed point, the smallest model of universality. For $d = 4$ it goes to 0 slowly, for $d > 4$ fast (7-08) Physics . A figure placing the $[g^2]$ of (2) and the fixed point of (11) next to other quantities that switch at four is the dimension knob of 1-12 § 10.

Figure 2 Moving the dimension d with the knob changes the flow of (11). The lines are $\hat{u}_0 = 1, 10, 60, 150$, the dotted line is the fixed point. For d<4 they gather to one point; for d≥4 they fall to 0. "Play" moves d back and forth between 3 and 5



`d = 3.50: û* = 26.32, at length ×10⁸: û = 26.25, 26.31, 26.32, 26.32`

Computed Physics The numbers in this figure were computed on this machine (closed-form solution; agreeing with RK4). Static values (length $\times 10^8$, $\hat{u}_0 = 1$ / 100): at $d = 3$, 52.6 / 52.6 ($\hat{u}_* = 52.64$); at $d = 4$, 0.741 / 2.78; at $d = 5$, 9.8×10^{-9} / 3.5×10^{-7}. The formula is a one-loop approximation, reliable when $4 - d$ is small.

Detour: non-renormalisable theories are useful too — the Fermi interaction

A subsection you may skip.

In the low-energy description of the weak interaction, the term in which four fermions meet at one point ($\Delta = 6$) carries $G_F = 1.1663788(6) \times 10^{-5}\ \mathrm{GeV}^{-2}$ (PDG 2024), and it is non-renormalisable. Even so, $G_F E^2$ is 1.2×10^{-5} at 1 GeV and reaches 1 only at $E = 292.8$ GeV Computed . Before that, the gauge theory of the W and Z (3-07) takes over. Such theories are useful as effective theories below some scale Known .

07

In gauge theories — subtracting without breaking the symmetry

One would like the counterterms of Yang–Mills, once added, to stay within the Yang–Mills form. The gauge-invariant local terms of dimension at most 4 are only $\mathrm{tr}F^2$ and $\mathrm{tr}F\tilde{F}$ (the θ term; a total derivative that does not enter perturbation theory; 3-05), and the mass term $\mathrm{tr}A_\mu A_\mu$ cannot be written (3-02). However, perturbation theory is done with the gauge-fixed action. There, instead of gauge invariance, BRST symmetry (the symmetry that remains after gauge fixing) constrains the form of the counterterms (4-08). As a result, the gauge field part is kept in the form $\mathrm{tr}F^2$. The divergences are absorbed into g and the field normalisations (including those of the ghost and gauge-fixing terms), and since there is one parameter, the integration constant of the running becomes the only scale Λ_{YM} (a different quantity from the cutoff Λ; 4-07). Gauge invariance also reduces divergences. The two-point correction of the gluon is

restricted to transverse form by the Ward (in the non-abelian case, Slavnov–Taylor) identities Known .

$$\Pi_{\mu\nu}(P) = \left(P^2\delta_{\mu\nu} - P_\mu P_\nu\right)\Pi(P^2) \quad (12)$$

Transverse means $P_\mu\Pi_{\mu\nu} = 0$, that is, there is no component along the momentum. Since the second power of P factors out, of the $\omega = 2$ in Table 2 only the logarithmically divergent Π remains. $\Lambda^2\delta_{\mu\nu}$, corresponding to a mass term, does not appear in gauge-invariant regularisations. A momentum cutoff breaks this, and the construction of Magnen–Rivasseau–Sénéor 1993 also says it 'breaks gauge invariance and requires gauge restoring counterterms'. The abstract of 't Hooft 1971 says that regulator fields that are not gauge invariant can also be used if gauge invariance can be restored by finitely many counterterms in the limit of large regulator mass, and that massless Yang–Mills fields 'appear to be renormalizable' in the usual sense. Gauge fixing and ghosts are in 4-08. The lattice keeps gauge invariance exactly at the stage of regularisation (5-01).

08

Renormalisation seen from mathematics — where the theorems are, and where they are not yet

The key is not to mix up three levels.

Table 3 Three levels of renormalisation

Level	Status	Chapter
That each order of perturbation theory is finite	Known Theorem. BPHZ (Bogoliubov–Parasiuk 1957, Hepp 1966, Zimmermann 1969), Epstein–Glaser 1973. For gauge theories, 't Hooft 1971, 't Hooft–Veltman 1972, Lee–Zinn-Justin 1972, Becchi–Rouet–Stora 1976. Another proof by the BV formalism, Costello 2011	4-05
Summing the series (four dimensions)	Physics For both ϕ^4 and Yang–Mills the series is thought to diverge, and no meaning is given to its sum	4-12
Summing the series (two and three dimensions)	Known For ϕ^4 in two dimensions, divergence of the series (Jaffe 1965) and Borel summability (Eckmann–Magnen–Sénéor 1975); for ϕ^4 in three dimensions, Borel summability (Magnen–Sénéor 1977); for the massive two-dimensional Gross–Neveu model, the Borel sum (Feldman–Magnen–Rivasseau–Sénéor 1986)	4-12

Level	Status	Chapter
The theory itself with the cutoff removed	Known ϕ^4 in two and three dimensions has been constructed. Four-dimensional Yang–Mills: uniform bounds on the effective action in finite volume, on the lattice, at small coupling (Bałaban's series of papers in the 1980s; the 1989 theorem carries an assumption on the effective coupling; what can be taken from uniform bounds is a convergent subsequence by compactness, and the limit is not necessarily unique). Foundations of a continuum construction with an infrared cutoff (Magnen–Rivasseau–Sénéor 1993). The continuum theory as the scaling limit of Wilson's lattice was built by the construction published in 2026-11 (at the level of the committee's assessment; 6-00). The uniqueness of the infinite-volume limit has not been shown (6-04)	7-01, 7-07

An early example of the first way of thinking is Dyson 1949 for quantum electrodynamics (an argument removing the divergences at each order by renormalising mass and charge; the complete proof came later with Weinberg 1960 and BPHZ; convergence of the series was not shown). On the third, the problem statement writes that for ϕ^4 in two and three dimensions the coefficients are chosen 'to depend on κ in the way that perturbation theory suggests' on the cutoff κ, and 'One then proves that the expectations converge as $\kappa \to \infty$, even though the specified constants α, . . . diverge.' (§ 6.2). This is the constructive version of (7) being finite even though the δu of (6) diverges.

For four-dimensional Yang–Mills, it is required that the correlation functions agree at short distances with the textbook predictions of asymptotic freedom and perturbative renormalisation theory (§ 4; quoted in 0-03 § 07). § 2 writes 'Asymptotic freedom ensures a decisive regularity in the case when classical Sobolev inequalities are borderline'. In this article's reading, "borderline" corresponds to "the coupling is dimensionless in four dimensions" of § 03 (18-04), and the sentence is an outlook, not a theorem. That handling the infinities is not itself the centre of the difficulty is in 16-08.

What this chapter can say

	Content
What is said	The calculations (1)–(11), the properties of the regularisations in Table 1, the threefold classification of renormalisability — as calculations and theorems of perturbation theory Known . Table 2, Figures 1 and 2, the numerical differentiation of the β coefficient, the running of QED and the numbers for G_F Computed
Conditional	(11) and Figure 2 are an approximation extending the one-loop formula to real dimensions. Universality remains a premise for four-dimensional Yang–Mills Physics . (12) is quoted at textbook level, and the source is bibliographic data only

	Content
What is not said	The sum of the perturbation series. The existence of the four-dimensional theory with the cutoff removed, the continuum limit of Yang–Mills, the mass gap (this site shows none of them)

Sources and reproduction

Item	Kind	Source / tools
Read in full text	Known	A. Jaffe, E. Witten, *Quantum Yang–Mills Theory* (the Clay problem statement) § 2, § 4, § 6.2, § 6.5. D. Tong, *Statistical Field Theory* (lecture notes) Chapter 3 § 3.2.2, § 3.4, § 3.5, § 3.5.3. J. Erler, A. Freitas, Electroweak Model…, in S. Navas et al. (PDG), Phys. Rev. D 110 (2024) 030001 (eqs. (10.8)(10.11), $\hat{\alpha}(M_Z)$) and Summary Tables
Read in abstract only	Known	Gell-Mann–Low, Phys. Rev. 95 (1954) 1300; Weinberg, Phys. Rev. 118 (1960) 838; Callan, Phys. Rev. D 2 (1970) 1541; Symanzik, CMP 18 (1970) 227; 't Hooft, Nucl. Phys. B 33 (1971) 173; 't Hooft–Veltman, Nucl. Phys. B 44 (1972) 189; 't Hooft–Veltman, Nucl. Phys. B 50 (1972) 318; Bollini–Giambiagi, Nuovo Cim. B 12 (1972) 20; Lee–Zinn-Justin, Phys. Rev. D 5 (1972) 3121; Jaffe, CMP 1 (1965) 127; Eckmann–Magnen–Sénéor, CMP 39 (1975) 251; Magnen–Sénéor, CMP 56 (1977) 237; Wilson, Phys. Rev. B 4 (1971) 3174; Wilson–Fisher, Phys. Rev. Lett. 28 (1972) 240 (INSPIRE)
Checked against the survey material	Known	Dyson, Phys. Rev. 75 (1949) 486, 1736 (abstracts); Magnen–Rivasseau–Sénéor, CMP 155 (1993) 325; Bogoliubov–Parasiuk, Acta Math. 97 (1957) 227; Hepp, CMP 2 (1966) 301; Zimmermann, CMP 15 (1969) 208; Epstein–Glaser, Ann. IHP A 19 (1973) 211; Costello, AMS (2011); Bałaban, CMP 122 (1989) 175, 355 (CMP 109 (1987) 249 bibliographic data only); Feldman–Magnen–Rivasseau–Sénéor, CMP 103 (1986) 67 (abstract)
Bibliographic data only	Known	Becchi–Rouet–Stora, Ann. Phys. 98 (1976) 287 (from the bibliography of the problem statement); Stueckelberg–Petermann, Helv. Phys. Acta 26 (1953) 499; Peskin–Schroeder, *An Introduction to Quantum Field Theory* (1995; the form of (12))
Table 2, Figures 1 and 2, numerical examples	Computed	python3 (scipy, enumeration, RK4; each under a minute). The figures' JS checked with node

Next chapter to read: 4-05 The algebra of divergences (BPHZ, dimensional regularisation, Connes–Kreimer). The sign of the β function is in 4-06.

The algebra of divergences — nested subtraction becomes a decomposition in a group, yet still does not leave formal power series

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — Three prescriptions for subtracting nested divergences (BPHZ, Epstein–Glaser, dimensional regularisation). That subtraction becomes a Birkhoff decomposition of a group in the Hopf algebra of Connes–Kreimer (computed in a small model). Costello's families of effective actions, and where all this falls short of the mass gap.

Prerequisite chapters — 4-04 Divergences and renormalisation. Helpful: 1-11 (distributions), 4-03 (Feynman diagrams).

Said first — This is a guide to known results at the level of formal power series, and shows nothing about the existence of Yang–Mills theory or the mass gap. The small model is for seeing the algebraic structure, not a four-dimensional calculation.

Order of this chapter

1. Nested divergences — seen in a small model: subtracting only the whole leaves the outer scale
2. The forest formula — BPHZ: subtract from the inside, and count the sets to subtract by forests
3. Extension in position space — Epstein–Glaser: the freedom is δ and its derivatives
4. Moving the dimension continuously — dimensional regularisation: divergences become poles 1/ε. The MS scheme
5. The algebra of diagrams — the Hopf algebra of Connes–Kreimer: the coproduct writes the extraction
6. Birkhoff decomposition — subtraction becomes a decomposition in a group: counterterms are one factor. Detour: rooted trees and Runge–Kutta methods

7. Families of effective actions — Costello: a family of actions, one for each scale
8. Where the algebra reaches and where it does not: the measure and $\exp(-1/g^2)$ lie outside

---

01

Nested divergences — seen in a small model

In 4-04 we removed one-loop divergences by subtracting the value at a reference scale μ. What was subtracted was a number independent of the external momenta (a counterterm). From two loops on, a part of a diagram (a **subdiagram**) diverges on its own (a **subdivergence**). Two subdivergences are either **nested**, one containing the other, **overlapping**, sharing a part without either containing the other, or disjoint.

We use a small model with nesting only (Kreimer's toy model; the definition in § 3.2 of Panzer 2012). "One loop" is $\int_0^\infty dy/(y+s)$, where $s>0$ is the outer scale and y the loop variable. Like the four-dimensional bubble it diverges logarithmically, so we **regularise** by multiplying by y^{-z} (z is a small positive number, corresponding to the ε of § 04), and build nesting by putting the inner value into the outer integrand.

$$\varphi(\bullet)(s)=\int_0^\infty \frac{y^{-z}\,dy}{y+s}=s^{-z}F(z),\qquad F(z)=\frac{\pi}{\sin\pi z}=\frac{1}{z}+\frac{\pi^2}{6}z+\cdots \tag{1}$$

$$\varphi(\ell_2)(s)=\int_0^\infty \frac{y^{-z}}{y+s}\,\varphi(\bullet)(y)\,dy=s^{-2z}F(z)\,F(2z) \tag{2}$$

Writing the part with negative powers of z (the **pole part**) as pole, with $L=\ln(s/\mu)$ and $\mu=1$,

$$\operatorname{pole}\varphi(\bullet)=\frac{1}{z},\qquad \operatorname{pole}\varphi(\ell_2)=\frac{1}{2z^2}-\frac{L}{z} \tag{3}$$

Computed The pole of the double loop contains L. Subtracting only the pole of the whole, the subtracted quantity contains $\ln s$ and is not a local term that can be added to the action. **Example:** at $z=0.1$, $L=1$, $\varphi(\ell_2)=44.49$, and its pole part is $50-10=40$.

02

The forest formula — BPHZ

The cure is to subtract from the inside. In Bogoliubov's prescription, one first adds the term in which the subdivergence is replaced by its counterterm $S(\bullet) = -1/z$, and then subtracts the whole.

$$\bar{R}(\ell_2) = \varphi(\ell_2) + S(\bullet)\,\varphi(\bullet), \qquad \operatorname{pole}\bar{R}(\ell_2) = -\frac{1}{2z^2} \quad (4)$$

Computed L/z disappears, and the overall counterterm $S(\ell_2) = 1/(2z^2)$ does not depend on the outer scale ($\bar{R}$: Bogoliubov's map). For general diagrams the sum is over **forests** — sets of divergent subdiagrams any two of which are nested or disjoint (including the empty set and the whole diagram). Zimmermann's forest formula is

$$R\,\Gamma = \sum_{\mathcal{F}:\ \text{forests}} \prod_{\gamma\in\mathcal{F}} (-T_\gamma)\,\Gamma \quad (5)$$

where T_γ is the operation extracting the divergent part of γ (in momentum space, the low orders of the Taylor expansion in the external momenta), and the product acts from the inside (a form with details omitted). **Example:** the forests of ℓ_2 are $\varnothing, \{\bullet\}, \{\ell_2\}, \{\bullet, \ell_2\}$, and (5) is the expansion $(1 - T_{\ell_2})(1 - T_\bullet)\,\ell_2$. The $(1 - T_\bullet)$ part corresponds to the $\bar{R}$ of (4), and the remaining $-T_{\ell_2}$ to adding the overall counterterm $S(\ell_2) = 1/(2z^2)$, leaving a value without poles (the φ_+ of § 05). It is called BPHZ after the initials of Bogoliubov–Parasiuk (1957), Hepp (1966) and Zimmermann (1969), and the gaps in the first proof are said to have been filled by the work of the latter two (§ 1 of Ebrahimi-Fard–Kreimer 2005). The forest formula with overlaps also follows from the Hopf algebra of rooted trees (§ 05) (abstract of Kreimer 1999). Known

03

Extension in position space — Epstein–Glaser

This section uses the distributions of 1-11 § 04 and § 08. We give only the essentials. A **test function** g is a function that can be differentiated any number of times and goes to 0 quickly far away, and a **distribution** t is a rule assigning linearly a number $\langle t, g\rangle$ to each test function. $\langle \delta, g\rangle = g(0)$, and its derivative is defined by $\langle \partial^\alpha \delta, g\rangle = (-1)^{|\alpha|}\partial^\alpha g(0)$. A sequence of distributions "converges as distributions" if $\langle t_n, g\rangle$ converges for every test function g.

In position space, a divergence is the impossibility of extending a distribution to a singular point. We take examples in Euclidean $\mathbb{R}^4$. The square of the four-dimensional massless propagator $G = 1/(4\pi^2|x|^2)$ (the bubble diagram) behaves as $|x|^{-4}$ near the

origin, and its integral diverges logarithmically. The method of Epstein–Glaser (1973) was set up in Minkowski space (spacetime with time); it builds time-ordered products (whose expectations are the correlation functions) inductively in the number of vertices (§ 4 of Brunetti–Fredenhagen 2000), and **extends** values fixed off the **diagonal**, where points coincide, onto the diagonal. Renormalisation is exactly the extension of expectations to the diagonal (abstract of the same paper). Since it is formulated as extension, there is no need to write divergent integrals along the way. The extension problem itself has the same form in Euclidean $\mathbb{R}^d$, as below. The freedom is measured by the **scaling degree** ((33) of the same paper; numbering of the arXiv version; convergence as distributions).

$$\mathrm{sd}(t) = \inf\{\omega : \lambda^\omega\, t(\lambda x) \to 0 \ (\lambda \downarrow 0)\}, \qquad \mathrm{sd}(|x|^{-a}) = a, \quad \mathrm{sd}(\delta) = d \tag{6}$$

For t_0 on $\mathbb{R}^d \setminus \{0\}$ (Theorems 5.2 and 5.3 of the same paper) Known : if $\mathrm{sd}(t_0) < d$, the extension preserving the scaling degree is unique. If $d \le \mathrm{sd}(t_0) < \infty$, extensions exist, and the difference of two extensions is supported at the origin, of the form

$$t - t' = \sum_{|\alpha| \le \rho} c_\alpha\, \partial^\alpha \delta, \qquad \rho = \mathrm{sd}(t_0) - d \tag{7}$$

ρ corresponds to the superficial degree of divergence of 4-04, and the constants c_α are the freedom of local counterterms. **Example:** for $t_0 = |x|^{-4}$ ($d = 4$, $\rho = 0$), the extension that subtracts $g(0)$ only inside the ball of radius $1/\mu$,

$$\langle u_\mu, g\rangle = \int_{\mathbb{R}^4} \frac{g(x) - g(0)\,\mathbf{1}\{|x| < 1/\mu\}}{|x|^4}\, d^4x \tag{8}$$

agrees with t_0 for $x \neq 0$, and converges because the numerator is $O(|x|)$. The difference on changing μ is $2\pi^2 \ln(\mu'/\mu)\, g(0)$ ($2\pi^2$ is the area of S^3), a constant multiple of δ.

Example: for $g_s = e^{-|x|^2/s^2}$, $\langle u_\mu, g_s\rangle = 2\pi^2(\ln(\mu s) - \gamma/2)$ (γ is Euler's constant). At $\mu = 1$, for $s = 0.1, 1, 10$ it is −51.148, −5.697, 39.754, and the difference from $\mu = 2$ is $2\pi^2 \ln 2 = 13.682$ in each case. Removing $|x| < \eta$ without extending gives 85.21, 357.91 at $\eta = 10^{-2}, 10^{-8}$ ($s = 1$), growing. Computed

04

Moving the dimension continuously — dimensional regularisation

We regard momentum integrals as functions of the dimension d. The bubble of 4-04 (mass m) converges for $d < 4$, giving

$$I(d)=\int \frac{d^dk}{(2\pi)^d}\,\frac{1}{(k^2+m^2)^2}=\frac{\Gamma(2-d/2)}{(4\pi)^{d/2}}\,(m^2)^{d/2-2} \quad (9)$$

($1/(8\pi m)$ at $d=3$). The right-hand side extends **meromorphically** to complex d (as a function complex-differentiable except at isolated points called poles), and has a pole at $d=4,6,\dots$. Using this is **dimensional regularisation**. Here we write $d=4-2\varepsilon$ (half the ϵ of the $d=4-\epsilon$ of 4-04). Then

$$16\pi^2\,I=\frac{1}{\varepsilon}-\gamma+\ln 4\pi-\ln m^2+O(\varepsilon) \quad (10)$$

and the divergence becomes the pole $1/\varepsilon$. At $\varepsilon=0.1, 0.01, 0.001$, $16\pi^2 I$ ($m=1$) is 12.2535, 101.9814, 1001.9565, and the differences from the approximation (10) are 0.2997, 0.0276, 0.0027. The role of $\ln k_{\max}^2$ in the $\ln(k_{\max}^2/m^2)-1$ cut at $|k|<k_{\max}$ in momentum is played by $1/\varepsilon$. Computed The **MS scheme** subtracts only $1/\varepsilon$, and the $\overline{\text{MS}}$ **scheme** subtracts $1/\varepsilon-\gamma+\ln 4\pi$. Since a scale μ is introduced (the coupling is multiplied by $\mu^{2\varepsilon}$), $-\ln(m^2/\mu^2)$ remains at $\overline{\text{MS}}$. The $\Lambda_{\overline{\text{MS}}}$ of 4-06 is the Λ of this scheme.

In 1972 there were papers by Bollini–Giambiagi and by 't Hooft–Veltman (titled "Regularization and renormalization of gauge fields"), and it is said to be used as standard in Yang–Mills perturbation theory as a regularisation that does not break gauge symmetry. The price is that (9) is an analytic continuation diagram by diagram, and gives no measure for non-integer d. The problem statement too writes that there are few ways to recover the reflection positivity lost by dimensional regularisation and the like (p.11; quoted in 0-03 § 03; 4-10).

05

The algebra of diagrams — the Hopf algebra of Connes–Kreimer

Subtraction is the repetition of "extracting a subdiagram γ" and "the remainder Γ/γ with γ collapsed to a point". Before making this an operation, we prepare three words.

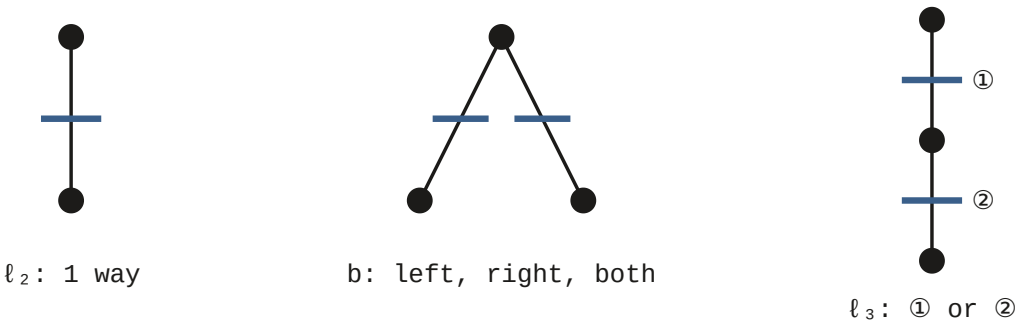
- **One-particle irreducible**: a diagram that does not split in two when any one internal line is cut.
- $\otimes$ (tensor product): $\gamma\otimes\Gamma/\gamma$ is a symbol that formally sets a pair of diagrams side by side. It is linear in each of left and right (bilinear), and does not multiply left and right. When values are put in later, it is multiplied as $\varphi(\gamma)\,\varphi(\Gamma/\gamma)$.
- **Hopf algebra**: an algebra with a product, a **coproduct** (a map splitting one element into a sum of pairs $\otimes$), and auxiliary maps called the counit and the antipode. This chapter uses only the coproduct.

Making subtraction an operation is the Hopf algebra of Kreimer (1998) and Connes–Kreimer (1998, 2000): the generators are divergent one-particle irreducible diagrams, the product is setting diagrams side by side, and extraction is the coproduct ((15) of Ebrahimi-Fard–Kreimer 2005). The sum in (11) is over divergent one-particle irreducible proper subdiagrams and their disjoint sets (set side by side as a product). For a set, Γ/γ is the remainder with all of them collapsed to points. Writing the coproduct as Δ is a convention of the literature, unrelated to the Δ of the mass gap.

$$\Delta(\Gamma) = \Gamma \otimes 1 + 1 \otimes \Gamma + \sum_{\gamma \subsetneq \Gamma} \gamma \otimes \Gamma/\gamma \quad (11)$$

With nesting only, **rooted trees** suffice in place of diagrams. Vertices are divergent subdiagrams, and children are inner subdivergences. The sum in (11) becomes a sum over **admissible cuts** (choices of edges that cut every path from the root to a leaf at most once), with the cut-off side going on the left and the root side on the right (Figure 1).

Figure 1 Rooted trees and admissible cuts. The top is the root (the whole diagram), lower is further inside. Coloured lines are cuts; the lower side cut off is the left factor of (12), the root side the right factor. In ℓ_3, ① and ② are on the same path, so they are not cut together



$$\Delta(\ell_2) = \ell_2 \otimes 1 + 1 \otimes \ell_2 + \bullet \otimes \bullet$$
$$\Delta(b) = b \otimes 1 + 1 \otimes b + 2\ \bullet \otimes \ell 2 + \bullet\bullet \otimes \bullet \quad (12)$$

b is the tree with two children at the root. Regularised values φ (maps taking values in Laurent series in z and carrying products to products; **characters**) form a group under the convolution $\varphi_1 \star \varphi_2 = m \circ (\varphi_1 \otimes \varphi_2) \circ \Delta$ (m is the product of values). The unit is the character that gives 1 to the empty diagram and 0 to other diagrams. Since there are inverses, the problem can be posed as multiplying φ by a factor S_R that cancels its divergences to make $S_R \star \varphi = \varphi_+$ finite. The counterterm S_R plays the role of the inverse of φ (§ 06). Fixing a scheme R (for MS the pole part, for MOM the value at $s = \mu$), the counterterm S_R and the renormalised value φ_+ are fixed in order of tree size ((21)(25)(26) of Ebrahimi-Fard–Kreimer 2005).

$$S_R(\Gamma) = -R\big[\bar{R}(\Gamma)\big], \qquad \varphi_+(\Gamma) = (1-R)\big[\bar{R}(\Gamma)\big]$$
$$\bar{R}(\Gamma) = \varphi(\Gamma) + \sum_{\gamma} S_R(\gamma)\,\varphi(\Gamma/\gamma) \tag{13}$$

In the toy model, writing the subtree rooted at vertex v as t_v and the number of vertices as $|t|$, the values take closed form (Proposition 3.2.2 of Panzer 2012).

$$\varphi(t)(s) = s^{-z|t|} \prod_{v} F\big(z\,|t_v|\big) \tag{14}$$

Computing (13) for the 8 trees with up to 4 vertices, (i) none of the counterterms contains L (locality), (ii) the renormalised values are polynomials of degree $|t|$ in L with top-degree term $(-L)^{|t|}/t!$ ($t! = \prod_v |t_v|$ is the **tree factorial**), and (iii) the MOM values agreed with Panzer's (3.4.2)–(3.4.5). The difference between MS and MOM is finite. Computed

Computed Table 1 Renormalisation in the toy model (computed exactly in rationals). The right-hand column is the value at $z \to 0$. Subtracting by MS changes the constants (ℓ_2 becomes $\frac{L^2}{2} + \frac{\pi^2}{4}$, ℓ_3 becomes $-\frac{L^3}{6} - \frac{5\pi^2 L}{12}$).

Tree	$t!$	Pole part of the bare value	Counterterm (MS)	φ_+ (MOM)
$\bullet$	1	$\frac{1}{z}$	$-\frac{1}{z}$	$-L$
ℓ_2	2	$\cancel{\frac{1_2}{2z^2}} - \frac{L}{z}$	$\cancel{\frac{1_2}{2z^2}}$	$\frac{L^2}{2}$
b	3	$\cancel{\frac{1_3}{3z^3}} - \cancel{\frac{L_2}{z^2}} + \frac{27L^2+11\pi^2}{18z}$	$-\cancel{\frac{1_3}{3z^3}} + \frac{\pi^2}{18z}$	$-\frac{L^3}{3} - \frac{\pi^2 L}{3}$
ℓ_3	6	$\cancel{\frac{1_3}{6z^3}} - \cancel{\frac{L_2}{2z^2}} + \frac{27L^2+14\pi^2}{36z}$	$-\cancel{\frac{1_3}{6z^3}} - \frac{\pi^2}{18z}$	$-\frac{L^3}{6} - \frac{\pi^2 L}{6}$

06

Birkhoff decomposition — subtraction becomes a decomposition in a group

The main theorem of Connes–Kreimer (2000) (abstract) is that the dimensionally regularised bare theory gives a loop $\gamma(z)$ in the group of characters on a small circle around the integer spacetime dimension D, and 'the renormalized theory is just the evaluation at z = D of the holomorphic part γ+ of the Birkhoff decomposition of γ'. $z = D$ is the integer dimension of spacetime, γ is the loop given by the regularised theory (the family of characters as the dimension z moves on the circle), and γ_+ is the holomorphic factor of its decomposition. Known Below we write z as the deviation from D (the same role as the z of the toy model, with $z = 0$ the integer dimension).

$$\varphi(z) = \varphi_-(z)^{\star -1} \star \varphi_+(z) \tag{15}$$

φ_- is a polynomial in $1/z$ (constant term the unit) and is the MS counterterm; φ_+ is holomorphic at $z = 0$, and $\varphi_+(0)$ is the renormalised value. As an analogy, it is the

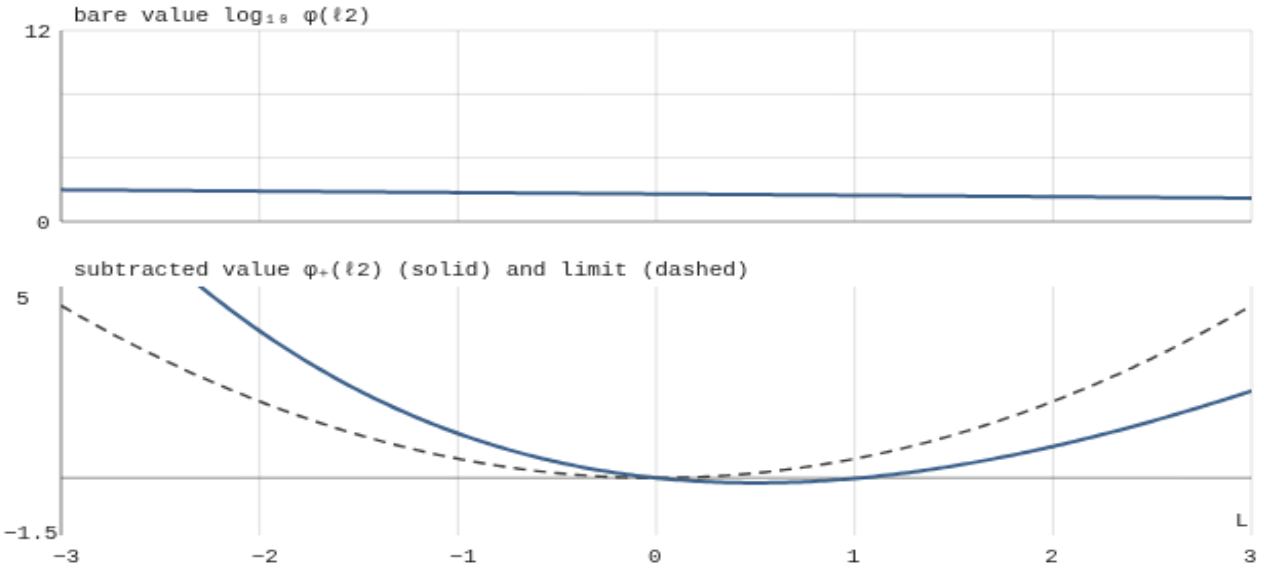
multiplicative version of splitting a Laurent series into its pole part and its holomorphic part. Precisely, since the group is not commutative, they are fixed in order of tree size by the induction of (13). That S_R and φ_+ are again characters is because R satisfies $R^2 = R$ and the **Rota–Baxter identity**

$$R(x)\,R(y) + R(xy) = R\big(R(x)\,y + x\,R(y)\big) \quad (16)$$

((20) of Ebrahimi-Fard–Kreimer 2005). Known **Example:** for the pole part R and $x = 1/z + a$, $y = 1/z + b$, both sides are $2/z^2 + (a+b)/z$.

Figure 2 is the MOM of the ladder ℓ_n (n vertices in a row). The sum in (13) becomes $\bar{R}(\ell_n) = \varphi(\ell_n) + \sum_{j<n} S_R(\ell_j)\,\varphi(\ell_{n-j})$; as $z \to 0$ the bare value runs away at the rate z^{-n}, while the subtracted value settles to a polynomial. Computed

Figure 2 Making the knob z smaller, the bare value $\log_{10}\varphi(\ell_n)$ in the upper panel runs away upwards, and the subtracted value $\varphi_+(\ell_n)$ in the lower panel (solid line) approaches a polynomial in $z \to 0$ (dashed line). The horizontal axis is L, and n is the number of rungs of the ladder. "Play" lowers z



n = 2, z = 0.100, at L = 1, bare value 44.49, subtracted value −0.0141, limit 0.5000

Computed The numbers in this figure were computed on this machine. The closed forms of (13)(14) are evaluated in the figure and checked against python3.

If JavaScript is disabled: at $L = 1$, the subtracted values at $z = 0.02,\ 0.005$ are, for $n = 1$: −0.991, −0.998 (limit −1); $n = 2$: 0.394, 0.473 (0.5); $n = 3$: −1.696, −1.776 (−1.812); $n = 4$: 0.655, 1.419 (1.687). The bare values at $z = 0.02$ are 49, 1205, 19802, 245195. The difference from the limit shrinks roughly in proportion to z. Computed

Detour: rooted trees and Runge–Kutta methods

A subsection you may skip.

Expanding the solution of $y' = f(y)$ in h, reading the derivative of order k as "a vertex with k children", $f,\ f'f,\ f''(f,f),\ f'f'f$ becomes $\bullet,\ \ell_2,\ b,\ \ell_3$.

$$y(h) = y + hf + \frac{h^2}{2} f'f + \frac{h^3}{6}\big(f''(f,f) + f'f'f\big) + O(h^4) \tag{17}$$

The coefficient of h^3 is $1/t!$ for ℓ_3, and for b the reciprocal of the product of 2, from exchanging left and right, and $t! = 3$; this is the same tree factorial as in Table 1. Runge–Kutta methods are built to reproduce these coefficients up to a given order, and the group handling the expansion is called Butcher's group (Butcher 1972). Brouder (2000), in his abstract, points out the connection between the algebra of rooted trees in renormalisation and Runge–Kutta methods. Known

07

Families of effective actions — Costello

Another route is to regard the family of actions, one for each scale, as the theory (Wilson's renormalisation group; 7-05). Costello (a 2007 paper, a 2011 book) restricts the t (the square of a length) of the propagator $1/k^2 = \int_0^\infty e^{-tk^2}\,dt$ from ε to T.

$$P(\varepsilon, T)(k) = \int_\varepsilon^T e^{-tk^2}\,dt = \frac{e^{-\varepsilon k^2} - e^{-Tk^2}}{k^2} \tag{18}$$

Precisely, the paper is written with the heat kernel on a compact manifold, which on flat space becomes e^{-tk^2}. **Example:** at $\varepsilon = 0.01$, $T = 100$, $k^2 P$ is 0.990, 0.368, 1.2×10^{-4} at $k = 1, 10, 30$. Computed Actions at different scales are linked by the renormalisation group equation ($\Gamma(P, S)$ is the sum of diagrams joining vertices of S by P).

$$S_{\mathrm{eff}}(T_2) = \Gamma\big(P(T_1, T_2),\, S_{\mathrm{eff}}(T_1)\big) \tag{19}$$

The Counterterm Uniqueness Theorem of the paper (Theorem B of Costello) is that, once a scheme is fixed (a choice of which functions of ε to call the singular part), there is a unique counterterm $S^{\mathrm{CT}}(\varepsilon)$ such that $\lim_{\varepsilon\to 0} \Gamma(P(\varepsilon, T), S - S^{\mathrm{CT}})$ exists, and the counterterms are local and independent of T. The counterterms are power series in $\hbar$ (the letter counting loops). The **effective action** at scale T is the limit as $\varepsilon \to 0$ after subtracting these counterterms,

$$\Gamma^R\big(P(0,T), S\big) = \lim_{\varepsilon\to 0} \Gamma\big(P(\varepsilon, T),\, S - S^{\mathrm{CT}}(\varepsilon)\big) \tag{20}$$

and building diagrams with $P(0,T)$ as it is diverges. The effective action too is a formal power series in $\hbar$ (§ 08). The Classification Theorem for Families of Effective Actions (Theorem C of Costello) is that, once a scheme is fixed, local actions correspond one-to-one to families of effective actions satisfying (19) and locality as $T \to 0$. Changing the scheme amounts to putting different coordinates on the same family. Known

Gauge symmetry is expressed by the quantum master equation of the Batalin–Vilkovisky (BV) formalism (here only the name; it gathers the condition of keeping gauge symmetry into one equation with auxiliary fields added). The 2007 paper distinguishes renormalisation (constructing counterterms) from quantisation (constructing solutions of the quantum master equation), and gives pure Yang–Mills on compact four-dimensional manifolds as an example of the former, and Chern–Simons theory as a non-trivial example of the latter. The 2011 book is described as containing a proof of the perturbative renormalisability of Yang–Mills by cohomology (an algebraic tool that looks at $\ker Q / \operatorname{im} Q$ for maps Q with $Q^2 = 0$; here only the name). Renormalisability itself is said to have been established by 't Hooft (1971), 't Hooft–Veltman (1972), Lee–Zinn-Justin (1972), Becchi–Rouet–Stora (1976) and others (7-12 § 06). Known

08

Where the algebra reaches and where it does not

All the tools so far work on **formal power series** in the coupling constant, make each order finite, and confine the freedom to finitely many local terms. Known Three things lie out of reach. **(i) The sum of the series:** the perturbation series of four-dimensional Yang–Mills is thought to diverge (4-12). Physics **(ii) Quantities that are 0 at every order:** the one-loop scale (4-07; for pure $SU(N)$, $\beta_0 = 11N/3$; conversion between four ways of writing the coefficient in Table 1 of 4-06)

$$\Lambda = \mu \, \exp\!\Big(-\frac{8\pi^2}{\beta_0 \, g^2}\Big) \quad (21)$$

(with $b_0 = \beta_0/(16\pi^2)$, the same formula as (3) of 4-07) has all its Taylor coefficients in g equal to 0. The mass gap is expected to be proportional to Λ Physics , and such quantities cannot be distinguished from 0 within formal power series (7-12 § 07). **(iii) The measure and positivity:** what the problem statement asks for is not a sequence of coefficients but a theory satisfying the axioms (0-03), whose correlation functions should agree at short distances with asymptotic freedom and the predictions of perturbative renormalisation theory (p.6; quoted in 0-03 § 07). The algebra of this chapter is on the side of the answers that a construction must reproduce.

An example of the same algebra being used inside a tool outside formal power series is the renormalisation of Hairer's regularity structures. Regularity structures are a framework for solving the equations of **stochastic quantisation** (7-09; here only the name; a method that solves a time-evolution equation shaken by noise and tries to obtain the measure as its distribution after a long time), and their renormalisation is written with a pair of Hopf algebras and a variant of the Birkhoff decomposition (abstract of Bruned–Hairer–Zambotti 2019). Regularity structures have been used to

construct local-in-time solutions (solutions only for a short time) of the stochastic quantisation equation of three-dimensional Yang–Mills–Higgs (abstract of Chandra–Chevyrev–Hairer–Shen 2024). Known What they can handle is subcritical equations (the non-linear term weakens at shorter scales), and four-dimensional Yang–Mills is **critical** (the strength of the non-linear term does not depend on scale), so it lies outside.

What this chapter can say

	Content
What is said	The five formulations and the theorems quoted (classification of extensions, the main theorem on Birkhoff decomposition, Costello's Counterterm Uniqueness Theorem and Classification Theorem for Families of Effective Actions) Known . In the toy model: that subtracting only the whole leaves poles containing L and that Bogoliubov's prescription removes them; that for 8 trees the counterterms are local and the values are polynomials in L. That the difference of extensions of $\lvert x\rvert^{-4}$ is a constant multiple of δ. The pole of the bubble Computed
Conditional	The toy model has nesting only, and locality is checked within it (in general it rests on the theorems in the literature). The divergence of the perturbation series and $\Delta \propto \Lambda$ are expectations Physics
What is not said	Multi-loop calculations in four-dimensional Yang–Mills. The sum of the series. The non-perturbative measure, reflection positivity, the mass gap. The content of the proofs in literature cited by bibliographic data only

Sources and reproduction

Item	Kind	Source / tools
Read in full text	Known	Jaffe–Witten, Quantum Yang–Mills Theory (the Clay problem statement) p.6, p.11. Panzer, arXiv:1202.3552 (2012) § 3.2–3.4. Ebrahimi-Fard–Kreimer, J. Phys. A 38 (2005) R385 § 1–2. Brunetti–Fredenhagen, arXiv:math-ph/9903028 (Commun. Math. Phys. 208 (2000) 623), abstract, § 4, § 5.1–5.2 (numbering of formulas and theorems from the arXiv version). Costello, arXiv:0706.1533 (2007) § 1
Read in abstract only	Known	Kreimer, Adv. Theor. Math. Phys. 2 (1998) 303. Connes–Kreimer, Commun. Math. Phys. 199 (1998) 203; 210 (2000) 249. Kreimer, Commun. Math. Phys. 204 (1999) 669. Brouder, Eur. Phys. J. C 12 (2000) 521. Bruned–Hairer–Zambotti, Invent. Math. 215 (2019) 1039. Chandra–Chevyrev–Hairer–Shen, Invent. Math. 237 (2024) 541 (arXiv:2201.03487)

Item	Kind	Source / tools
Bibliographic data only	Known	Bogoliubov–Parasiuk, Acta Math. 97 (1957) 227. Hepp, Commun. Math. Phys. 2 (1966) 301. Zimmermann, ibid. 15 (1969) 208. Epstein–Glaser, Ann. Inst. H. Poincaré A 19 (1973) 211. Bollini–Giambiagi, Nuovo Cim. B 12 (1972) 20. 't Hooft–Veltman, Nucl. Phys. B 44 (1972) 189 (standard use in gauge theories not checked). 't Hooft, ibid. 33 (1971) 173. Lee–Zinn-Justin, Phys. Rev. D 5 (1972) 3121. Becchi–Rouet–Stora, Ann. Phys. 98 (1976) 287. Costello, Renormalization and Effective Field Theory, AMS (2011) (from the publisher's description). Butcher, Math. Comp. 26 (1972) 79
Equations (3)(4), Table 1	Computed	python3 (Laurent series in z computed exactly with the standard library's rationals; 8 trees, MS and MOM)
Figure 2, the numbers in § 03, § 04, § 06 and § 07	Computed	python3 (SciPy numerical integration and the Γ function, closed formulas). The JavaScript of Figure 2 was checked against the python3 values with node

Next chapter to read: 4-06 Asymptotic freedom. Detours: 7-12, 4-12.

Asymptotic freedom — the sign that weakens the coupling at short distances, and what is inside 11N/3

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — The β function of Yang–Mills is negative, and the coupling weakens logarithmically at shorter distances. The coefficient 11/3 splits into 4 from spin and −1/3 from orbital motion. b_0, b_1 do not depend on the definition of the coupling, and on that footing lattice numbers can be compared with the prediction.

Prerequisite chapters — 4-04 (the running coupling). Helpful: 17-01, 3-02, 5-01. $\hbar = c = 1$, $\mathrm{tr}(T^a T^b) = \frac{1}{2}\delta^{ab}$. We write the β function as $\beta(g)$ and the inverse lattice coupling as $\beta_{\mathrm{L}} = 2N/g_0^2$, to keep them apart.

Said first — The coefficients are perturbative calculations (Known here means "known as a perturbative calculation"). The four-dimensional continuum theory was built by the construction published in 2026-11, and it includes the one-loop running of the coupling and several short-distance checks as statements (the committee's assessment is that it is proved in Lean; 10-06, 6-01). Agreement with perturbation theory at all orders has not been shown (6-04). The comparison with the lattice in this chapter is a comparison of numbers, and does not in itself show that the continuum limit exists. This site shows neither the mass gap nor the construction of the continuum theory.

Order of this chapter

1. The running coupling — the sign of the β function
2. The one-loop coefficient — four ways of writing it (Table 1)
3. What is inside 11/3 — spin and orbital motion (Figure 1, Table 2) / Detour: the vacuum is a paramagnet
4. Adding matter — Coleman–Gross
5. Two loops and universality — b_0, b_1 do not depend on the definition
6. How it was discovered — a timeline (Table 3)
7. Agreement with the lattice — the ratio of Λs and r_0/a (Figure 2, Table 4)

8. Asymptotic freedom in the problem statement — what it says and what it does not (Table 5)

01

The running coupling — the sign of the β function

The renormalised coupling g has a value only once the scale μ at which it is measured is fixed (a momentum; as a distance, roughly $1/\mu$) (4-04). How it changes is the **β function**, and the right-hand side is an expansion in small g.

$$\beta(g) = \mu \frac{dg}{d\mu} = -b_0\, g^3 - b_1\, g^5 - \cdots \quad (1)$$

If $\beta > 0$, the coupling grows at short distances (large μ); QED and four-dimensional ϕ^4 are on this side (7-08). If $b_0 > 0$, it weakens at short distances. This is **asymptotic freedom**. Dropping b_1 and higher and integrating, $1/g^2$ becomes a linear function of $\ln \mu$ (the same formula as in 17-01).

$$\frac{1}{g^2(\mu)} = \frac{1}{g^2(\mu_0)} + 2b_0 \ln \frac{\mu}{\mu_0} \quad (2)$$

Formula (2) means that $1/g^2$ grows in proportion to the logarithm of the factor by which μ has grown. Whether μ is multiplied by 10 or by 100, it grows in fixed steps, and the coupling weakens only slowly.

Example. Writing α for the square of the coupling divided by 4π, one-loop QED with electrons only gives $d(1/\alpha)/d\ln\mu = -2/(3\pi) \approx -0.212$ (a rewriting of the formula in § 2.4.1 of Tong's lecture notes), and pure SU(3) gives $+1.751$. In the latter, $1/\alpha$ grows by about 4.03 each time μ grows by a factor of 10 (0-03) Computed .

02

The one-loop coefficient — four ways of writing 11N/3

For a gauge group with n_f Dirac particles (representation R_f) and n_s complex scalars (representation R_s) added, the one-loop coefficient is as follows (§ 2.4 and § 2.7 of Tong's lecture notes; formula (5.3) of 't Hooft 1999 is the same formula with a different normalisation) Known .

$$16\pi^2\, b_0 = \frac{11}{3}\, C_A - \frac{4}{3}\, T(R_f)\, n_f - \frac{1}{3}\, T(R_s)\, n_s \quad (3)$$

$T(R)$ is the number fixed by $\mathrm{tr}_R(T^a T^b) = T(R)\,\delta^{ab}$: $\frac{1}{2}$ for the fundamental representation and C_A for the adjoint representation (N for SU(N); 1-04). For pure SU(N), $b_0 = \frac{11N}{48\pi^2}$. The apparent numbers change with the choice of variable, so we list them in Table 1 (the four rows say the same thing).

Table 1 Four ways of writing the one-loop coefficient of pure SU(N)

Variable	Flow equation	Coefficient	SU(2)	SU(3)
g	$\mu\, dg/d\mu = -b_0 g^3$	$\frac{11N}{48\pi^2}$	0.04644	0.06966
$1/g^2$	$d(1/g^2)/d\ln\mu = 2b_0$	$\frac{11N}{24\pi^2}$	0.09288	0.1393
$\alpha = g^2/(4\pi)$	$\mu^2 d\alpha/d\mu^2 = -b_0^{(\alpha)}\alpha^2$	$\frac{11N}{12\pi}$	0.5836	0.8754
$\beta_{\mathrm{L}} = 2N/g^2$	$d\beta_{\mathrm{L}}/d\ln\mu = \kappa$	$\kappa = 4Nb_0 = \frac{11N^2}{12\pi^2}$	0.3715	0.8359

Computed Evaluated with python3. $b_0^{(\alpha)} = 33/(12\pi)$ for SU(3) agrees with formula (9.3) of the PDG and the formula for b_0 right after it.

The last row is in the language of the lattice (5-01). The one-loop prediction is that the same physics is kept if β_{L} is raised by $\kappa \ln 2$ (0.258 for SU(2), 0.579 for SU(3)) each time the lattice spacing a is halved Computed . $\kappa = 11/(3\pi^2)$ for SU(2) is the same number as the slope κ in the record of this book's investigation, and is the known one-loop coefficient written in lattice conventions (23-02) Known .

03

What is inside 11/3 — 4 from spin and −1/3 from orbital motion

Why is the sign reversed? The **background field method** places a slowly varying external field (the background field), integrates out the quantum fluctuations around it at one loop, and looks at how the coefficient $1/g^2$ of the action of the background field changes. Calculated this way (§ 2.4.2 of Tong's lecture notes), the gluon contribution splits in two. The term that counts the components of the field like scalars gives, after the excess is subtracted with ghosts, $-\frac{1}{3}$ in the screening direction, the same as in electromagnetism. The term that comes from spin 1 (the lecture notes' 'the paramagnetic piece') gives $+4$ in the antiscreening direction. The overall minus sign comes from here ('This is where the overall minus sign comes from') Known .

$$\frac{11}{3} = 4 - \frac{1}{3} \quad (4)$$

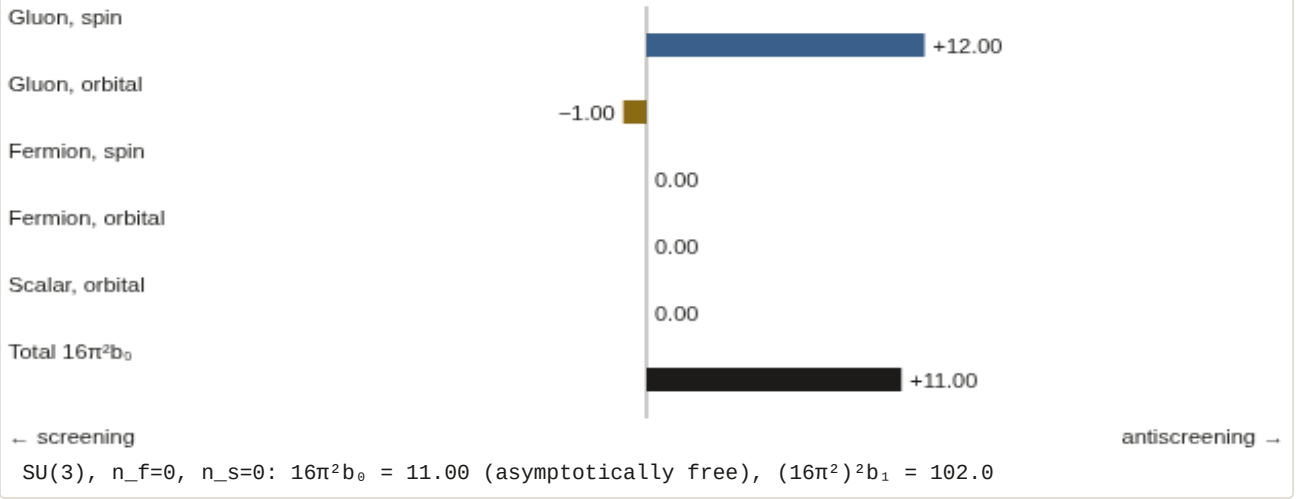
The photon carries no charge and takes no part in the vacuum polarisation of QED, but the gluon carries colour charge, so it does take part (3-02). Wilczek's Nobel lecture also writes 'it is the effect of virtual gluons that is responsible for antiscreening'.

The same split can be used for matter. We count each state of a massless particle (spin s, component s_z along the magnetic field) as contributing to $16\pi^2 b_0$ the amount

$$\tfrac{1}{2}\,(-1)^{2s}\big((2s_z)^2-\tfrac{1}{3}\big)\,T(R) \quad (5)$$

each. $(2s_z)^2$ is the spin (paramagnetic) part, $-\frac{1}{3}$ the orbital (diamagnetic) part, and $(-1)^{2s}$ the sign of a fermion loop. A gluon ($s_z=\pm1$) gives $\frac{11}{3}C_A$, a Dirac particle $-\frac{4}{3}T$, and a complex scalar $-\frac{1}{3}T$, and all of formula (3) comes out (Table 2). In this chapter we set down formula (5) as a way of counting that reproduces the known coefficients. That this rule reproduces all the coefficients of formula (3) was checked on this machine Computed . The derivation of the rule itself is not treated in this chapter. The analogy that views asymptotic freedom as paramagnetism is said to go back to Nielsen 1981 (Tong's note on the literature; Tong also mentions a private attribution to 't Hooft).

Figure 1 Moving the sliders for N, the number n_f of Dirac particles in the fundamental representation and the number n_s of complex scalars changes the contributions to $16\pi^2 b_0$ and their total. The right (positive) is the asymptotically free side. "Play" moves n_f back and forth



SU(3), n_f=0, n_s=0: $16\pi^2 b_0$ = 11.00 (asymptotically free), $(16\pi^2)^2 b_1$ = 102.0

Table 2 Contribution per particle by formula (5) (before multiplying by $T(R)$)

Particle	States	Spin	Orbital	Total = known coefficient
Gluon ($\times C_A$)	2	+4	−1/3	+11/3
Dirac particle	4	−2	+2/3	−4/3
Complex scalar	2	0	−1/3	−1/3

Computed The numbers in this figure and table were computed on this machine (formula (5) evaluated with python3 rationals, agreeing with the known coefficients; the figure adds up the same formula on the spot). If JavaScript is disabled, this table can be read instead.

Detour: the vacuum is a paramagnet

A subsection you may skip.

In the relativistic vacuum the speed of light is 1, so the product of permittivity and permeability is $\varepsilon\mu_{\mathrm{m}} = 1$. If it is paramagnetic ($\mu_{\mathrm{m}} > 1$), then $\varepsilon < 1$, and charge is antiscreened (Tong § 2.4.1). In a free electron gas, Pauli paramagnetism from spin and Landau diamagnetism from orbital motion are in the ratio $1 : -\frac{1}{3}$ (Tong's statistical physics § 3.6), the same ratio as the bracket $1 - \frac{1}{3}$ in formula (5). For fermions in the vacuum, however, the loop sign $(-1)^{2s}$ flips the whole, so the contribution of the electron is on the screening side ($-\frac{4}{3}$ in Table 2). For the gluon, $4 - \frac{1}{3}$. Speaking in metaphor, the vacuum of gluons is a strong paramagnet. Strictly, "spin" and "orbital" in formula (5) are names of terms in the calculation, not an observable such as the magnetisation of the vacuum.

04

Adding matter — $n_f < 11N/2$ and Coleman–Gross

Matter lowers b_0. With n_f Dirac particles in the fundamental representation, formula (3) is $16\pi^2 b_0 = \frac{11N}{3} - \frac{2n_f}{3}$, and the theory is asymptotically free only when $n_f < 11N/2$ (less than 16.5 for SU(3)) Computed . 't Hooft's recollection (1999) has the same number. There are 6 real quarks, and QCD is on the asymptotically free side.

For an abelian group $C_A = 0$, so only the matter terms remain in (3), and $b_0 < 0$ (the direction of QED in § 01). For $U(1)_Y$ of the Standard Model (normalised by $Q = T_3 + Y$; the hypercharges are in Table 1 of 3-07), replacing $T(R)$ by Y^2 and counting one left-handed Weyl component as half a Dirac particle, the three generations of fermions ($\sum Y^2 = 10/3$ per generation) and the Higgs doublet give

$$16\pi^2\, b_0\big|_{U(1)_Y} = -\frac{2}{3}\cdot 3\cdot\frac{10}{3} - \frac{1}{3}\cdot 2\cdot\frac{1}{4} = -\frac{41}{6}$$

and at one loop the coupling grows at short distances Computed . This is a different number from b_1 in § 05 (the two-loop coefficient).

For theories without Yang–Mills fields, the abstract of Coleman and Gross (1973) states 'No renormalizable field theory without non-Abelian gauge fields can be asymptotically free.' (a claim at the level of calculating β functions) Known . § 6.2 of the problem statement also places the fact that four-dimensional ϕ^4 is not asymptotically free next to triviality (7-08), and calls Yang–Mills 'the natural candidate for four-dimensional CQFT'. This is one reason why Yang–Mills is the protagonist of the problem.

05

Two loops and universality — b_0, b_1 do not depend on the definition

The two-loop coefficient was found by Caswell and by Jones, each in a 1974 paper [Known] . For pure SU(N) it is positive, and rewriting the PDG formula for α in terms of g gives the following form (checked for SU(3) [Computed]).

$$b_1 = \frac{34}{3}\,\frac{N^2}{(16\pi^2)^2} \quad (6)$$

b_0 and b_1 do not change when the definition of the coupling (the renormalisation scheme) is changed. Changing to another coupling $\tilde{g} = g + c\,g^3 + d\,g^5 + \cdots$ and rewriting $\mu\, d\tilde{g}/d\mu = (1 + 3cg^2 + 5dg^4)\beta(g)$ in terms of $\tilde{g}$ gives

$$\tilde{\beta}(\tilde{g}) = -b_0\tilde{g}^3 - b_1\tilde{g}^5 - \tilde{b}_2\tilde{g}^7 - \cdots$$
$$\tilde{b}_2 = b_2 - 2c\,b_1 + (2d - 3c^2)\,b_0 \quad (7)$$

and c, d enter only from $\tilde{b}_2$ on (checked with series of rationals [Computed]). The PDG also writes that b_2 and beyond are 'renormalization-scheme-dependent', and Capitani et al. 1999 call b_0, b_1 'The universal coefficients' [Known] . This universality is the footing for § 07.

Example. For pure SU(3), starting from the same integration constant Λ (4-07), α at $\mu/\Lambda = 10$ is 0.248 at one loop and 0.186 at two loops, and even at 10^{10} it is 0.0248 and 0.0232, a difference of 7% [Computed] . The correction is a "log of a log" and is slow to disappear. Adding matter to make $b_1 < 0$, the β function vanishes at $g_*^2 = -b_0/b_1$, and for SU(3) with $n_f = 16$ an infrared fixed point with $\alpha_* \approx 0.042$ appears [Computed] (Banks–Zaks 1982 [Physics]). The coefficients have been computed up to five loops (the PDG's list of references; for the meaning of the series as a whole, 4-12).

06

How it was discovered — the sign had been seen several times

The negative sign had appeared in calculations before 1973. According to 't Hooft's recollection (1999), at the time it was thought to be universal that the β function is positive, and quantum field theory was thought unable to explain Bjorken scaling (the regularity, in high-energy scattering, that the constituents inside a hadron appear to move almost freely at short distances).

Table 3 A timeline of asymptotic freedom

Year	Event	Source
1964	Vanyashin–Terentyev: a negative sign for charged vector mesons, put down to non-renormalisability (paper from 1965)	't Hooft 1999

Year	Event	Source
1969	Khriplovich: the opposite sign for Yang–Mills in Coulomb gauge, not connected with asymptotic freedom	Same
1972	At the Marseille conference 't Hooft writes the one-loop formula on the blackboard (not made into a paper)	Same
1973	Gross–Wilczek, Politzer, Coleman–Gross	Abstract
1974	Caswell, Jones: two loops	Abstract
1980	Creutz: lattice numbers. Hasenfratz–Hasenfratz: lattice and continuum Λ	Abstract
2004	Nobel Prize in Physics to Gross, Politzer and Wilczek	nobelprize.org

Politzer's title carried a question mark: 'Reliable Perturbative Results for Strong Interactions?' Known . 't Hooft writes of the 1973 papers 'I was surprised, not about the result, which I had known all along, but about the stir they caused', and goes on to say that the real problem is confinement. The problem of this book also lies beyond that (17-01).

07

Agreement with the lattice — the running of the bare coupling and the ratio of Λs

On the lattice (5-01) the lattice spacing a is the cutoff, and the bare coupling g_0 runs with a. By the universality of § 05 its β function also begins with the same b_0, b_1, and integrating to two loops gives

$$a\Lambda_{\mathrm{L}} = (b_0 g_0^2)^{-b_1/(2b_0^2)}\, e^{-1/(2b_0 g_0^2)}\, \left(1 + O(g_0^2)\right) \quad (8)$$

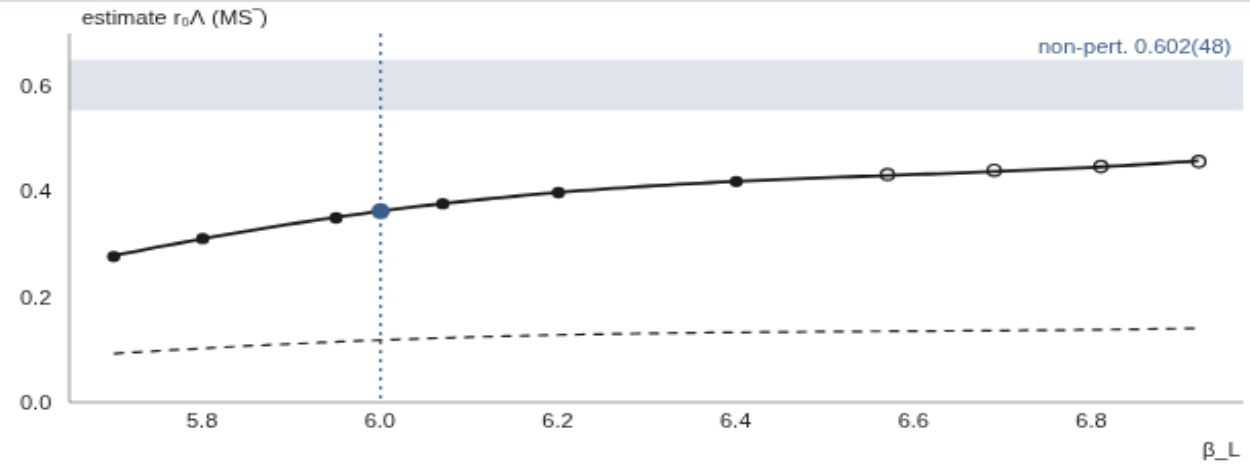
(the form of formula (2.7) of Capitani et al. 1999 with $\mu = 1/a$) Known . We compare the lattice Λ_{L} with the Λ of the continuum MS scheme. MS is the standard definition of the coupling in continuum theory, which in dimensional regularisation subtracts only the poles of the divergences and the constants that come with them (4-05). The two Λ differ only by a constant factor fixed at one loop; for the Wilson action $\Lambda_{\overline{\mathrm{MS}}}/\Lambda_{\mathrm{L}} = 28.81$ (SU(3)) and 19.82 (SU(2)). Computed from the coefficient k_1 of Lüscher–Weisz 1995, the SU(3) value agreed with 28.80934 in formula (4.1) of the same paper Computed . That formula (8) holds for lattice numbers is called **asymptotic scaling** (5-05; an early test is Creutz 1980).

At today's precision — r_0/a for pure SU(3)

The length standard r_0 is fixed from the force F between a pair of static colour charges by $r_0^2 F(r_0) = 1.65$ (17-01). The naive estimate of $r_0\Lambda_{\overline{\mathrm{MS}}}$, obtained by multiplying r_0/a

of Necco–Sommer 2002 by formula (8) and the ratio of Λs, is a quantity that should be constant if asymptotic scaling holds.

Figure 2 Moving the slider for β_L shows below the r_0/a at that point and the step in β_L that halves a (data and two loops). The curves are the estimate of $r_0\Lambda_{\overline{\rm MS}}$ (solid two loops, dashed one loop), the points are lattice values (filled: direct values of r_0/a; open: converted from r_c/a, the * in Table 4), and the band is the non-perturbative value



β_L = 6.00: r₀/a = 5.37, step: data 0.475, two loops 0.614, estimate r₀Λ = 0.363

Table 4 Pure SU(3), Wilson action: r_0/a, the estimate of $r_0\Lambda_{\overline{\rm MS}}$, and the step that halves a

β_L	r_0/a	Estimate (two loops)	Step (data)	Step (two loops)
5.70	2.922	0.277	0.354	0.616
5.95	4.898	0.350	0.456	0.614
6.20	7.380	0.398	0.524	0.613
6.40	9.74	0.419	—	0.612
6.57	12.18*	0.432	—	0.611
6.92	19.13*	0.457	—	0.609

Computed The numbers in this figure and table were computed on this machine. r_0/a is from Table 1 of Necco–Sommer 2002 (* converts r_c/a with $r_c/r_0 = 0.5133$), and the step (data) is from the interpolation formula (2.6) of the same paper (— is out of range). The estimate is $(r_0/a) \times a\Lambda_L \times 28.81$. The values in the figure agree with python3. If JavaScript is disabled, this table can be read instead.

The estimate keeps growing from 0.28 to 0.46 and has not yet reached the non-perturbative value 0.602(48). The expansion in the bare coupling converges slowly (with one loop alone, 0.09 to 0.14). Still, the direction and the order of magnitude agree, and the higher β_L, the closer to the prediction. In the method that measures the running coupling non-perturbatively with the size of the box as the scale (Lüscher–Weisz–Wolff 1991), the measured coupling is 'closely matched by the perturbative evolution', and 0.602(48) comes from there (Capitani et al. 1999) Known . All of these, however, are comparisons of finite-lattice numbers with perturbation theory, and show neither the existence nor the uniqueness of the continuum limit (5-09, 16-08).

08

Asymptotic freedom in the problem statement — what it says and what it does not

The problem statement first introduces asymptotic freedom as a physical explanation (§ 1), and gives it two roles. One is a requirement: the correlation functions must agree at short distances with the predictions of asymptotic freedom and perturbative renormalisation theory (§ 4; quoted in 0-03 § 07). Its role is to check at short distances that the theory built is "Yang–Mills". The other is a tool: 'Asymptotic freedom ensures a decisive regularity in the case when classical Sobolev inequalities are borderline' (§ 2) Known . This can be read as saying that four is the borderline dimension in which classical estimates fall just short, and asymptotic freedom makes up the shortfall (1-12; for Bałaban's work, 7-07).

Table 5 What asymptotic freedom says, and what it does not

What it says	What it does not say
The coupling weakens logarithmically at short distances (as a perturbative calculation). A handle on ultraviolet control (7-05)	That this holds in a constructed theory. Behaviour at long distances (as $\mu \to \Lambda$, g becomes large and formula (2) cannot be used)
An integration constant Λ appears and does not depend on μ (4-07)	The mass gap. A quantity expected to be proportional to Λ has the form $e^{-1/(2b_0 g^2)}$ and shows up at no order of perturbation theory (4-12)
Asymptotic freedom in four dimensions needs non-abelian gauge fields	Confinement (17-02)

Asymptotic freedom says "short distances are easy", and at the same time says "at long distances there is no handle for an expansion". The difficulty of the problem lies on the latter side (7-10, 7-11).

What this chapter can say

	Content
What is said	Formulas (3)(6)(7)(8), the threshold on the amount of matter, Coleman–Gross, as perturbative calculations Known . The checks of formulas (5)(7), the numbers in the tables and figures, the ratio of Λs Computed
Conditional	The agreement of the lattice with the prediction, as a comparison of finite-lattice numbers with perturbative formulas. The infrared fixed point within the range of the two-loop formula Physics

	Content
What is not said	That asymptotic freedom agrees with perturbation theory at all orders in the constructed four-dimensional theory (the statements of the construction published in 2026-11 go up to the one-loop running; 10-06). The uniqueness of the infinite-volume limit (6-04). "Spin" and "orbital" in formula (5) are names of terms. The events of 1964 to 1969 are taken from recollections (the original papers not read). Confinement and the mass gap

Sources and reproduction

Item	Kind	Source / tools
Read in full text	Known	PDG: J. Huston, K. Rabbertz, G. Zanderighi, in S. Navas et al., Phys. Rev. D 110 (2024) 030001, § 9.1.1. G. 't Hooft, Nucl. Phys. B Proc. Suppl. 74 (1999) 413, arXiv:hep-th/9808154. F. Wilczek 2005, arXiv:hep-ph/0502113. S. Capitani, M. Lüscher, R. Sommer, H. Wittig 1999, arXiv:hep-lat/9810063. S. Necco, R. Sommer 2002, arXiv:hep-lat/0108008. M. Lüscher, P. Weisz 1995, arXiv:hep-lat/9502001, hep-lat/9505011. D. Tong's lecture notes *Gauge Theory* § 2.4 and *Statistical Physics* § 3.6. A. Jaffe, E. Witten (the Clay problem statement)
Read in abstract only	Known	Gross–Wilczek, PRL 30 (1973) 1343; Politzer, PRL 30 (1973) 1346; Coleman–Gross, PRL 31 (1973) 851; Caswell, PRL 33 (1974) 244; Jones, Nucl. Phys. B 75 (1974) 531; Banks–Zaks, Nucl. Phys. B 196 (1982) 189; Creutz, Phys. Rev. D 21 (1980) 2308; Hasenfratz–Hasenfratz, Phys. Lett. B 93 (1980) 165; Lüscher–Weisz–Wolff, Nucl. Phys. B 359 (1991) 221. The prize citation is on nobelprize.org
Bibliographic data only	Known	Vanyashin–Terentyev, Zh. Eksp. Teor. Fiz. 48 (1965) 565; Khriplovich, Yad. Fiz. 10 (1969) 409 (from 't Hooft's list of references); N. K. Nielsen, Am. J. Phys. 49 (1981) 1171; the literature on higher coefficients (from the PDG's list of references)
Formulas (5)(7), Tables 1, 2 and 4, Figures 1 and 2, the ratio of Λs, the numerical examples, $-41/6$ for $U(1)_Y$ in § 04	Computed	python3 (series of rationals, bisection; each under 1 second). The JavaScript values of the figures were checked against the python3 values with node

Next chapter to read: 4-07 Dimensional transmutation and Λ (the Λ that appears when formula (2) is extended to the long-distance side). The continuation on the lattice side is 5-05.

Dimensional transmutation and Λ — quantum theory brings one scale into a classical theory without scales

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — From a classical theory without scales, one scale Λ appears in quantum theory as the integration constant of the running coupling. Masses are expected to be constant multiples of Λ, and their form $e^{-1/(2b_0 g^2)}$ does not show up in perturbation theory. The two-dimensional delta well as an exact example.

Prerequisite chapters — 4-06 (the β function and b_0, b_1). Helpful: 1-12, 2-05, 3-05. The notation is the same as in 4-06 ($\hbar = c = 1$, $\beta(g) = \mu\, dg/d\mu$, $b_0 = 11N/(48\pi^2)$).

Said first — Dimensional transmutation is a physical argument read off from the perturbative renormalisation group, and in four dimensions it is not a theorem. The renormalisation group says only "if there is a mass, it is a constant multiple of Λ", and does not say that the coefficient is non-zero (the mass gap). This site shows neither the mass gap nor the construction of the continuum theory.

Order of this chapter

1. The classical theory has no scale
2. Λ is the integration constant of the running (Table 2)
3. One parameter disappears (Figure 1)
4. Physical quantities are constant multiples of Λ (Table 3)
5. $\exp(-1/g^2)$ does not show up in perturbation theory (Table 4)
6. The two-dimensional delta well (Figure 2, Table 5; Detour: Cooper pairs in a metal)
7. Λ changes with the convention (Table 6)
8. How scale symmetry is broken
9. What can be said and what cannot (Table 7)

01

The classical theory has no scale

With $\hbar = c = 1$, length is the inverse of mass, and every quantity can be measured in powers of mass. The dimension of the coupling in the action $S = \frac{1}{2g^2}\int \operatorname{tr} F_{\mu\nu}F_{\mu\nu}\, d^dx$ is

$$[g^2] = (\text{mass})^{4-d} \quad (1)$$

(2-04, 1-12). In three dimensions g^2 is itself a mass, and one can write $m = c\,g^2$ (7-06). In four dimensions g is dimensionless, so no mass can be made from the classical ingredients. Classical waves have the dispersion $\omega = |\mathbf{k}|$, and the problem statement also writes 'massless waves that travel at the speed of light' (§ 1) Known . There are no static finite-energy solutions (lumps) either (3-03; Deser 1976, Coleman 1977) Known .

Example. The action $8\pi^2/g^2$ of the BPST instanton does not depend on its size ρ (3-05). Classical solutions have no preferred size. If $\Delta > 0$ in the quantum theory, that scale must come from outside the classical action. This is what 'the mass gap is not classically visible' in § 6.6 of the problem statement is about (§ 05).

02

Λ is the integration constant of the running

The $\beta(g) = \mu\, dg/d\mu$ of 4-06 is a first-order ordinary differential equation. Separating the variables and integrating gives

$$\ln\frac{\mu}{\Lambda} = \int^{g(\mu)} \frac{dg'}{\beta(g')} \quad (2)$$

The freedom in the lower limit on the right-hand side, taken to the left-hand side as a constant, is Λ; being an integration constant, it has the same value whichever point of the flow it is computed from. At one loop ($\beta = -b_0 g^3$), choosing the constant to be 0 gives

$$\Lambda = \mu\, \exp\!\Big(-\frac{1}{2b_0\, g^2(\mu)}\Big) \quad (3)$$

and differentiating $\ln\Lambda$ with respect to $\ln\mu$ gives $1 + \beta/(b_0 g^3) = 0$. Rewritten for SU(2) on the lattice ($\beta = 4/g^2$) it is $1/(a\Lambda) \propto e^{3\pi^2\beta/11}$, the same formula as the one-loop part of the bookkeeping length ξ_{RG} of Part 23 (23-01 The line of construction and the

diffusive-relaxation-time assumption) Known . From two loops on, a convention for the lower limit is needed. The one widely used is formula (2.7) of Capitani et al. 1999.

$$\Lambda = \mu\,(b_0 g^2)^{-b_1/(2b_0^2)}\, e^{-1/(2b_0 g^2)} \exp\!\Big(-\int_0^{g} \Big[\frac{1}{\beta(x)} + \frac{1}{b_0 x^3} - \frac{b_1}{b_0^2 x}\Big]\, dx \Big) \quad (4)$$

The bracket is finite as $x \to 0$, so 0 can be used as the lower limit. The PDG review explains Λ as 'a constant of integration, which corresponds to the scale where the perturbatively-defined coupling would diverge' Known .

Example (Table 2). Treating the two-loop β as exact, we integrated the flow numerically for SU(3) from $g^2(\mu_0) = 2$. Formula (4) is constant to 9 digits. Formula (3) drifts by a factor of 2.7 while μ grows by a factor of 10^{16}, and the formula with only the prefactor power attached merely shrinks the difference from 4.6% to 0.43%. The numbers are given with their defining formulas.

Table 2 Three "Λ/μ_0" along the two-loop flow of pure SU(3)

μ/μ_0	$g^2(\mu)$	Formula (3)	With the power only	Formula (4)
1	2.000	0.02763	0.06341	0.066444376
10^{16}	0.1733	0.01028	0.06616	0.066444376

Computed python3 (RK4). The integral in formula (4) agrees between the closed form and numerical integration to 9 digits.

For how this book uses this, see 24-07 § 01 (reading the dimensionless coupling as the first source of the difficulty of four dimensions).

03

One parameter disappears — Λ is a unit

Turning formula (3) around, the whole flow is fixed by Λ alone.

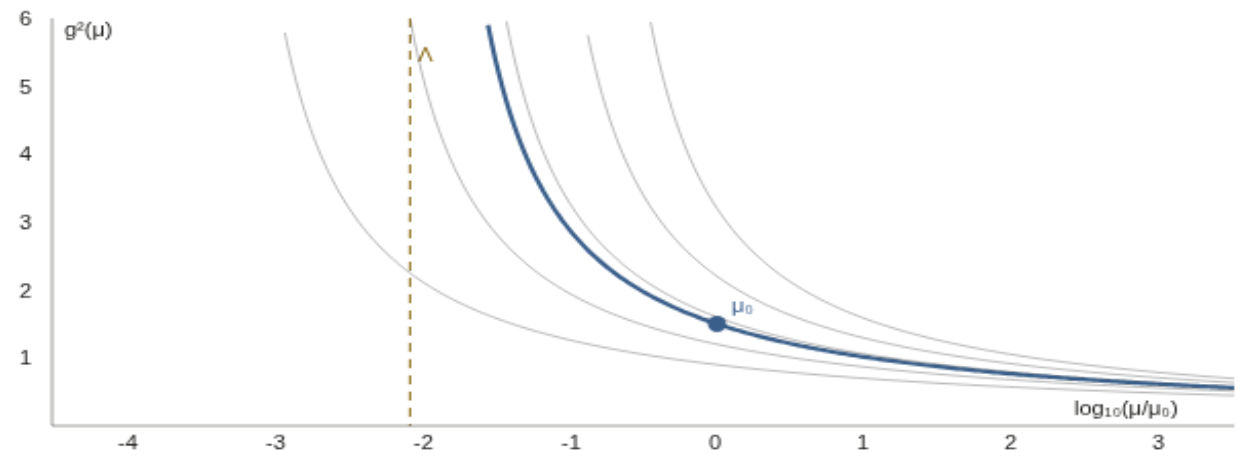
$$g^2(\mu) = \frac{1}{2b_0 \ln(\mu/\Lambda)} \qquad (\mu > \Lambda,\ \text{one loop}) \quad (5)$$

Λ enters only in the form μ/Λ. Doubling Λ is the same as halving the unit of mass, and the one-parameter family "choose the coupling at μ_0" is, up to units, a single theory. The dimensionless g has been replaced by the dimensionful Λ. This is called **dimensional transmutation** Physics .

Tong writes that this idea was known from the 1973 work of Coleman and E. Weinberg (Tong § 2.9). E. Weinberg's thesis of the same year writes, for the electrodynamics of a massless scalar field, that changing the vacuum expectation value $\langle\phi\rangle$ generated by quantum effects (the value of the scalar field in the vacuum, which serves as the scale of mass) is 'completely equivalent to changing the scale in which masses are measured'

Known . Pure Yang–Mills has no dimensionless coupling left, and the only remaining inputs are the group (N) and the θ angle (3-05).

Figure 1 Moving the slider for $g^2(\mu_0)$ shifts the one-loop flow of pure SU(3) (thick line) and Λ (dashed line) sideways (the thin lines are $g^2(\mu_0)$ = 0.9 to 3.2). "Horizontal axis μ/Λ" makes the six coincide. "Play" moves back and forth



g²(μ₀) = 1.50 → Λ/μ₀ = 8.35e-3

Computed The numbers in this figure were computed on this machine (the figure evaluates formulas (3)(5) on the spot). Λ/μ_0 of the thin lines ranges from 3.44×10^{-4} to 0.106, but g^2 at $\mu = 10\Lambda$, $10^3\Lambda$ and $10^6\Lambda$ is always 3.117, 1.039 and 0.520.

The scale comes from the cutoff. On the lattice Λ is made from the pair (a, g_0) (formula (8) of 4-06).

04

Physical quantities are constant multiples of Λ — nothing is said about the sign of the coefficient

Computing a physical quantity with the dimension of mass, dimensional analysis gives the form $m = \mu\, F(g(\mu))$. m does not depend on the choice of μ, so

$$\mu \frac{d}{d\mu}\big(\mu\, F(g(\mu))\big) = \mu\big(F(g) + \beta(g)\, F'(g)\big) = 0 \quad (6)$$

and the solution is $F(g) = C \exp\big(-\int^g dg'/\beta(g')\big)$. Comparing with formula (2),

$$m = C\,\Lambda \quad (7)$$

where C is a pure number. Formula (6) is the renormalisation group equation (4-04) for a physical quantity without anomalous dimension. For the string tension $\sigma = C'\Lambda^2$, and the ratio $m/\sqrt{\sigma}$ becomes a pure number fixed by N and θ alone, with no value left to adjust Physics . This is what "one scale" means (17-01, 17-12).

The solutions of formula (6) include $C = 0$. What can be said is only "if there is a mass, it is a constant multiple of Λ", and $C > 0$ cannot be said. The mass gap problem can be read as the problem of constructing the theory and showing $C > 0$ for the state next to the vacuum (0-03, 4-11).

Example (Table 3). Lattice numbers come out as dimensionless ratios (r_0 is the length of 4-06). $m(0^{++})$ is the mass of the lightest glueball with spin 0, positive parity and positive charge conjugation (5-08).

Table 3 Dimensionless ratios of pure SU(3)

Ratio	Value	Source
$r_0 \Lambda_{\overline{\text{MS}}}$	0.602(48)	Capitani et al. 1999
$r_0\, m(0^{++})$	4.21(11)(4)	Morningstar–Peardon 1999
$m(0^{++})/\sqrt{\sigma}$	3.405(21)	Athenodorou–Teper 2020
$m(0^{++})/\Lambda_{\overline{\text{MS}}}$	6.99(59)	Division of the two rows above (a rough guide combining values from different papers)

Computed The last row was divided with python3, propagating the errors as independent. Physics All are extrapolations from the lattice, to be read on the assumption that the continuum theory exists (17-01).

05

exp(−1/g²) does not show up in perturbation theory

Rewriting formula (7) at one loop gives

$$\frac{m}{\mu} = C\,\exp\!\Big(-\frac{1}{2b_0\, g^2(\mu)}\Big) \quad (8)$$

$f(x) = e^{-1/x}$ $(x > 0)$, $f(x) = 0$ $(x \le 0)$ can be differentiated any number of times at $x = 0$, and all its derivatives are 0. This is because $f^{(n)}(x) = P_n(1/x)\, e^{-1/x}$ ($P_1(u) = u^2$, $P_2(u) = u^4 - 2u^3$), and polynomials cannot beat e^{-u} Known . Expanding (8) in powers of g^2, all the coefficients are 0.

Perturbation theory (4-03) computes a power series in g^2, so a quantity of the form (8) looks like 0 at every order. § 6.6 of the problem statement writes that, since the gap is not classically visible, showing it may need a non-classical change of variables ('change of variables or “duality transformation.”') Known . The abstract of Gross–Neveu 1974 also says that a mass generated dynamically in an asymptotically free theory 'must depend on the coupling constants in a nonanalytic fashion, vanishing exponentially when these vanish' (at the level of the renormalisation group). The tunnelling splitting of the double well is of the same type (2-05).

Example (Table 4). As a rough guide with coefficients set to 1, the order $n^* = \ln(\Lambda/\mu)/\ln\alpha$ at which the n-th power of $\alpha = g^2/4\pi$ drops to Λ/μ is 2.8 at $g^2 = 1$, and grows without bound as $g \to 0$. The actual coefficients grow with the order, and the series is thought to diverge (4-12).

Table 4 Pure SU(3), one loop

g^2	Λ/μ	n^*	$\log_{10} e^{-8\pi^2/g^2}$
2	2.76×10⁻²	1.95	−17.15
1	7.63×10⁻⁴	2.84	−34.29
0.25	3.39×10⁻¹³	7.33	−137.16
0.1	6.71×10⁻³²	14.85	−342.91

Computed python3. The rightmost column (§08) agrees with $(\Lambda/\mu)^{11}$ in all digits. P_n was checked up to $n = 6$.

06

The two-dimensional delta well — exact dimensional transmutation

An exact example exists in two-dimensional quantum mechanics. In units with $\hbar^2/2m = 1$, consider the delta-function well in d dimensions

$$H = -\nabla^2 - g\,\delta^d(\mathbf{x}) \tag{9}$$

δ^d has the dimension of length to the power $-d$, and ∇^2 to the power -2, so g has the dimension of length to the power $d-2$, and is dimensionless at $d = 2$. This is the same situation as four-dimensional Yang–Mills. Taking the Fourier transform as $\hat\psi(\mathbf{k}) = \int \psi(\mathbf{x})\, e^{-i\mathbf{k}\cdot\mathbf{x}}\, d^dx$ (1-11), the transform of δ^d is 1. A bound state $E = -\kappa^2$ satisfies $(k^2 + \kappa^2)\,\hat\psi(\mathbf{k}) = g\,\psi(0)$, so dividing by $k^2 + \kappa^2$ and integrating gives

$$1 = g \int_{|\mathbf{k}|<K} \frac{d^dk}{(2\pi)^d}\,\frac{1}{k^2+\kappa^2} \tag{10}$$

(K is the momentum cutoff). For $d = 2$ the right-hand side is $(g/4\pi)\ln(1 + K^2/\kappa^2)$, which diverges logarithmically, as does one-loop four-dimensional Yang–Mills (4-04). Solving,

$$E_B = \kappa^2 = \frac{K^2}{e^{4\pi/g} - 1} \approx K^2\, e^{-4\pi/g} \tag{11}$$

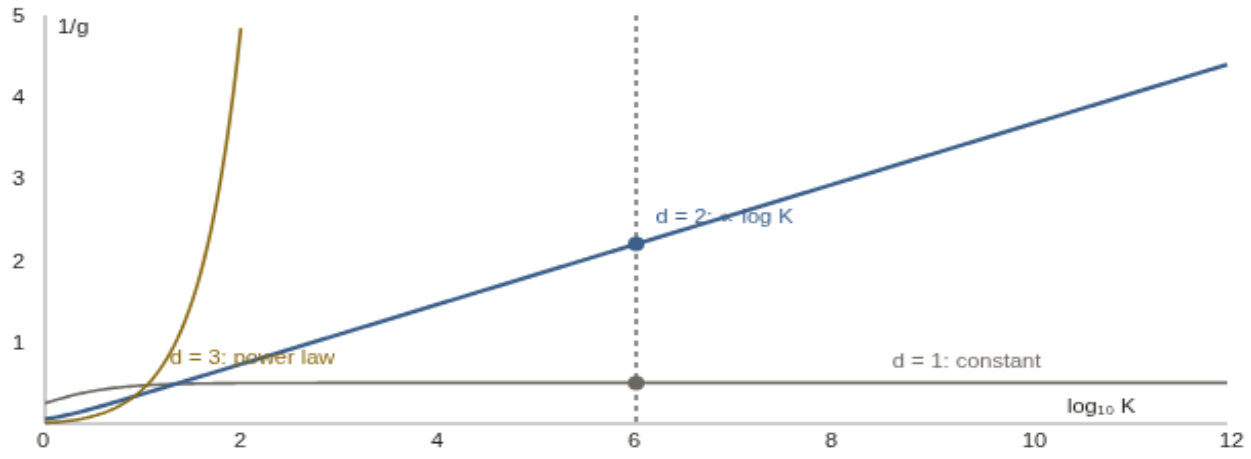
the same form as formula (8). Renormalisation is the operation of fixing E_B and taking $K \to \infty$, which sends the coupling towards 0 as $g(K) = 4\pi/\ln(1 + K^2/E_B)$. For $K^2 \gg E_B$,

$$\beta(g) = K\frac{dg}{dK} = -\frac{g^2}{2\pi} \quad (12)$$

the sign is negative, that is, asymptotically free. The only parameter left in the limit is E_B. Up to here this is an elementary calculation with a momentum cutoff Computed .

On the mathematical side, it is known that in the limit where the cutoff is removed the Hamiltonian converges (in the sense of the resolvent $(H - z)^{-1}$) to a self-adjoint operator with exactly one bound state (energy $-E_B$) (Albeverio et al., Part I, Chapter 5; bibliographic data only) Known . The limit there is taken by shrinking a short-range potential while renormalising the coupling logarithmically, a different route from the momentum cutoff above. As teaching material there is Gosdzinsky–Tarrach 1991 (bibliographic data only).

Figure 2 Moving the slider for the cutoff K shows how the inverse 1/g of the coupling that keeps E_B = 1 moves for d = 1, 2, 3. For d = 2 it is proportional to log K (asymptotically free). "Play" moves K back and forth



K = 10^6.00 : g_1 = 2.0000 g_2 = 0.4548 g_3 = 1.97e-5, g_2 up 1%: E_B ×1.31

Table 5 The coupling $g_d(K)$ that keeps $E_B = 1$ (in units of $\sqrt{E_B}$)

K	$g1$	$g2$	$2\pi/\ln K$	$g3$
10	2.1355	2.7229	2.7288	2.314
10^6	2.0000	0.4548	0.4548	1.974×10^{-5}
10^{12}	2.0000	0.2274	0.2274	1.974×10^{-11}

Computed The numbers in this figure were computed on this machine. The closed form of formula (10) agrees with numerical integration to 11 digits, and $K\,dg/dK$ agrees with $-g^2/2\pi$ for $K \geq 10^4$ (ratio 1.000000). The factor by which E_B changes when g_2 is moved by 1% is 1.05 at $K = 10$ and 1.73 at 10^{12}.

For $d = 1$, $E_B = g^2/4$ is a power of g, and $g(K)$ settles at 2. For $d = 3$, g has to be brought to the critical value $2\pi^2/K$ as a power. These are the cases where the mass dimension of the coupling is positive, zero and negative; four-dimensional Yang–Mills is on the $d = 2$ side, and three-dimensional Yang–Mills on the $d = 1$ side (an analogy).

Detour: Cooper pairs in a metal

A subsection you may skip.

The same logarithm appears for pairs of electrons in a metal. In a metal, the electron states are filled up to the Fermi surface (the surface of the filled sea of electrons). Let ξ be the energy of one electron measured from the Fermi surface, and $N(0)$ the density of states per spin at the Fermi surface (the number of states per unit energy and unit volume). If two electrons on the surface feel an attraction of strength V only within the shell $0 < \xi < \omega_D$ (ω_D is the maximum frequency of lattice vibrations, the cutoff here), the binding condition, by the same steps as formula (10), is

$$1 = N(0)V \int_0^{\omega_D} \frac{d\xi}{2\xi + E_B} = \frac{N(0)V}{2} \ln\left(1 + \frac{2\omega_D}{E_B}\right) \quad (15)$$

Since $N(0)$ is constant, a logarithm appears, just as with $d^2k \propto d(k^2)$ in two dimensions. Solving gives $E_B = 2\omega_D / (e^{2/(N(0)V)} - 1) \approx 2\omega_D\, e^{-2/(N(0)V)}$: even a weak attraction has a bound state, and its size does not show up in the series in powers of V (Cooper 1956; bibliographic data only) Known . The 2 in the exponent comes from the convention of taking $N(0)$ per spin.

Computed python3. The closed form and numerical integration of formula (15) agree within double precision for $N(0)V = 0.05$ to 0.5.

07

Λ changes with the convention, the ratios do not

The value of Λ changes with the definition of the coupling (the scheme). If g of scheme A and $\tilde{g}$ of scheme B are related by $\tilde{g} = g + c\,g^3 + O(g^5)$, then $1/\tilde{g}^2 = 1/g^2 - 2c + O(g^2)$, so from formula (3)

$$\frac{\Lambda_B}{\Lambda_A} = e^{c/b_0} \quad (13)$$

Since b_0, b_1 do not depend on the scheme (4-06), this ratio holds exactly for the Λ of formula (4). The abstract of Celmaster–Gonsalves 1979 writes that the ratio of Λ in two schemes is 'a pure number which we show is exactly deducible from our one-loop results' Known .

Example (Table 6). MS (minimal subtraction: the scheme that subtracts only the poles appearing in dimensional regularisation; 4-04) and $\overline{\text{MS}}$, the PDG's standard (the scheme that also subtracts the constant $\ln 4\pi - \gamma_E$ along with the poles; Bardeen et al. 1978), differ only by the rescaling $\mu^2 \to \mu^2 e^{\gamma_E}/4\pi$, and the ratio is $\sqrt{4\pi e^{-\gamma_E}} = 2.656$. The ratio 28.81 to the lattice Wilson action (SU(3), 4-06) also comes from a shift in $1/g^2$

of only 0.468. The one-loop difference enters the exponent multiplied by $1/b_0$, so Λ changes by an order of magnitude or more.

Table 6 Changing the scheme and the ratio of Λs

Ratio	Group	c	Shift in $1/g^2$	e^{c/b_0}	Ratio from formula (4)
MS / MS	SU(3)	0.0681	−0.136	2.656	–
MS / lattice	SU(3)	0.2341	−0.468	28.81	28.8100
MS / lattice	SU(2)	0.1387	−0.277	19.82	19.8200

Computed python3. Building the β of scheme B from $\tilde{g} = g(1 + cg^2)$ and integrating formula (4) numerically, the ratio is constant to 6 digits for $\mu/\mu_0 = 1$ to 10^{16}. The ratio for the formula with the prefactor power only is slow: 6.76 → 23.08 (SU(3)).

Ratios of physical quantities such as $m/\sqrt{\sigma}$ do not depend on the scheme. $r_0\Lambda$ is a single number only once the scheme is named. To attach MeV, another quantity is borrowed (17-01).

08

How scale symmetry is broken — instantons and the trace anomaly

The classical scale symmetry of § 01 leaves traces in two places in the quantum theory. One is the size of instantons: ρ, free in the classical theory, is distinguished by the running of the coupling (Tong § 2.3.1– § 2.3.2). Using the coupling at the scale $1/\rho$, from formula (3) and $16\pi^2 b_0 = 11N/3$, at one loop

$$e^{-8\pi^2/g^2(1/\rho)} = (\rho\Lambda)^{11N/3} \quad (14)$$

holds identically (checked against Table 4 Computed). Small instantons are suppressed by a power of $\rho\Lambda$, and the suppression stops working at $\rho \sim 1/\Lambda$. The one-loop calculation is 't Hooft 1976 Physics .

The other is the trace of the energy–momentum tensor. The divergence of the conserved current of scale transformations (2-01) is the trace $T^{\mu}{}_{\mu}$, which is 0 classically in four dimensions (3-03). In the quantum theory a term in $F^a_{\mu\nu}F^{a\,\mu\nu}$ proportional to $\beta(g)$ remains (Collins–Duncan–Joglekar 1977; for the formula, 18-04) Physics . § 4 of the problem statement also includes 'the existence of a stress tensor' in the short-distance requirements Known .

09

What can be said and what cannot

On the lattice, a test from formula (7) is whether ratios of correlation lengths $\xi_{\text{lat}} = 1/(am)$ approach constant values (5-05, 5-09). We list the degree of certainty example

by example. The top four are two-dimensional models with a dimensionless coupling, and what has become a theorem is either a model with finitely many degrees of freedom or a form that keeps the cutoff.

Row by row, we separate what is a theorem from what is physics. For the delta well, the binding condition (10) with a cutoff is an elementary calculation, and the operator in the limit with the cutoff removed is in the mathematical literature. For the two-dimensional Gross–Neveu model (a model of interacting fermions), keeping the ultraviolet cutoff and taking the number of components N large (but finite), it is a theorem that the two-point function decays exponentially, that is, that a mass is generated. For the two-dimensional O(N) σ model (a model of fields valued in N-component vectors of length 1), mass generation in a regularised version is also a theorem, but there is still no construction of the theory with the cutoff removed. The $8/e$ of O(3) is a physical value obtained by matching the result of the Bethe ansatz (a method that uses the exact solvability of scattering in two-dimensional integrable models) with perturbation theory. For pure SU(3) in four dimensions, the ratio m/Λ comes only from lattice numbers. The construction of the continuum theory was given by the construction published in 2026-11 (at the level of the committee's assessment; 6-00), but $\Delta > 0$ has not been shown.

Table 7 Examples of dimensional transmutation, and their status in mathematics

Model	Scale	Status in mathematics
Two-dimensional delta well	$K^2 e^{-4\pi/g(K)}$	With a cutoff, an elementary calculation. The limit with the cutoff removed is a self-adjoint operator with exactly one bound state (Albeverio et al.; bibliographic data only)
Two-dimensional Gross–Neveu	Non-analytic mass	With a cutoff and large finite N, exponential decay is a theorem (Kopper–Magnen–Rivasseau 1995)
Two-dimensional O(3) σ model	$m = (8/e)\Lambda_{\overline{\text{MS}}}$	A physical value from matching the Bethe ansatz with perturbation theory (Hasenfratz–Maggiore–Niedermayer 1990)
Two-dimensional O(N) σ model (N large and finite)	Mass m	With an ultraviolet cutoff, mass generation (exponential decay of correlations) is a theorem (Kopper 1999). A construction at the level required did not exist as of the 2004 report (Douglas 2004)
Four-dimensional pure SU(3)	$m(0^{++}) \approx 7\Lambda_{\overline{\text{MS}}}$ (the rough guide of Table 3)	Lattice numbers. The construction of the continuum theory: the construction published in 2026-11 (at the level of the committee's assessment). $\Delta > 0$ is not proved

The problem statement does not mention Λ; it asks for $\Delta > 0$ and for agreement with the predictions of asymptotic freedom at short distances (0-03). If the theory exists and agrees with those predictions, in physics one considers that Λ can be read off from them Physics . The mass gap is then the claim that C, which shows up neither in perturbation theory nor in the classical theory, is positive. How this single scale is used is treated in 17-01 and 17-12.

What this chapter can say

	Content
What is said	The μ-invariance of formulas (2) to (4), the solution (7), that the Taylor coefficients of $e^{-1/x}$ are 0, formulas (10) to (15). These are Known calculations, and the numbers in the tables and figures are Computed
Conditional	"Masses are constant multiples of Λ" assumes that the theory exists and that the renormalisation group describes short distances correctly Physics . Table 3 is an extrapolation from the lattice. The limiting operator of the delta well is taken from the literature (text not checked). The correspondence for each d is an analogy
What is not said	That Λ is defined in four dimensions, $C > 0$, that the values of Table 3 exist as quantities of the continuum theory (they are extrapolations from the lattice, not theorems). This site shows neither the mass gap nor the construction of the continuum theory

Sources and reproduction

Item	Kind	Source / tools
The problem statement, the sentence on Λ, formula (4)	Known	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory" (Clay); PDG, PRD 110 (2024) 030001 § 9.1.1; S. Capitani et al., NPB 544 (1999) 669, hep-lat/9810063 (text of all the above). W. A. Bardeen et al., PRD 18 (1978) 3998 (abstract)
Origin, instantons	Known	D. Tong, *Gauge Theory* § 2.3.1– § 2.3.2, § 2.9 (text); E. J. Weinberg, thesis (1973), hep-th/0507214 Chapter IV (text); S. Coleman, E. Weinberg, PRD 7 (1973) 1888 (abstract only; the first appearance of the term dimensional transmutation not checked)
The classical theorems of § 01	Known	S. Deser, PLB 64 (1976) 463; S. Coleman, CMP 55 (1977) 113 'There are no classical glueballs' (abstracts of both)
Literature for § 04 to § 08	Known	D. J. Gross, A. Neveu, PRD 10 (1974) 3235; W. Celmaster, R. J. Gonsalves, PRD 20 (1979) 1420; G. 't Hooft, PRD 14 (1976) 3432; J. C. Collins et al., PRD 16 (1977) 438 (abstracts of all the above). S. Albeverio et al., *Solvable Models in Quantum Mechanics* (Springer, 1988) Part I, Chapter 5, P. Gosdzinsky, R. Tarrach, Am. J. Phys. 59 (1991) 70, and L. N. Cooper, Phys. Rev. 104 (1956) 1189 are bibliographic data only
Numbers and examples in Tables 3 and 7	Physics	C. Morningstar, M. Peardon, PRD 60 (1999) 034509; A. Athenodorou, M. Teper, JHEP 11 (2020) 172; P. Hasenfratz, M. Maggiore, F. Niedermayer, PLB 245 (1990) 522 (abstract); C. Kopper et al., CMP 169 (1995) 121; C. Kopper, CMP 202 (1999) 89 (the O(N) σ model with large finite N; abstract); M. R. Douglas's report to Clay (2004). Beyond the bibliographic data, based on the material of the checked survey
Tables 2 to 6, the figures, formula (14), formula (15)	Computed	python3 (RK4, scipy). The figures evaluate the same formulas in JavaScript. 28.81 and 19.82 are the values of 4-06

Next chapter to read: 4-08 Gauge fixing and Faddeev–Popov.

Gauge fixing and Faddeev–Popov — correct locally, and globally the signs cancel

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — Perturbation theory needs gauge fixing, and its procedure (Faddeev–Popov) produces a determinant and ghosts. After fixing, BRST symmetry remains and formally guarantees independence from the choice of gauge. When an orbit meets the condition surface many times (Gribov copies) the procedure breaks down, and on the lattice the cancellation of signs gives 0/0.

Prerequisite chapters — 4-03 (perturbation theory, propagators) and 3-04 (orbit space, the geometry of Gribov copies, Singer's theorem). Grassmann numbers (§ 03), BRST cohomology (§ 04) and the Euler characteristic (§ 08) are defined on the spot, as far as needed.

Said first — Faddeev–Popov and BRST are tools established at each order of perturbation theory, not a non-perturbative definition. The figures and tables are calculations in a planar model and for constant fields in a box. This chapter shows neither the mass gap nor the construction of the theory.

Order of this chapter

1. Why fix the gauge — the action is flat along the orbits, and no propagator can be made
2. A finite-dimensional model — "inserting 1" with rotations of the plane
3. In field theory — writing the determinant as an integral over anticommuting fields (ghosts) (Detour: in the abelian case the ghosts disappear)
4. BRST symmetry — the symmetry that remains after fixing, and how physical states are chosen
5. What happens when there are copies — comparing four ways of counting (Figure 1, Table 1; Detour: counting intersections with signs)
6. The Gribov region and the fundamental modular region — convex and bounded in every direction

7. The horizon in a box — the larger the box, the closer the horizon is to the origin (Figure 2, Table 2)
8. The 0/0 on the lattice — counting with signs gives the Euler characteristic 0
9. Gauge fixing seen from the mass gap problem — its place in the axioms and in attempts at construction

---

01

Why fix the gauge — no propagator can be made

Conventions: Euclidean $\mathbb{R}^4$, $\hbar = c = 1$, and we write the Hermitian $A_\mu = A_\mu^a T^a$ and the coupling g separately (the style of 3-02; the connection of the geometry chapters is $\mathcal{A} = -igA$).

$$(D_\mu \omega)^a = \partial_\mu \omega^a + g f^{abc} A_\mu^b \, \omega^c, \qquad \delta A_\mu^a = \frac{1}{g} (D_\mu \lambda)^a, \qquad S = \frac{1}{4} \int F_{\mu\nu}^a F_{\mu\nu}^a \, d^4x \quad (1)$$

Perturbation theory (4-03) uses the inverse of the quadratic part of the action as the propagator. S does not change under gauge transformations, so the quadratic part is flat in the direction $A = \partial\lambda$ and has no inverse.

$$S_2 = \frac{1}{2} \int \frac{d^4k}{(2\pi)^4} \, \tilde{A}_\mu^a(-k) \, K_{\mu\nu}(k) \, \tilde{A}_\nu^a(k), \qquad K_{\mu\nu} = k^2 \delta_{\mu\nu} - k_\mu k_\nu, \qquad Kk = 0 \quad (2)$$

Example. If $k = (0, 0, 0, \kappa)$, then $K = \kappa^2 \, \mathrm{diag}(1, 1, 1, 0)$, and the fourth component does not appear in the action Computed . On Wilson's lattice, where the volume of the orbits (4-02 § 07) is finite, no fixing is needed (5-01 § 06) Known . It is needed when one expands around $A = 0$ while staying in the continuum.

Let us put into words why there is no inverse. The path integral $\int \mathcal{D}A \, e^{-S}$ (4-02) adds up all fields related by gauge transformations (one orbit). Along the orbit e^{-S} does not change, so the Gaussian integral in that direction does not decay and diverges. Speaking in metaphor, the bottom of the valley is a single trench, flat without end along the trench. Precisely, the kernel K of the quadratic form (2) has eigenvalue 0 in the direction of k. So we would like to take the integral along the orbit outside, choose one representative from each orbit and add up the rest. We first see how to take it out in a finite-dimensional model.

02

A finite-dimensional model — Faddeev–Popov with rotations of the plane

The weight $e^{-S(|x|)}$ on the plane does not change under the rotation R_θ, and the orbits are circles centred at the origin. Here the angle of rotation θ plays the role of "the direction of the orbit". The aim is to bring the integral over θ to the surface without changing the value of Z, and for that we put a rewritten 1 into Z. We use the rule for the one-variable δ function.

$$\int d\theta\, \delta\big(G(\theta)\big)\, \big|G'(\theta_0)\big| = 1 \qquad (G \text{ has a single point } \theta_0 \text{ where it equals 0 and no other})$$

If the condition $G(x) = 0$ meets each circle exactly once, $\theta \mapsto G(R_\theta x)$ vanishes at a single point only, so the following "insertion of 1" holds.

$$1 = \Delta_G(x) \int_0^{2\pi} d\theta\, \delta\big(G(R_\theta x)\big), \qquad \Delta_G(x) = \Big|\tfrac{\partial}{\partial\theta} G(R_\theta x)\Big|_{G(R_\theta x)=0} \quad (3)$$

Putting this into $Z = \int d^2x\, e^{-S}$ and changing variables $x \to R_\theta^{-1} x$, since d^2x, S and Δ_G do not change under rotation, the integral over θ leaves the volume of the group 2π.

$$Z = 2\pi \int d^2x\, \delta\big(G(x)\big)\, \Delta_G(x)\, e^{-S(|x|)} \quad (4)$$

This is the whole of the Faddeev–Popov procedure, and the weight Δ_G on the surface is the **Faddeev–Popov determinant**.

Example. Taking the condition to be the positive x_1 axis ($G = x_2$, $x_1 > 0$), $\Delta_G = r$, and (4) is $Z = 2\pi \int_0^\infty r\, e^{-S} dr$: the Jacobian of polar coordinates. The determinant measures the size of the orbit (the same role as the Weyl factor in 3-04 § 04).

03

In field theory — the determinant and ghosts

We repeat (3)(4) with the group taken to be all pointwise gauge transformations and the condition $G^a = \partial_\mu A^a_\mu - \omega^a$. Since $\delta G^a = \frac{1}{g}\partial_\mu (D_\mu \lambda)^a$, the content of the determinant is the following $\mathcal{M}$. The constant factor $1/g$ and the sign become an overall constant of the determinant, absorbed into the overall constant of Z (which cancels in ratios of expectation values), so we drop them.

$$\mathcal{M} = -\partial_\mu D_\mu, \qquad Z = \int \mathcal{D}A\, \det\mathcal{M}\, \delta(\partial_\mu A_\mu - \omega)\, e^{-S} \quad (5)$$

Averaging ω with the weight $e^{-\int \omega^2/2\alpha}$ (many references write the parameter as ξ, but this book uses ξ for the correlation length, so we write α; the same letter as in 4-

03), the δ function turns into a term of the action, and the quadratic form acquires an inverse.

$$S + \frac{1}{2\alpha}\int (\partial_\mu A^a_\mu)^2, \qquad D^\alpha_{\mu\nu}(k) = \frac{1}{k^2}\left(\delta_{\mu\nu} - (1-\alpha)\frac{k_\mu k_\nu}{k^2}\right) \quad (6)$$

$\alpha = 1$ is the Feynman gauge and $\alpha \to 0$ the Landau gauge (the inverse checked numerically Computed). To make $\det \mathcal{M}$ a term of the action as well, we use anticommuting numbers (Grassmann numbers: $\theta_i\theta_j = -\theta_j\theta_i$, $\theta_i^2 = 0$). The rules are $\int d\theta\, 1 = 0$ and $\int d\theta\, \theta = 1$. Since the expansion stops at $e^{-a\bar\theta\theta} = 1 + a\theta\bar\theta$,

$$\int d\bar\theta\, d\theta\, e^{-a\bar\theta\theta} = a, \qquad \int \prod_i d\bar\theta_i d\theta_i\, e^{-\bar\theta M\theta} = \det M \quad (7)$$

Known . The determinant comes out **with its sign** (the source of the 0/0 of § 08). In this way the determinant becomes an integral over anticommuting fields $c^a, \bar c^a$ (**ghosts**; a different meaning from the "ghost" in 4-10, which refers to the negative-weight states of Pauli–Villars).

$$\det \mathcal{M} = \int \mathcal{D}\bar c\, \mathcal{D}c\, e^{-\int \bar c^u (\mathcal{M}c)^u}, \qquad \bar c^a (\mathcal{M}c)^a \simeq \partial_\mu \bar c^a \partial_\mu c^a + g f^{abc} (\partial_\mu \bar c^a) A^b_\mu c^c \quad (8)$$

($\simeq$ is integration by parts.) A ghost propagator δ^{ab}/k^2 and a vertex proportional to $g f^{abc} p_\mu$ come out Known . Ghosts are not particles that come out, and appear only inside loops (they also enter the β function: 4-06 § 03). **Example.** A real Gaussian integral in n components gives $(\det A)^{-1/2}$ (2-07 § 02). In the Feynman gauge ($\alpha = 1$) with $g = 0$ the kernel is $k^2\delta_{\mu\nu}$, so at each momentum the 4 components of the gauge field give $(k^2)^{-4/2}$ and the ghosts give $(k^2)^{+1}$ by (7), together $(k^2)^{-2/2}$: only the two transverse polarisations remain ($4 - 2 = 2$).

Detour: in the abelian case the ghosts disappear

A subsection you may skip.

For $U(1)$, $f^{abc} = 0$, so $\mathcal{M} = -\partial^2$ does not depend on A, and $\det \mathcal{M}$ cancels in ratios as a constant. No ghosts are needed. In the non-abelian case $\mathcal{M}$ depends on A, so the determinant couples to the field and ghost loops join the diagrams. The abstract of Faddeev–Popov (1967) writes that Feynman and DeWitt showed that in gauge-invariant field theories the rules for closed loops must be changed Known . The change of rules is needed when this $\mathcal{M}$ depends on A.

04

BRST symmetry — the symmetry that remains after fixing

The fixed action still has an anticommuting symmetry s that mixes the gauge field and the ghosts (with an auxiliary field b^a added).

$$sA^a_\mu = (D_\mu c)^a, \quad sc^a = -\tfrac{g}{2} f^{abc} c^b c^c, \quad s\bar{c}^a = ib^a, \quad sb^a = 0; \qquad S + \int s\Psi, \quad \Psi = \bar{c}^a (\partial_\mu A^{c}_{\mu}$$

$s\Psi = ib\,\partial A + \frac{\alpha}{2} b^2 + \bar{c}\,\mathcal{M} c$, and integrating out b returns to (6)(8). From $sS = 0$ and nilpotency $s^2 = 0$ (the Jacobi identity), two things follow formally Known .

- **Independent of the choice of gauge.** Changing Ψ changes the action only by $s(\delta\Psi)$, and for O annihilated by s, $\langle O\, s(\delta\Psi)\rangle = \langle s(O\delta\Psi)\rangle = 0$. The abstract of Becchi–Rouet–Stora (1976) writes that they showed that, for semisimple groups, in linear gauges in which all fields are massive, the Slavnov identities (the s symmetry written as relations between correlations) hold to all orders of renormalised perturbation theory, and that in the SU(2) Higgs–Kibble model examined as an example they showed, in the perturbative sense, that the physical S operator is unitary and independent of the gauge parameters (the 1975 paper is on the abelian Higgs–Kibble model). It is not a conclusion about massless pure Yang–Mills.
- **Physical states.** s becomes an operator Q_B acting on states (the BRST charge; $Q_B^2 = 0$). In covariant gauges the state space has an indefinite metric (the inner product $\langle \psi, \psi \rangle$ can be negative or 0), and as it stands it has no probabilistic interpretation. So we take the physical states to be $\mathbf{Ker}\, Q_B / \mathbf{Im}\, Q_B$ (BRST cohomology). This is "the states annihilated by Q_B, identified modulo the image of Q_B". Since $Q_B^2 = 0$, the image $\mathbf{Im}\, Q_B$ is contained in the kernel $\mathbf{Ker}\, Q_B$, and the quotient can be formed. The aim is that states in the image have inner product 0 with physical states, so dropping them by the identification does not change measurable quantities. States of negative or zero norm (the time and longitudinal polarisations, the ghosts) are dropped in this way, and the inner product on the remaining space is expected to be positive. The S operator is the operator that connects the states before and after scattering, and unitarity means preserving the total probability (the same property as e^{-itH} in 2-05 § 02). The abstract of Kugo–Ojima (1979) writes that, within a locally covariant operator formalism (a framework that assumes the theory and its asymptotic fields exist), they showed the unitarity of the physical S matrix from the algebra of the BRS charge Physics . The existence of the theory itself is assumed, and the identification in the continuum theory has not been shown (§ 09).

"BRST" is said to be the initials of Becchi, Rouet, Stora and Tyutin. **Example.** Computing s^2 symbolically on the part with constant fields, $s^2 A^a$ and $s^2 c^a$ had 0 terms for both SU(2) and SU(3). Changing the coefficient $-\frac{1}{2}$ of sc to -1 leaves 2 terms and 25 terms in $s^2 A^a$ Computed .

05

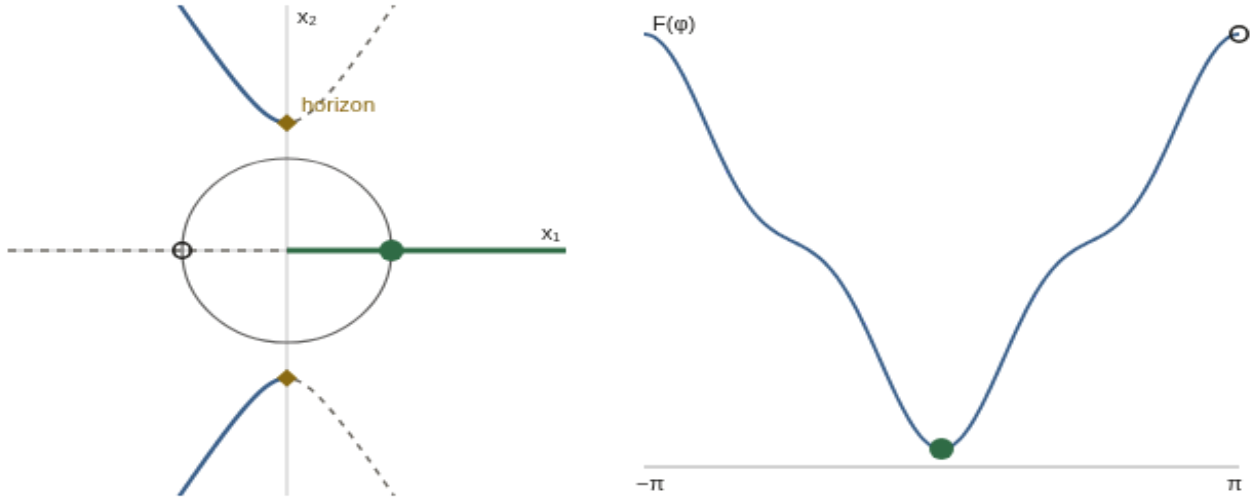
What happens when there are copies — four ways of counting

We build a case in which the "exactly once" of (3) breaks down. Following the Landau gauge (3-04 § 05), we take critical points of F on the circle as representatives, $M = \partial_\varphi^2 F$ as the determinant, and the set of local minima as the analogue of the Gribov region.

$F = -r\cos\varphi - c\,r^2\cos 3\varphi$: $\quad r < r_* = \frac{1}{3c}$ gives 2 copies (1 minimum, 1 maximum), $\quad r >$

For large fields pairs of a local minimum and a local maximum are born, and at those points $(0, \pm r_*)$, $M = 0$ (the analogue of the horizon). With the weight $e^{-r^2/2}$ we count r^2 (true expectation value 2) in four ways (for the difference between $|M|$ and M, see the detour of this section): (i) all copies with $|\det|$, (ii) only the Gribov region (local minima), (iii) only the global minimum of F (the fundamental modular region), (iv) with signs (the BRST way of counting).

Figure 1 Moving c and r changes the intersections (copies) of the condition surface with the orbit in the plane (left or top) and F(φ) on the circle (right or bottom). ● local minimum (Gribov region), ○ local maximum, green ● global minimum (fundamental modular region), ◇ horizon



c = 0.20, r = 1.20: copies 2 (min 1, max 1), sum of signs 0 r* = 1.667

⟨r²⟩: true 2.000, |det| 2.924, Gribov region 2.924, global min 2.000, signed 0/0

Computed The numbers in this figure were computed on this machine (the closed formulas of formula (10) and Table 1 evaluated on the spot; Table 1 agrees with a brute-force count of critical points to within 0.004).

Table 1 Four ways of counting ⟨r²⟩ (true value 2). Closed formula $\frac{4+8e^{-u}(1+u)}{2+4e^{-u}}$, $u = 1/(18c^2)$

c	r_*	(i) $\lvert\det\rvert$	(ii) Gribov region	(iii) Global minimum	(iv) With signs
0.08	4.167	2.005896	2.005896	2	0/0
0.20	1.667	2.924325	2.924325	2	0/0
0.50	0.667	2.273601	2.273601	2	0/0

When the number of copies varies from orbit to orbit, (i) is off, and (ii), which leaves copies inside the region, is not fixed either. (iii) is correct, but choosing the global minimum is a global comparison and cannot be written as a local condition. (iv) is 0/0, since local minima and local maxima come in equal numbers.

Example. The error of (i), $4ue^{-u}/(1+2e^{-u})$ (2×10^{-8} at $c=0.05$), is 0 at every order in powers of c, the same type as $e^{-1/x}$ in 4-07 § 05 Computed .

Detour: counting intersections with signs

A subsection you may skip. We look at the difference between $|M|$ and M in (10) with one variable on a circle.

Take the group to be the circle $\theta\in[0,2\pi)$, the condition a periodic function $f(\theta)=0$, and its zeros θ_i (with $f'\neq0$ there). From the rule for the δ function, $\int\delta(f)\,u\,d\theta=\sum_i u(\theta_i)/|f'(\theta_i)|$,

$$\int_0^{2\pi}d\theta\,\delta\big(f(\theta)\big)\,|f'(\theta)|=\#\{\theta_i\},\qquad\int_0^{2\pi}d\theta\,\delta\big(f(\theta)\big)\,f'(\theta)=\sum_i\operatorname{sign}f'(\theta_i)=0 \tag{10a}$$

The right-hand side is 0 because a periodic function crosses 0 upwards as many times as it crosses downwards. Example: for $f(\theta)=\sin(\theta-\theta_A)$ there are two intersections; the left-hand side is 2 and the right-hand side $1-1=0$. With absolute values the number of intersections remains; without them, the cancellation of signs. The relation to the obstruction to BRS-invariant gauge fixing done non-perturbatively on the lattice (Neuberger 1987; the abstract writes that it is formally present in the continuum as well) is treated in § 08.

06

The Gribov region and the fundamental modular region — convex and bounded in every direction

The abstract of Gribov (1978) writes that in the non-abelian case fixing $\partial_\mu A_\mu$ does not fix the gauge, and that restricting to the range where the FP determinant is positive seems to be enough Known . Reading the Landau gauge as the stationary points of the norm $\|A^g\|^2$,

$$\Omega=\{\partial_\mu A_\mu=0,\ \mathcal{M}(A)>0\}\ \text{(local minima)},\qquad\Lambda=\{A:\ \|A\|\le\|A^g\|\ \forall g\}\ \text{(global minim}$$

Ω is the **Gribov region**, its boundary the **horizon**, and Λ the **fundamental modular region** Known . $\bar\Omega$ is the closure of Ω (including its boundary). Ω is an open set given by $\mathcal{M}>0$, and Λ a closed set given by $\le$; the boundary of Λ has points touching the horizon (van Baal, § 07), so we write it with the closure.

- **Convex.** $\mathcal{M}(A)$ is linear in A, so "$\langle\omega, \mathcal{M}(A)\omega\rangle > 0$ for every ω" is an intersection of half-spaces. The abstract of Zwanziger (1982) also writes 'convex region Ω which is bounded in every direction'. Λ is also convex (the section "Complete gauge fixing" of van Baal 1997).
- **Bounded in every direction** (the abstract of Zwanziger 1982). The outline in finite dimensions: $\mathcal{M}_1$ in $\mathcal{M}(tA) = \mathcal{M}_0 + t\mathcal{M}_1$ is a symmetric matrix with trace 0 ($f^{aba} = 0$), so it has a negative eigenvalue, and for large t the operator stops being positive.
- **Every orbit passes inside the horizon** (the abstract of Dell'Antonio–Zwanziger 1991: the norm attains its minimum on each orbit, and there the orbit meets the inside of the horizon).
- **Copies remain even inside Ω.** Zwanziger (2004) argues that they do not affect expectation values Physics .

The solid line in Figure 1 is the analogue of Ω, and the positive x_1 axis the analogue of Λ. For the effect on the propagator, see 17-02 § 07.

07

The horizon in a box — eigenvalues in a constant background field

Taking a constant SU(2) field $A^a_\mu = a\,\delta^{a3}\delta_{\mu 1}$ in a periodic box of side L, $\mathcal{M}$ is diagonalised by plane waves $p = 2\pi n/L$, and splits into colour component 3 (neutral) and $\omega^1 \pm i\omega^2$ (charged).

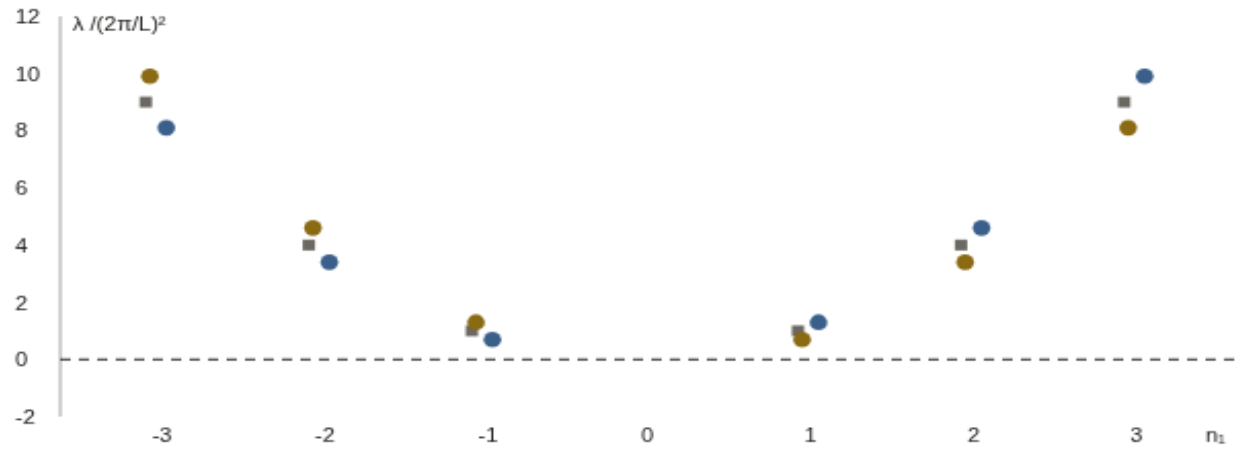
$$\lambda_{\text{neutral}} = \left(\tfrac{2\pi}{L}\right)^2 |n|^2, \qquad \lambda_\pm = \left(\tfrac{2\pi}{L}\right)^2 \left(|n|^2 \pm s\, n_1\right), \qquad s = \frac{gaL}{2\pi} \quad (12)$$

The smallest non-trivial eigenvalue (excluding $n = 0$) is $1 - s$ at $n = (\pm 1, 0, 0, 0)$, and the horizon is at

$$g\,a_{\text{h}} = 2\pi/L \;\longrightarrow\; 0 \quad (L \to \infty) \quad (13)$$

van Baal (the section "Gauge fields on the three-torus" of 1997) does the same calculation in the Coulomb gauge on the three-torus with $C = gaL$, and shows that the Gribov region is the cube $|C_k| \le 2\pi$ and that copies of $A = 0$ under twisted (antiperiodic) gauge transformations lie on the horizon Known . The fundamental modular region is $|C_k| \le \pi$ ($s \le \tfrac{1}{2}$) when antiperiodic transformations are also identified (pure Yang–Mills). When there are fields in the fundamental representation (quarks) and only periodic transformations are allowed, it is $|C_k| \le 2\pi$ ($s \le 1$), and its boundary touches the horizon Known .

Figure 2 Moving s (the size of the background) moves the eigenvalues of the FP operator (in units of $(2\pi/L)^2$, with transverse n equal to 0). Grey is neutral, blue and orange charged, red negative. The "boundary of Λ (s = 0.5)" in the readout is the value when antiperiodic transformations are also identified (s = 1 with periodic transformations only)




```
s = 0.30: smallest non-trivial eigenvalue 0.70 ×(2π/L)²  → inside Λ and Ω
```

Computed The numbers in this figure were computed on this machine (formula (12) evaluated on the spot). At $s = 0$, 0.5, 1 and 2 the smallest non-trivial eigenvalue is 1, 0.5, 0 and −1 (in order: $A = 0$, the boundary of the fundamental modular region [when antiperiodic transformations are also identified], the horizon, and the copy of $A = 0$ under a periodic transformation).

We also checked on the lattice. On an L^4 lattice, setting the links in direction 1 to $e^{i\theta\sigma 3/2}$ and diagonalising the second derivative of the Landau gauge functional $\sum(1 - \frac{1}{2}\mathrm{Re}\,\mathrm{tr}\, g_x U_{x,\mu} g^\dagger_{x+\mu})$ momentum by momentum, the charged eigenvalues agreed with $\frac{1}{2}[\cos\frac{\theta}{2} - \cos(p_1 \pm \frac{\theta}{2})] + \frac{1}{2}\sum_{\nu\neq 1}(1 - \cos p_\nu)$ (with lattice spacing 1, $\theta = ga$).

Table 2 The horizon of the lattice FP matrix (bisection)

L	4	8	16	32
$\theta_{\mathrm{h}} L/2\pi$	1.000000	1.000000	1.000000	1.000000

Computed python3, NumPy. The difference from the closed formula is at most 3×10^{-16}.

The larger the box, the closer to the origin the horizon is in long-wavelength directions, and the abstract of Gribov writes that this restriction cancels the infrared singularity and leads to a linear interaction at long distances Physics . It has not become a theorem about the four-dimensional measure.

08

The 0/0 on the lattice — it vanishes when counted with signs

The ghosts gave $\det \mathcal{M}$ without absolute value (formula (7)). So the integral of BRST-symmetric gauge fixing counts the copies on an orbit **with** $\mathrm{sign}\det \mathcal{M}$.

$$Z_{\mathrm{GF}}[U] = \int \prod_x dg_x\, \delta\big(G(U^g)\big)\, \det \mathcal{M}(U^g) = \sum_{\text{copies}} \mathrm{sign}\det \mathcal{M} = \chi(G)^N = 0 \quad (14)$$

We define the words used here. The **Euler characteristic** χ: the sum over the critical points (those whose matrix of second derivatives is non-degenerate) of a smooth function on a compact manifold without boundary (1-05), each counted with the sign of the determinant of that matrix ($+1$ if the number of negative eigenvalues is even, -1 if odd). This number does not depend on the choice of function (the Poincaré–Hopf theorem). On a circle a minimum counts $+1$ and a maximum -1, giving $\chi(S^1) = 0$; on the sphere $\chi(S^2) = 2$. For a compact connected Lie group G of positive dimension $\chi(G) = 0$, and for product spaces $\chi(G^N) = \chi(G)^N$ Known . **Morse theory** is the name of the theory that reads the shape of a manifold by counting critical points in this way.

The last two equalities read the sum of signs as the Euler characteristic of G^N (N is the number of lattice points) (Schaden 1999's interpretation via Morse theory, formula (2) of Mehta et al. 2014; on orbits whose critical points are non-degenerate), and $\chi(G) = 0$ Known (7-11 § 05). Expectation values of gauge-invariant quantities are 0/0. Mehta et al. call this "Neuberger's 0/0 problem" and cite Neuberger's papers of 1986 and 1987, and the abstract of the 1987 paper says that this obstruction met non-perturbatively is formally present in the continuum formulation as well Known .

Example. The "sum of signs 0" in Figure 1 is its smallest version ($\chi(S^1) = 0$). The same abstract also argues that at every order of the perturbative expansion the BRS invariance of the lattice does the job of local gauge invariance. This is because the cancellation of signs does not show up at any order of perturbation theory, the same type as $e^{-1/(18c^2)}$ in the example of § 05. The ways around it change the way of counting Known .

- **Do not fix** (Wilson's lattice; Elitzur's theorem: 18-02 § 09), or the maximal tree gauge, which has no Gribov problem (3-04 § 08).
- **Modified Landau gauge by stereographic projection**: map the link variables to flat coordinates by stereographic projection before building the functional (von Smekal et al. 2007; abstract: for compact $U(1)$ it solves the Gribov problem, and used on the maximal abelian subgroup it may avoid the 0/0 of $SU(N)$; the continuum limit of the Landau gauge is unchanged).
- **Orbifolding**: replace the group manifold used for gauge fixing by a generalisation of manifolds that allows singular points (an orbifold), and count with Morse theory on it (Mehta et al. 2014; abstract: the partition function becomes independent of the orbit).
- **Regularising with a mass**: count with an action in which mass terms (the Curci–Ferrari mass) are added to the gauge field and the ghosts, regularising the 0/0 (Ghiotti–von Smekal–Williams 2006, abstract). **Coset space**: BRST counting not on the whole group but on $SU(2)/U(1)$ (the sphere S^2, with $\chi = 2$) (Schaden).

For a comparison with the algebra of divergences, see 7-12.

09

Gauge fixing seen from the mass gap problem

What the problem statement asks for are correlations of gauge-invariant local fields (such as $\operatorname{tr} F_{\mu\nu}F_{\rho\sigma}$) and the gap of the reconstructed Hamiltonian (0-03). Gauge fixing is an indispensable tool in perturbation theory (§ 04). Elsewhere it has the following place Known .

- **Relation to the axioms.** The positivity needed for a probabilistic interpretation (reflection positivity, defined in 4-10) cannot be expected for fields fixed in a covariant gauge. That covariant gauges use a state space with an indefinite metric (§ 04) is a known framework, and Magnen–Rivasseau–Sénéor 1993 write that covariant gauges such as the Feynman gauge lack positivity and so cannot be used from the start of a construction (4-10 § 09). So the reconstruction is done from gauge-invariant quantities. The identification of BRST cohomology with that positive-definite space has not been shown in the continuum (the entry "physical states" of 18-02).
- **Attempts at construction** use it locally. On Bałaban, the exposition of the problem statement writes 'The choices vary locally in space-time, as well as on different length scales.' (Jaffe–Witten § 6.5, 7-05, 7-07). Chevyrev–Shen in two dimensions also use lattice gauge fixing (abstract, 7-09).

What Singer's theorem forbids is only continuous global slices, and the 0/0 too is a phenomenon of fixing a covariant gauge BRST-symmetrically over the whole lattice. The four-dimensional continuum limit of gauge-invariant correlations was built at the level of correlation functions by the construction published in 2026-11 (at the level of the committee's assessment; 6-00). The remaining questions are whether its measure exists and whether the spectrum has a gap (6-04).

What this chapter can say

	Content
What is said	The formal derivation of (3) to (9) Known . Table 1, Table 2, the numbers in the figures, the symbolic calculation of s^2 Computed . (10a) in the detour Paper (elementary)

	Content
Conditional	The Slavnov identities to all orders of perturbation theory in massive linear gauges; unitarity and gauge independence in the perturbative sense in the Higgs–Kibble model. The unitarity of Kugo–Ojima within a framework that assumes the theory exists Physics . The properties of the Gribov region checked at the level of abstracts. (14) on non-degenerate orbits. The effect on the infrared Physics
What is not said	The arrangement of copies in the four-dimensional continuum theory, the measure on the fundamental modular region, the identification of the BRST and OS spaces. Neuberger's argument only as far as the abstract. This site shows neither the mass gap nor the construction of the continuum theory

Sources and reproduction

Item	Kind	Source / tools
The procedure, Gribov copies	Known	L. D. Faddeev, V. N. Popov, Phys. Lett. B 25 (1967) 29; V. N. Gribov, Nucl. Phys. B 139 (1978) 1 (abstracts)
BRST, unitarity	Known	C. Becchi, A. Rouet, R. Stora, Commun. Math. Phys. 42 (1975) 127 (the abelian Higgs–Kibble model; bibliographic data only); Ann. Phys. 98 (1976) 287 (abstract); T. Kugo, I. Ojima, Prog. Theor. Phys. Suppl. 66 (1979) 1 (abstract). I. V. Tyutin, arXiv:0812.0580 (bibliographic data only)
The Gribov region	Known	D. Zwanziger, Nucl. Phys. B 209 (1982) 336; G. Dell'Antonio, D. Zwanziger, Commun. Math. Phys. 138 (1991) 291; D. Zwanziger, Phys. Rev. D 69 (2004) 016002 (abstracts); P. van Baal, hep-th/9711070 (text of the sections "Complete gauge fixing" and "Gauge fields on the three-torus")
The 0/0 on the lattice and ways around it	Known	H. Neuberger, Phys. Lett. B 175 (1986) 69 (bibliographic data only); Phys. Lett. B 183 (1987) 337 (abstract only); D. Mehta, N. S. Daleo, J. D. Hauenstein, C. Seaton, Phys. Rev. D 90 (2014) 054504, arXiv:1406.6678 (abstract, text § I, § II); M. Schaden, Phys. Rev. D 59 (1999) 014508 (bibliographic data only); L. von Smekal et al. arXiv:0710.2410 (abstract); M. Ghiotti et al. hep-th/0611058 (abstract)
Quotations in § 09	Known	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory" (Clay) § 6.5; I. Chevyrev, H. Shen, arXiv:2302.12160 (abstract); J. Magnen, V. Rivasseau, R. Sénéor, Commun. Math. Phys. 155 (1993) 325 (via the quotation in 4-10 § 09)
Tables, figures, the numbers in § 01, § 03, § 04	Computed	python3, NumPy (toy_fp, box_horizon, lattice_formula_check, brst_nilpotent, propagator_check). The figures evaluate formulas (10)(12) in JavaScript

Next chapter to read: 4-09 The axioms of quantum field theory.

The axioms of quantum field theory — "the theory exists" said in seven conditions, and infinitely many degrees of freedom force a choice of representation

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — The seven conditions of the Wightman axioms, which the problem statement names as the content of "existence", and what each of them excludes. That the state space and the fields can be rebuilt from the totality of correlation functions. That with infinitely many degrees of freedom the vacuum may fail to lie in the Fock space, and that this, as Haag's theorem, excludes the interaction picture.

Prerequisite chapters — 1-10 Hilbert spaces and self-adjoint operators, 2-06 The harmonic oscillator and creation and annihilation. We also use test functions and distributions (1-11) and light cones and Lorentz transformations (2-04), but define each in a line where it is used. $\hbar = c = 1$. We write the squared length of a spacetime point $x = (t, \underline{x})$ as $x^2 = -t^2 + |\underline{x}|^2$; the metric signature $(-,+,+,+)$ refers to this arrangement of signs.

Said first — This is an introduction to a known framework, and the proofs are given only in outline. The problem statement of 2000, § 6, writes that no non-trivial theory satisfying the axioms in four dimensions is known. For four-dimensional Yang–Mills, the construction published in 2026-11 built one (at the level of the committee's assessment; section 09). This chapter shows nothing about Yang–Mills. The numbers are for free fields.

Order of this chapter

1. Why axioms — conditions for judging whether a theory has been made
2. Fields have no values at points — average first, then make an operator (Table 1)
3. The Wightman axioms — seven conditions (Table 2)
4. Wightman functions and reconstruction — building the theory from correlations

5. Locality — commuting is not the same as uncorrelated (Figure 1, Table 3)
6. What follows from the axioms — PCT, spin and statistics, clustering. Detour: the vacuum reaches everywhere
7. Representations with infinitely many degrees of freedom — with infinitely many particles, they do not fit in the same space (Figure 2, Table 4)
8. Haag's theorem — an interacting field cannot be placed in the space of the free field
9. What happens in gauge theories — observables, positivity, the form of the axioms

---

01

Why axioms — conditions for judging whether a theory has been made

Physics textbooks calculate on the assumption that an interacting field theory exists (4-03). The mathematical question comes before that. Unless we first decide what must be made for us to say it "exists", we cannot judge whether it has been made. The axioms are the list of conditions for that judgement.

§ 3 of the problem statement defines a quantum field as 'an operator-valued generalized function on space-time obeying certain axioms', and continues 'Gårding and Wightman gave mathematically precise axioms for quantum field theories on R4 with a Minkowski signature, see [45]'. § 4 takes the content of existence to be properties at least as strong as the axioms of [45] (the book of Streater–Wightman) and [35] (the two papers of Osterwalder–Schrader) (0-03) Known . What must be made is the following quadruple.

$$\mathcal{T} = (\mathcal{H},\ U,\ \Omega,\ \phi) \qquad \text{state space, representation of the symmetry, vacuum, field} \ (1)$$

Example. In the quantum mechanics of one particle (2-05), $L^2(\mathbb{R})$, e^{-itH}, the ground state and the position operator are these four, and they are at hand from the start. What is new in field theory is that the symmetry is the Poincaré group (translations and Lorentz transformations; 2-04), that fields have no value at a point (section 02), and that with infinitely many degrees of freedom $\mathcal{H}$ is not determined uniquely (section 07). The axioms on the imaginary-time side are treated in 4-10.

02

Fields have no values at points — average first, then make an operator

Placing the free field of 2-06 § 08 over the whole of space, the vacuum expectation value of the square of the value at a point $\phi(\underline{x})$ becomes the sum of the zero-point fluctuations of all modes, and diverges. Averaged with a function f, $\phi(f) = \int \phi(\underline{x}) f(\underline{x})\, d^D x$ (D is the dimension of space) is finite Known .

$$\langle \Omega, \phi(f)^2\, \Omega \rangle = \int \frac{d^D p}{(2\pi)^D} \frac{|\hat{f}(\underline{p})|^2}{2\omega_p}, \qquad \omega_p = \sqrt{\underline{p}^2 + m^2} \quad (2)$$

$\hat{f}(\underline{p}) = \int f(\underline{x})\, e^{-i\underline{p}\cdot\underline{x}}\, d^D x$ is the Fourier transform: the amplitude of f split into waves of wave number $\underline{p}$ (1-11). An f of width ε removes the modes with $|\underline{p}| \gtrsim 1/\varepsilon$, and as $\varepsilon \to 0$ (2) diverges. Example. For a normalised Gaussian ($\hat{f} = e^{-\varepsilon^2 p^2/2}$), (2) grows logarithmically for $D = 1$, as $e^{\varepsilon^2 m^2/2} K_0(\varepsilon^2 m^2/2)/(4\pi)$ (K_0 is from section 05), and like $1/(8\pi^2\varepsilon^2)$ for $D = 3$.

Computed Table 1 $\langle \phi(f)^2 \rangle$ of the free field averaged with a Gaussian of width ε ($m = 1$, numerical integration of (2)). In brackets, $\frac{1}{2\pi} \ln \frac{1}{m\varepsilon}$ and $\frac{1}{8\pi^2\varepsilon^2}$

$m\varepsilon$	$D = 1$	$D = 3$
0.1	0.43302 (0.36647)	1.2383 (1.2665)
0.01	0.79736 (0.73294)	126.59 (126.65)
0.001	1.16379 (1.09940)	12665 (12665)

The difference for $D = 1$ approaches $(\ln 2 - \gamma/2)/(2\pi) = 0.06439$ (γ is Euler's constant) Computed . It is finite for every $\varepsilon > 0$, and only $\varepsilon = 0$ has no meaning. So we define a field as a linear map from test functions $f \in \mathcal{S}(\mathbb{R}^4)$ to operators $\phi(f)$. This is an operator-valued distribution. $\mathcal{S}(\mathbb{R}^4)$ is the space of functions that are smooth and that, together with their derivatives of every order, decay faster than any power of $|x|$ (1-11). We average over spacetime and not only over space because, for interacting fields, a field at fixed time is thought not necessarily to make sense. In theories that need renormalisation of the field strength (4-04), fixing the time and averaging only over space does not make the quantity corresponding to (2) finite Known .

03

The Wightman axioms — seven conditions, and what they exclude

The axioms are conditions imposed on each part of (1). The numbering differs from book to book, so in this book we call them by the names of the conditions rather than by numbers. Table 2 is for a single real scalar field Known .

Table 2 The Wightman axioms (a neutral scalar field; a summary in the form of Streater–Wightman)

No.	Condition	Content	What it excludes
1	States and symmetry	A strongly continuous unitary representation $U(a, \Lambda)$ of the Poincaré group on a separable Hilbert space $\mathcal{H}$	Indefinite inner products (section 09)
2	Spectral condition	The joint spectrum of $P^\mu = (H, \boldsymbol{P})$ lies in V_+ (formula (3))	Negative energy
3	Uniqueness of the vacuum	A unit vector Ω invariant under U, unique up to a phase	Mixtures of vacua
4	Field axiom	For each $f \in \mathcal{S}(\mathbb{R}^4)$, $\phi(f)$ is defined on a dense $\mathcal{D} \ni \Omega$ with $\phi(f)\mathcal{D} \subset \mathcal{D}$; $\langle \psi, \phi(f)\chi \rangle$ is a tempered distribution; $\phi(f)^* \supset \phi(\bar{f})$	Values at points (section 02)
5	Covariance	$U(a,\Lambda)\phi(f)U(a,\Lambda)^{-1} = \phi(f_{a,\Lambda})$, $f_{a,\Lambda}(x) = f(\Lambda^{-1}(x-a))$	Special coordinate systems
6	Local commutativity	If the supports of f, g are spacelike separated, $[\phi(f), \phi(g)] = 0$	Influences faster than light (section 05)
7	Cyclicity	Linear combinations of $\phi(f_1)\cdots\phi(f_n)\Omega$ are dense in $\mathcal{H}$	States unrelated to the field

We fix the words of the table one line at a time. **Dense**: every element of $\mathcal{H}$ can be approached arbitrarily closely by elements of the set. **Separable**: the linear combinations of countably many vectors are dense. $U(a, \Lambda)$ is the unitary operator assigned to the pair of a spacetime translation a and a Lorentz transformation Λ (a linear transformation preserving x^2; 2-04 § 02), and **strongly continuous** means that as (a, Λ) approaches the identity, $U(a, \Lambda)\psi \to \psi$ for every ψ. A **tempered distribution** is a continuous linear functional on $\mathcal{S}(\mathbb{R}^4)$. Two points x, y are **spacelike separated** if $(x-y)^2 > 0$, that is, if they cannot be connected by a light signal, and the surface $x^2 = 0$ is the **light cone** of the origin (2-04 § 03).

$$\sigma(P) \subset V_+ = \{\, p = (p^0, \boldsymbol{p}) \;:\; p^0 \geq |\boldsymbol{p}| \,\} \quad (3)$$

From (3) follows $H \geq 0$, and the mass operator $M = \sqrt{H^2 - \boldsymbol{P}^2}$ makes sense (4-11 § 05). The summary in § 3 of the problem statement names the states-and-symmetry axiom, the spectral condition ('positive energy, 0 ≤ H'), the uniqueness of the vacuum ('unique up to a phase'), covariance and local commutativity Known . For spin-1/2 fields one uses the double cover of the Poincaré group (a group in which two elements correspond to each Lorentz transformation; the four-dimensional version of $SU(2) \to SO(3)$ in 1-02), and local commutativity becomes anticommutation between fermions, $\phi(f)\phi(g) + \phi(g)\phi(f) = 0$. Example. The free field of mass m satisfies all seven, and its spectrum is the union of $\{0\}$, the hyperboloid $p^0 = \sqrt{\boldsymbol{p}^2 + m^2}$ and $p^0 \geq \sqrt{\boldsymbol{p}^2 + 4m^2}$ (Figure 4 of 2-06 § 08).

04

Wightman functions and reconstruction — building the theory from correlations

The vacuum expectation values of products of fields are called Wightman functions. By the field axiom (Table 2) they are tempered distributions.

$$W_n(x_1,\dots,x_n) = \langle \Omega,\ \phi(x_1)\cdots\phi(x_n)\,\Omega\rangle \quad (4)$$

The seven conditions translate into properties of W_n: Poincaré invariance, the spectral condition, locality (spacelike-separated neighbours may be swapped) and positivity. Positivity says that the length of $\Psi = \sum_n \phi(f_n)\Omega$ ($\phi(f_n) = \int f_n\,\phi(x_1)\cdots\phi(x_n)\,dx$) is not negative, and from $(\phi(x_1)\cdots\phi(x_n))^* = \phi(x_n)\cdots\phi(x_1)$

$$\|\Psi\|^2 = \sum_{n,k} \int f_n(x_1,\dots,x_n)\ W_{n+k}(x_n,\dots,x_1,y_1,\dots,y_k)\ f_k(y_1,\dots,y_k)\,dx\,dy\ \ge 0 \quad (5)$$

The reconstruction theorem Known . If $\{W_n\}$ satisfying these exist, a theory (1) with them as its Wightman functions can be made, unique up to unitary equivalence. The abstract of Wightman 1956 writes 'The problem of determining a theory of a neutral scalar field given its vacuum expectation values is posed and solved.' The construction is of the same type as in 4-10 § 05: put (5) as an inner product on finite sequences of test functions $(f_0, f_1, \dots)$, divide out the vectors of length 0 and complete, and the result is $\mathcal{H}$. $\Omega = (1,0,0,\dots)$, and $\phi(g)$ is the operation that replaces each component by the function with g attached in front ($(g\otimes f_1)(x_0,x_1) = g(x_0)f_1(x_1)$).

$$\phi(g):\ (f_0,f_1,f_2,\dots)\ \longmapsto\ (0,\ g\,f_0,\ g\otimes f_1,\ g\otimes f_2,\dots)$$

Analyticity follows from the spectral condition. Put $\psi = \phi(f)\Omega$ with a spacetime test function f, and let ρ be the spectral measure of H for ψ (1-10 § 05). Since Ω does not change under time evolution, the left-hand side is the two-point Wightman function with one of the fields shifted by time t and averaged.

$$\langle \psi,\ e^{-itH}\,\psi\rangle = \int_{[0,\infty)} e^{-iEt}\,d\rho(E) \ \xrightarrow{\ t=-i\tau\ }\ \int_{[0,\infty)} e^{-E\tau}\,d\rho(E) \quad (6)$$

Since $E \ge 0$, it extends analytically to $\operatorname{Im} t < 0$, and at $t = -i\tau$ oscillation turns into decay. This is the imaginary-time correlation, and the abstract writes 'The vacuum expectation values are shown to be boundary values of analytic functions.' This direction is the rigorous form of the Euclideanisation of 4-02, and the reverse direction is 4-10 Known . Example. For the free field, W_n is the sum of products of W_2 over all pairings, and reconstruction yields the Fock space of 2-06.

05

Locality — commuting is not the same as having no correlation

Local commutativity is the requirement that fields at two points that cannot be connected by light do not disturb each other. The time order of such two points is reversed for some observers (2-04 § 03), so if the order of the operators mattered there would be a contradiction. Local commutativity does not say that the correlation is 0. We compare with the free field in one space dimension.

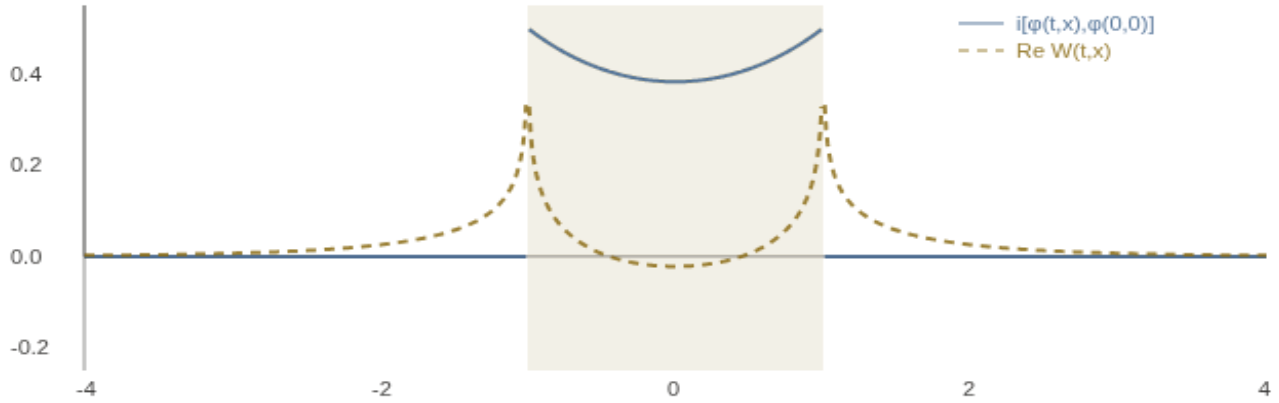
$$W(t,x) = \langle \Omega, \phi(t,x)\phi(0,0)\Omega \rangle = \int_{-\infty}^{\infty} \frac{dp}{2\pi \cdot 2\omega_p} e^{-i(\omega_p t - px)} \quad (7)$$

$$i\,[\phi(t,x),\phi(0,0)] = i\big(W(t,x) - W(-t,-x)\big) = \tfrac{1}{2}\,\mathrm{sgn}(t)\,\theta(t^2 - x^2)\,J_0\big(m\sqrt{t^2-x^2}\big) \quad (8)$$

$$W(0,x) = \frac{1}{2\pi} K_0(m|x|), \qquad K_0(z) \approx \sqrt{\frac{\pi}{2z}}\, e^{-z} \quad (z \to \infty) \quad (9)$$

θ is the step function, and $J_0(z) = \frac{1}{\pi}\int_0^\pi \cos(z\sin\vartheta)\,d\vartheta$ and $K_0(z) = \int_0^\infty e^{-z\cosh u}\,du$ are Bessel functions. The θ in (8) is local commutativity: outside the light cone the commutator is 0. But the correlation (9) is not 0 even outside, and decreases with distance as e^{-mr}. This decay is the spatial face of the mass gap (4-11 § 06). Passing to Euclidean space, the spatial direction and the imaginary-time direction are related by rotation, so it decays at the same rate as the imaginary-time decay of (6) (4-11 § 02). In four dimensions $W(0,\underline{x}) = mK_1(mr)/(4\pi^2 r)$ Computed .

Figure 1 Changing the time t and the mass m changes the shape of the commutator $i[\phi(t,x),\phi(0,0)]$ (solid line) and the correlation $\mathrm{Re}\,W(t,x)$ (dashed line) of the free field in 1+1 dimensions. The shading is the inside of the light cone. "Play" advances t



t = 1.00, m = 1.00: x = 1.50 (outside): commutator 0, Re W = 0.0567 / at x = 0: commutator 0.3826, Re W = −0.0221

Computed The numbers in this figure were computed on this machine ((8)(9) and the series and integral for Y_0 evaluated in JavaScript; agreeing with SciPy to within 10^{-13}).

Computed Table 3 The cross-section at $m = 1$, $t = 1$. Inside, $\mathrm{Re}\, W$ is $-\frac{1}{4} Y_0(m\sqrt{t^2 - x^2})$ (a Bessel function of the second kind). Agrees with numerical integration of (7) to 5 digits

x	Position	$i[\phi(t,x), \phi(0,0)]$	$\mathrm{Re}\, W(t,x)$
0	Inside	0.38260	−0.02206
0.9	Inside	0.47653	0.13611
1.1	Outside	0	0.15875
2	Outside	0	0.02529
4	Outside	0	0.00205

On the light cone $\mathrm{Re}\, W$ diverges logarithmically. The divergence at $t = x = 0$ is the same as the logarithm for $D = 1$ in section 02, a sign that fields have no value at a point.

06

What follows from the axioms — PCT, spin and statistics, clustering

From the seven conditions alone, properties known from experience come out as theorems. The title of [45], 'PCT, Spin and Statistics and all That', refers to two of them. The PCT theorem says that the combination of spatial reflection, time reversal and the exchange of particles and antiparticles is a symmetry of every theory; the spin–statistics theorem says that, at spacelike-separated points, fields of integer spin commute and fields of half-integer spin anticommute Known . The third is clustering: under the uniqueness of the vacuum, when groups of fields are moved apart in a spatial direction, correlations split into products.

$$W_{n+k}(x_1, \dots, x_n,\ y_1 + \lambda a, \dots, y_k + \lambda a) \xrightarrow[\lambda \to \infty]{} W_n(x_1, \dots, x_n)\, W_k(y_1, \dots, y_k) \quad (10)$$

With a mass gap the convergence is exponentially fast. § 5 of the problem statement writes 'An important consequence of the existence of a mass gap is clustering' (4-11 § 06) Known . Example. For the free field, the difference in (10) is a sum of terms e^{-mr} from (9).

Detour: the vacuum reaches everywhere

A subsection you may skip.

The Reeh–Schlieder theorem says that the density in cyclicity (Table 2) holds even with only test functions supported in an arbitrarily small open set of spacetime (attributed to Reeh–Schlieder 1961) Known . Operating only on the fields inside a room, one can make distant states as closely as one likes. Strictly it is an approximation, and the operators used are in general not unitary. This can happen because, as the dashed line in Figure 1 shows, the vacuum has

correlations reaching outside the light cone. On the other hand, the commutator (solid line) is 0 outside, so signals cannot be sent faster than light.

07

Representations with infinitely many degrees of freedom — with infinitely many particles, they do not fit in the same space

With finitely many oscillators, the irreducible representation of the commutation relations is unique (Stone–von Neumann; 2-06 § 09). With infinitely many it is not. We look at two free fields with different masses. For a single oscillator, the number n of quanta in the ground state of angular frequency ω', counted with the annihilation operator a_ω of ω, and the overlap between the ground states are

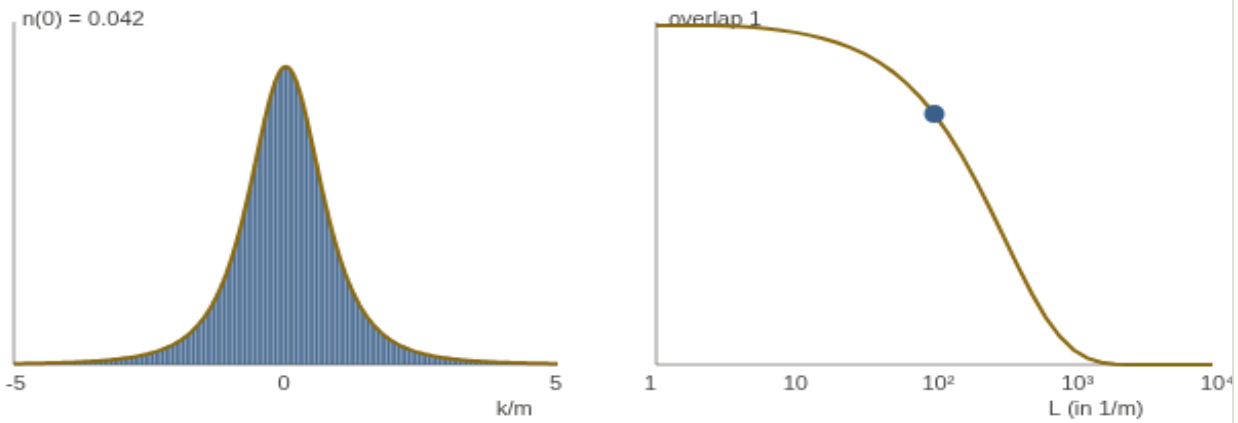
$$n = \langle 0\omega' | a_\omega^\dagger a_\omega | 0\omega' \rangle = \frac{(\omega - \omega')^2}{4\omega\omega'}, \qquad \langle 0\omega | 0\omega' \rangle = \left(\frac{2\sqrt{\omega\omega'}}{\omega + \omega'} \right)^{1/2} = (1+n)^{-1/4} \tag{11}$$

Known Computed . For a field on a ring of length L (2-06 (14)), each mode $k = 2\pi j / L$ has $\omega_k = \sqrt{k^2 + m^2}$ and $\omega'_k = \sqrt{k^2 + m'^2}$, and the vacuum is a product over modes, so

$$N(L) = \sum_k n_k, \qquad \langle \Omega_m, \Omega_{m'} \rangle = \prod_k (1+n_k)^{-1/4} \tag{12}$$

n_k is the n of (11) taken for mode k, and $N(L)$ is the expected number of quanta of mass m contained in the vacuum of mass m'. Since $n_k \approx (m'^2 - m^2)^2 / (16 k^4)$ decreases, for finite L the sum is finite and the overlap is positive. Enlarging L, both the number of modes and $N(L)$ are proportional to L, and the overlap goes to 0 as e^{-sL}.

Figure 2 Changing L and m'/m changes, on the left, n_k for each mode (vertical bars are the modes of the ring, the curve is $n(k)$), and on the right, the relation between the vacuum overlap and L (logarithmic scale). "Play" enlarges L



L = 100/m, m'/m = 1.50: number of quanta N = 1.269 (per 1/m: 0.0127), overlap 0.731

Computed The numbers in this figure were computed on this machine ((12) summed over $|k| \le 60m$, with the rest filled in by the k^{-4} tail; agrees with Table 4 to 6 digits).

Computed Table 4 $N(L)$ and the overlap on a ring of length L (in units of $1/m$) (formula (12))

L	N $(m' = 1.5m)$	Overlap	N $(m' = 2m)$	Overlap
10	0.12695	0.96915	0.43539	0.90040
100	1.2692	0.73104	4.3534	0.35025
1000	12.692	0.043592	43.534	2.778e−5
10000	126.92	2.478e−14	435.34	2.740e−46

Enlarging L, the overlap goes to 0. Precisely, in infinite volume there is no vector in the Fock space of m corresponding to the vacuum of m', and the two give unitarily inequivalent representations of the same commutation relations of the time-zero fields Known . From (12) one can read that the overlap is positive if and only if $\sum_k n_k < \infty$. A criterion of the same form is known for general Bogoliubov transformations (transformations that replace a_k by a linear combination of a_k and $a^\dagger_{-k}$; the relation between a_ω and $a_{\omega'}$ in (11) is an example) (attributed to Shale 1962). The infinite-volume density $\int \frac{d^D k}{(2\pi)^D} n_k$ $(m' = 1.5m)$ is finite: $0.012692,\ 0.0049736,\ 0.0030929$ (in units of m^D) for space dimension $D = 1, 2, 3$. Even for $D = 3$, which corresponds to real spacetime (four-dimensional spacetime), the density is finite, and the breakdown comes from the volume alone. For $D = 4$, where space is four-dimensional (spacetime five-dimensional), the density diverges logarithmically, and the overlap is 0 even in a finite box Computed . A figure for the three-dimensional lattice is in 7-11 § 07.

08

Haag's theorem — an interacting field cannot be placed in the space of the free field

Perturbation theory in physics takes the Fock space of the free field as its stage, splits $H = H_0 + V$ (H_0 free, V the interaction), and uses the picture of passing from the free vacuum to the interacting vacuum (the interaction picture). In section 07, even a mild change that only alters the mass gave a new vacuum that did not lie in the original Fock space.

Haag's theorem says this in general form Known . If the time-zero fields and conjugate momenta are unitarily related to those of the free field, and both theories have a unique vacuum invariant under spatial translations and rotations, then the first four Wightman functions agree with those of the free field, and a theory whose two-point function is the same as the free field's is the free field of the same mass (Haag 1955, Hall–Wightman 1957; the consolidated form is in Streater–Wightman, and a reading in Earman–Fraser 2006). The central lemma is elementary: in the Fock space of mass $m >$

0, the only vectors invariant under all spatial translations are constant multiples of the vacuum.

$$e^{-ia\cdot(p_1+\cdots+p_n)}\,\psi_n(p_1,\ldots,p_n)=\psi_n(p_1,\ldots,p_n)\ \ (\forall a)\ \Longrightarrow\ \psi_n=0\quad (n\geq 1)\ \text{(13)}$$

The n-particle component ψ_n vanishes off the surface where the total momentum is 0, and that surface has volume 0, so as an element of L^2 it is 0. Example. The vacuum of m' in (12) is invariant under translations and differs from the free vacuum, so it has no place in the Fock space.

In practice there are two ways of handling this. Physics calculates with a finite box and an ultraviolet cutoff, and takes as answers only the quantities with the cutoff removed by renormalisation (4-04). Constructive field theory builds the Hamiltonian in a space with a cutoff and obtains a new representation in the limit (7-03). Particles are also made from the spectrum. § 6 of the problem statement writes 'If m is an isolated eigenvalue of M, then one infers from the Wightman axioms and Haag–Ruelle scattering theory that asymptotic scattering states of an arbitrary number of particles exist' Known .

09

What happens in gauge theories — observables, positivity, the form of the axioms

It is known that the gauge field A_μ itself does not become a field satisfying the Wightman axioms (attributed to Strocchi 1967, 1970) Known . There are three reasons. Observables: A_μ changes under gauge transformations (3-01). § 4 of the problem statement asks that local operators correspond to 'the gauge-invariant local polynomials in the curvature F and its covariant derivatives, such as Tr Fij Fkl (x)', and in the footnote on local fields (footnote 1 of Jaffe–Witten) warns that, because of renormalisation, there is no natural one-to-one correspondence between classical polynomials and quantum operators Known .

Positivity: treating A_μ as a field in a Lorentz-covariant gauge needs a space with an indefinite inner product, and the states-and-symmetry axiom is broken. In the Gupta–Bleuler method (for the electromagnetic field), one takes from the large space the subspace of states satisfying a condition, and recovers positivity by dividing out the states of length 0 in it Known . For the non-abelian BRST method (4-08 § 04), a construction of the same form (take a subspace and divide; BRST cohomology) is known at the level of perturbation theory (attributed to Kugo–Ojima 1979). At the non-perturbative level this construction is not established, because of Gribov copies (4-08 § 05) Physics . The form of observables: Wilson loops (3-02) are quantities on curves and do not fit into the field axiom, which averages fields at points (4-10 § 09). § 3 of the

problem statement also mentions the Haag–Kastler framework, which places an algebra of operators on each region, and writes 'One hopes that the continued mathematical exploration of quantum field theory will lead to refinements of the axiom sets that have been in use up to now'.

Where things stand. § 6 of the problem statement writes 'at present we do not know any non-trivial relativistic field theory that satisfies the Wightman (or any other reasonable) axioms in four dimensions' (an assessment from 2000). In two and three dimensions, interacting scalar fields satisfying the axioms (ϕ^4_2, ϕ^4_3 and others) have been constructed (7-03) Known . Three-dimensional Yang–Mills has not been constructed (7-06). For four-dimensional Yang–Mills, the construction published in 2026-11 built, as the scaling limit of Wilson's lattice gauge theory, a theory satisfying the construction's own version of the Wightman axioms (the committee's assessment is that it is proved in Lean; 6-00, 12-08). The local fields are renormalised insertions corresponding to gauge-invariant polynomials in the field strength (8-04), and Wilson loops are not treated (6-04). The road that makes fields from the Euclidean side is the next chapter.

What this chapter can say

Content	What is said	What is not said
The axioms (section 03) and their correspondence with the problem statement	Known . § 3 to § 6 of the problem statement checked against the text	Table 2 is a summary of the standard form (the text of Streater–Wightman not obtained)
Reconstruction, analyticity (section 04)	Known (the abstract of Wightman 1956) and the outline of the construction	The details of the proofs
Tables 1 and 3, Figure 1, formulas (2)(7) to (9)	Computed : agreement in double precision for the free field	The form of correlations of interacting fields
Table 4, Figure 2, formulas (11)(12)	Computed : the overlap between vacua of free fields goes to 0 with the volume	The proof of Shale's criterion (text not read)
PCT, spin and statistics, Reeh–Schlieder, Haag's theorem	Presentation of the Known claims, and the lemma (13)	The proofs of the theorems (original papers bibliographic data only)

Content	What is said	What is not said
The axioms for gauge theories (section 09)	Organising the requirements of the problem statement and the known obstructions	Recovery of positivity by BRST at the non-perturbative level is not established. The form of axioms including Wilson loops is not settled. A theory satisfying the axioms for four-dimensional Yang–Mills was built by the construction published in 2026-11 (at the level of the committee's assessment; 6-00). This chapter shows nothing

Sources and reproduction

Item	Kind	Source / tools
Quotations from the problem statement	Known	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory" (Clay) § 3 to § 6, footnote 1. Checked word for word against the text
Axioms, PCT, spin and statistics, fields at fixed time (section 02)	Known	R. Streater, A. Wightman, *PCT, Spin and Statistics and all That*, Benjamin, 1964. Bibliographic data only
Reconstruction, analyticity	Known	A. S. Wightman, Phys. Rev. 101 (1956) 860. Checked in the abstract
Haag's theorem	Known	R. Haag, Mat.-fys. Medd. Dan. Vid. Selsk. 29 no. 12 (1955): bibliographic data only (secondary sources). D. Hall, A. S. Wightman, 'A theorem on invariant analytic functions with applications to relativistic quantum field theory', same journal 31 no. 5 (1957): bibliographic data only (consistent across secondary sources). J. Earman, D. Fraser, Erkenntnis 64 (2006) 305–344: bibliographic data only
The criterion for Bogoliubov transformations	Known	D. Shale, Trans. Amer. Math. Soc. 103 (1962) 149–167. Bibliographic data only
Detour	Known	H. Reeh, S. Schlieder, Nuovo Cim. 22 (1961) 1051–1068. Bibliographic data only
Haag–Kastler, Haag–Ruelle	Known	R. Haag, D. Kastler, J. Math. Phys. 5 (1964) 848–861: bibliographic data only (bibliographic data checked with Crossref). D. Ruelle, Helv. Phys. Acta 35 (1962) 147: bibliographic data only, not checked. R. Haag, *Local Quantum Physics*, Springer, 1992, and Phys. Rev. 112 (1958) 669. Bibliographic data only
Gauge fields and positivity (section 09)	Known, Physics	F. Strocchi, Phys. Rev. 162 (1967) 1429–1438, Phys. Rev. D 2 (1970) 2334–2340. T. Kugo, I. Ojima, Prog. Theor. Phys. Suppl. 66 (1979) 1–130. Bibliographic data only (bibliographic data checked with Crossref)
Tables 1, 3 and 4, Figures 1 and 2, (8)(9)(11)(12) and the formula with K_1	Computed	python3, SciPy (`smear_variance.py`, `wightman_1p1.py`, `free_3p1.py`, `bogoliubov.py`). The figures evaluate the same formulas in JavaScript

Next chapter to read: 4-10 The Osterwalder–Schrader axioms and reconstruction.

The Osterwalder–Schrader axioms and reconstruction — reflection positivity turns probability back into quantum mechanics

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — The five conditions imposed on Schwinger functions (regularity, Euclidean covariance, reflection positivity, permutation symmetry and clustering; the Osterwalder–Schrader axioms E0 to E4), and the procedure that builds from them a Hilbert space, $H \geq 0$ and a vacuum. That the key, reflection positivity, is for one-dimensional Gaussian processes the same as "the correlation is a positive superposition of decaying exponentials". Regularisations that break reflection positivity, and the lattice, which keeps it.

Prerequisite chapters — 4-02 The path integral and Euclideanisation, 4-09 The axioms of quantum field theory. Tools from 1-08, 1-10, 1-11, 2-06 and 2-07. $\hbar = c = 1$.

Said first — This is an introduction to known theorems, and the proofs are given only in one-dimensional outline. For four-dimensional Yang–Mills, the construction published in 2026-11 built a Hilbert space from the Schwinger functions of the limit by the reconstruction of this chapter (the committee's assessment is that it is proved in Lean; 6-00, 12-08). A probability measure satisfying the axioms has not been constructed (6-04). This chapter shows nothing. Section 06 is a calculation in a one-particle model.

Order of this chapter

1. Making the way back a theorem
2. Schwinger functions
3. Five conditions (Table 1)
4. Reflection positivity (Figure 1, Table 2)
5. Reconstruction
6. Running reconstruction on the smallest model (Figure 2, Table 3)

7. Detour: noisy data
8. The free field
9. Regularisations that break reflection positivity, and the lattice that keeps it
10. What gauge theories add

01

Making the way back a theorem — from imaginary time to Hilbert space

What can be made on the Euclidean side is a probability measure μ and its moments (4-02, One sign is different). What the problem statement asks for is a Hilbert space $\mathcal{H}$, $H \geq 0$ and a vacuum Ω (4-09), and § 4 writes 'Existence includes establishing axiomatic properties at least as strong as those cited in [45, 35].' [35] is the two papers of Osterwalder–Schrader (OS) Known .

The way there (from spacetime to imaginary time) is analytic continuation. On the way back, Schwinger said in the discussion at the 1958 CERN conference 'The question of to what extent you can go backwards, remains unanswered' (according to Jaffe 2018). Nelson built the way back under the Markov property (the property that, given the present, past and future are independent), and OS replaced it by reflection positivity (problem statement § 3) Known .

$$\mu \;\longrightarrow\; \{S_n\} \xrightarrow[\text{reconstruction}]{(E0)\sim(E4)} (\mathcal{H},\ H \geq 0,\ \Omega) \xrightarrow[\text{analytic continuation}]{(E0')} \{W_n\} \quad (1)$$

Example. For one particle, $\mathbb{E}[x(\tau)x(0)] = \langle x\psi_0, e^{-\tau(H-E_0)}x\psi_0\rangle$ is the way there (formula (8) of 2-06). Section 06 is the reverse.

02

Schwinger functions — energy is written into the correlations

We write points as $x = (\tau, \underline{x}) \in \mathbb{R}^d$ and call τ the Euclidean time. The n-point **Schwinger function** is a moment of μ; precisely, the object is $S_n(f)$ smoothed with test functions (1-11). It is related to the Wightman function formally by the substitution $x^0 \to -i\tau$, $S_n = W_n(x_1^E, \dots, x_n^E)$ (Summers § 3) Known .

$$S_n(x_1, \dots, x_n) = \int \phi(x_1)\cdots\phi(x_n)\, d\mu(\phi) \quad (2)$$

Example: the free field of mass m is the Gaussian measure with covariance $(-\Delta + m^2)^{-1}$ (Δ is the Laplacian $\sum_\mu \partial_\mu^2$, not the Δ of the mass gap) (problem statement § 6.2). Integrating over p_0 alone (formula (11) of 2-04) gives

$$S_2(x,y) = \int \frac{d^d p}{(2\pi)^d} \frac{e^{ip\cdot(x-y)}}{p^2+m^2}, \qquad \int \frac{dp_0}{2\pi} \frac{e^{ip_0\tau}}{p_0^2+\omega^2} = \frac{e^{-\omega|\tau|}}{2\omega} \quad (3)$$

where $\omega = \sqrt{\underline{p}^2 + m^2}$ is the energy of a particle of momentum $\underline{p}$. The energy is written into the rate at which the correlation decreases: this is the seed of reconstruction. For $d = 1$ the right-hand side is itself $S_2(\tau)$, which becomes the example of section 04. The higher S_n are sums of products of S_2 over pairings (1-08).

03

Five conditions — from regularity to clustering, and linear growth

Let $\Theta(\tau, \underline{x}) = (-\tau, \underline{x})$ be time reversal.

Known Table 1 The OS conditions (scalar field; the numbering is said to be that of OS 1975)

Condition	Content	On the spacetime side
Regularity (E0)	Tempered away from coinciding points	Operator-valued distributions
Euclidean covariance (E1)	Invariant under rotations and translations	Poincaré covariance
Reflection positivity (E2)	For f_n supported at positive times, $\sum_{n,m} S_{n+m}(\Theta f_n \otimes f_m) \geq 0$	Positive inner product, $H \geq 0$
Permutation symmetry (E3)	Invariant under permutations of the points	Separated fields commute
Clustering (E4)	$S_{n+m} \to S_n S_m$ when moved apart in a spatial direction	A unique vacuum
Linear-growth condition (E0′)	$\lvert S_n(f)\rvert \leq \sigma_n \lvert f\rvert_{C_n}$, $\sigma_n \leq A(n!)^B$ (the order of smoothness used in the estimate is proportional to n: the origin of the name "linear growth")	Continuation to W_n

Notation of the table. The **support** of a function f is (the closure of) the set of points where $f \neq 0$, and "supported at positive times" means $f = 0$ for $\tau \leq 0$. $\otimes$ is the function with separated variables multiplied, $(f \otimes g)(x,y) = f(x)g(y)$, and Θf is f with the time of each variable reversed. $\lvert f\rvert_k$ is the supremum of the derivatives of f up to order k multiplied by polynomial weights, measuring the smoothness and decay of rapidly decreasing functions (1-11).

That regularity excludes coinciding points agrees with the fact that S_2 of the free field diverges at $x = y$ (for $d \geq 2$). The proof in the 1973 paper had an error concerning the continuity of tensor products of distributions (Jaffe 2018), and the second paper of 1975 is said to have given a correct proof with the linear-growth condition added as an assumption (according to secondary sources; OS 1975 bibliographic data only). It has been shown that the S_n made from a Wightman theory satisfy the five conditions, but they do not necessarily satisfy the linear-growth condition. So it is not known whether the OS axioms including the linear-growth condition are equivalent to the Wightman axioms themselves (Summers § 3, footnote 15). What OS 1975 showed is equivalence with a modified form of Wightman theory (Jaffe 2018; the modification is said to add a growth condition) Known . Example. The free field satisfies all of them, and the linear-growth condition follows from S_{2k} being a sum over $(2k-1)!! \leq (2k)!$ pairings. The number of pairings $(2k-1)!!$ goes into σ_n, and each factor is bounded by a seminorm of fixed order.

04

Reflection positivity — measuring lengths with the mirror of time

Let F be a quantity determined by the fields at positive times only (for example $F = \sum_j c_j \phi(\tau_j, \underset{\sim}{x}_j)$, $\tau_j > 0$), and ΘF the same with time reversed. Reflection positivity has the following form.

$$\langle F, F \rangle := \mathbb{E}\big[\, \Theta F \; F \,\big] \;\geq\; 0 \qquad (4)$$

Speaking in metaphor, it is a condition that measures the length of a future quantity by its overlap with its mirror image. Precisely, it is the condition that the quadratic form of correlations between past and future is not negative, and unlike $\mathbb{E}[|F|^2] \geq 0$ it is not trivial. For a one-dimensional stationary Gaussian process (covariance $C(\tau - \tau')$), taking $F = \sum_j c_j \phi(\tau_j)$,

$$\sum_{i,j} c_i c_j C(\tau_i + \tau_j) \geq 0 \;\; (\tau_j > 0), \qquad \text{in particular} \;\; C(2\tau) \geq 0, \;\; C(\tau_1 + \tau_2)^2 \leq C(2\tau_1) C(2\tau_2$$

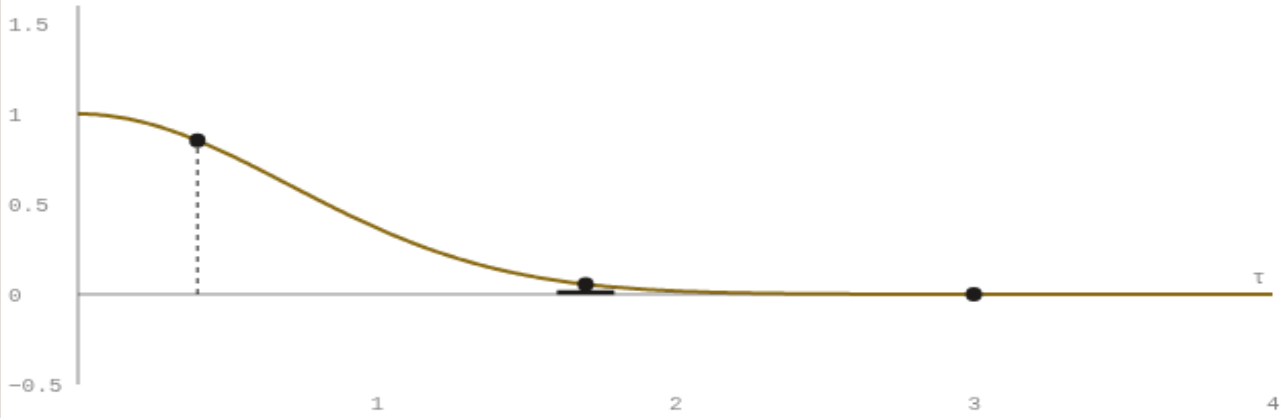
The right-hand side is the diagonal and determinant of the 2×2 matrix of two points, and requires the convexity of $\log C$. A sufficient condition: if $C(\tau) = \int_{[0,\infty)} e^{-E\tau}\, d\rho(E)$ ($\rho \geq 0$), the left-hand side becomes $\int |\sum_j c_j e^{-E\tau_j}|^2 d\rho \geq 0$.

Example: $e^{-\tau^2}$ is not reflection positive

$e^{-\tau^2}$ is a legitimate covariance (its Fourier transform is positive; **Bochner's theorem**: a continuous function C is positive definite, that is, $\sum_{i,j} c_i c_j C(\tau_i - \tau_j) \geq 0$ for any τ_i

and complex c_i, if and only if its Fourier transform is a positive measure), but $\log C$ is concave, and at $\tau 1 = 0.2$, $\tau 2 = 1.5$, from $C(0.4) = 0.852144$, $C(1.7) = 0.055576$, $C(3.0) = 1.234 \times 10^{-4}$ the determinant is -2.98×10^{-3} Computed . (Why it cannot be written in this form: for positive ρ, $C(\tau) \geq \rho([0, E])e^{-E\tau}$, so it cannot decrease faster than an exponential. That this is the cause of the violation becomes clear once necessity is shown in section 05.)

Figure 1 Choosing a covariance and moving τ_1, τ_2 moves the three points $C(2\tau_1), C(\tau_1 + \tau_2), C(2\tau_2)$ (circles) and the upper bound of (5) (horizontal bar). It is violated if an end circle is below 0, or if the absolute height of the middle circle exceeds the bar



e^−τ², τ₁=0.20, τ₂=1.50: C(2τ₁)=0.8521, C(τ₁+τ₂)=0.0556, C(2τ₂)=0.00012, smallest eigenvalue −0.00349 (violated)

Computed The numbers in this figure were computed on this machine (the matrix of (5) computed on the spot).

Computed Table 2 "Matrix" is the smallest eigenvalue of $[C(\tau_i + \tau_j)]$ ($\tau_i = 0.1i$, $i \leq 40$), and "two points" the minimum of the determinant of (5) for $\tau_1, \tau_2 \in (0, 2]$.

$C(\tau)$	Fourier transform (positive)	Matrix	Two points
$e^{-\lvert\tau\rvert}$	$\frac{2}{1+\omega^2}$	≈ 0	≈ 0
$e^{-\lvert\tau\rvert} + \frac{1}{2}e^{-3\lvert\tau\rvert}$	$\frac{2}{1+\omega^2} + \frac{3}{9+\omega^2}$	≈ 0	≈ 0
$e^{-\tau^2}$	$\sqrt{\pi}\, e^{-\omega^2/4}$	−0.462	−0.238
$(1 - \lvert\tau\rvert)_+$	$\left(\frac{\sin(\omega/2)}{\omega/2}\right)^2$	−0.366	−0.239
$e^{-\lvert\tau\rvert}\cos 3\tau$	$\frac{1}{1+(\omega-3)^2} + \frac{1}{1+(\omega+3)^2}$	−2.13	−0.376
$\frac{1}{2}e^{-\lvert\tau\rvert} - \frac{1}{6}e^{-3\lvert\tau\rvert}$	$\frac{8}{(1+\omega^2)(9+\omega^2)}$	−0.054	−0.011

All six are legitimate as covariances, and the bottom four violate reflection positivity ($(a)_+ = \max(a, 0)$). The damped oscillation is the form of a propagator with complex poles $\omega = \pm 3 \pm i$, and the Gribov–Zwanziger form of 17-02 also has complex poles. By (5), a field whose correlation in the time direction becomes negative cannot be read as

a field of a reflection-positive theory. In lattice calculations in the Landau gauge, the correlation of gluons in the time direction has been observed to become negative, and physics reads this as a sign that gluons are not observed particles Physics .

05

Reconstruction — inner product, contraction, Hamiltonian

Given (4), the rest proceeds by linear algebra Known . **(1) Inner product.** On the space $\mathcal{E}_+$ of all bounded functions of the fields at positive times, put $\langle F, G\rangle := \mathbb{E}[\Theta F\, G]$. By (4), $\langle F, F\rangle \geq 0$, but $\langle F, F\rangle = 0$ is possible even with $F \neq 0$ (positive semidefinite). Even so, Cauchy–Schwarz $|\langle F, G\rangle|^2 \leq \langle F, F\rangle\langle G, G\rangle$ holds (from the discriminant of the quadratic $\langle F + tG, F + tG\rangle \geq 0$ in t). From this the set $\mathcal{N}$ of elements of length 0 is a subspace, and we complete $\mathcal{E}_+/\mathcal{N}$, which identifies elements that differ only by something of length 0, to get $\mathcal{H}$ (for completion, 1-10). **(2) Shifting.** The shift U_s by $s \geq 0$ in the time direction takes quantities at positive times to quantities at positive times, and from the translation invariance of μ and $\Theta U_s \Theta = U_{-s}$,

$$\langle F, U_s G\rangle = \langle U_{s/2} F, U_{s/2} G\rangle = \langle U_s F, G\rangle \quad (6)$$

(3) Contraction. Repeating (6) and Cauchy–Schwarz n times gives

$$\|U_s F\|^2 = \langle F, U_{2s} F\rangle \leq \|F\|\,\|U_{2s} F\| \implies \|U_s F\| \leq \|F\|^{1-2^{-n}} \|U_{2^n s} F\|^{2^{-n}} \quad (7)$$

and since $\|U_u F\| \leq \sup|F|$, as $n \to \infty$ we get $\|U_s F\| \leq \|F\|$. Length 0 goes to length 0, and U_s defines $T(s)$ on $\mathcal{H}$. **(4) Hamiltonian.** $T(s)$ is symmetric with $0 \leq T(s) = T(s/2)^2 \leq 1$ and $T(s)T(u) = T(s+u)$, so if it is continuous, the spectral theorem (1-10) gives

$$T(s) = e^{-sH}, \qquad H = H^* \geq 0, \qquad \Omega := [1], \qquad H\Omega = 0 \quad (8)$$

This is the Hamiltonian.

Necessity in one dimension, and real time

In one dimension $C(\tau + 2\varepsilon) = \langle \psi_\varepsilon, e^{-\tau H} \psi_\varepsilon\rangle = \int e^{-E\tau}\, d\rho_\varepsilon(E)$ ($\psi_\varepsilon = [\phi(\varepsilon)]$, and ρ_ε is the spectral measure; ϕ is not bounded, but for Gaussians $\mathcal{E}_+$ may be extended to square-integrable functions), so the sufficient condition of section 04 is also necessary. Therefore a C that decreases faster than an exponential, like $e^{-\tau^2}$, cannot be reflection positive. The form $\int e^{-E\tau}\, d\rho(E)$ is called the **Laplace transform** of ρ, and on this side it corresponds to the characterisation of Widder 1934. The connection with reflection positivity is said to have been pointed out by Klein–Landau (Jaffe 2018). Since $H \geq 0$, e^{-zH} makes sense for $\operatorname{Re} z \geq 0$, and at $z = it$ it is time evolution e^{-itH}.

Boosts are obtained in the OS reconstruction from the analytic continuation of the Wightman functions. Fröhlich–Osterwalder–Seiler 1983 and Klein–Landau abstracted this into the form of extending the rotation angle to imaginary values (0-04 § 02) (Jaffe 2018 § 3.3.1). The linear-growth condition is used at the stage of obtaining W_n Known . Example. If $C = e^{-m|\tau|}/(2m)$ and $F = \phi(\tau_0)$, then $\|U_s F\| = e^{-ms}\|F\|$, and on this vector H acts as multiplication by m.

06

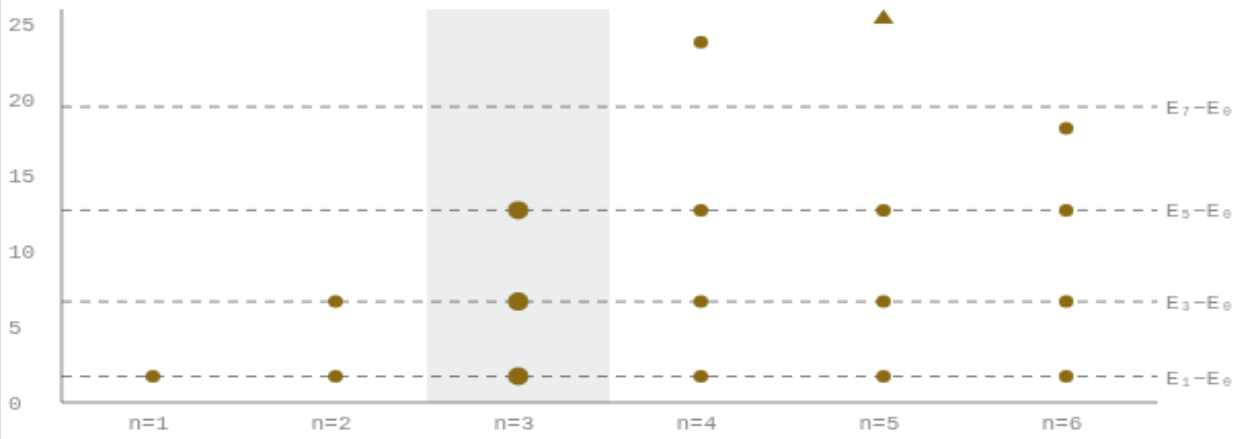
Running reconstruction on the smallest model — reading levels from correlations alone

With $H = \frac{1}{2}p^2 + x^4$ of Table 4 of 2-07 we make $C(\tau) = \mathbb{E}[x(\tau)x(0)]$ (diagonalising in a basis of 240 functions), then forget H and carry out section 05 with C alone. With $s = 0.25$ and $F_i = x((i+\frac{1}{2})s)$ $(i < n)$,

$$G_{ij} = \langle F_i, F_j\rangle = C\big((i+j+1)s\big), \qquad G^{(1)}_{ij} = \langle F_i, T(s) F_j\rangle = C\big((i+j+2)s\big) \ \text{(9)}$$

from the generalised eigenvalue problem $G^{(1)}v = \lambda G v$ (the same as the eigenvalues of $G^{-1/2}G^{(1)}G^{-1/2}$ if G is positive definite) we read $E = -\frac{1}{s}\log\lambda$. Since these are eigenvalues of $T(s)$ restricted to the subspace spanned by F_i, by the variational principle (1-10) they are **upper bounds** on the true levels Known . That G is positive definite (reflection positivity) is presupposed. x is odd, so what can be read are $E_1 - E_0$, $E_3 - E_0$, …, and the weights $|\langle \psi_0, x\psi_n\rangle|^2$ fall sharply, from 0.287 at $n = 1$ to 3.9×10^{-9} at $n = 7$ Computed .

Figure 2 Changing the number n of correlations and the noise ε changes the levels read (circles; one n per column). The dashed lines are the true levels, the triangles above 26



ε=0 (rounding only), n=3: smallest eigenvalue of G 3.6e−10. Levels read: 1.72566、6.66774、12.71344

Computed The numbers in this figure were computed on this machine (results of the same calculation as Table 3).

Computed Table 3 Levels read from the table of correlations alone (no noise), and the smallest eigenvalue of G.

n	E_1-E_0	E_3-E_0	E_5-E_0	E_7-E_0	min G
1	1.72807				1.9e-1
2	1.72566	6.67039			2.4e-5
3	1.72566	6.66774	12.71344		3.6e-10
4	1.72566	6.66774	12.71165	23.80937	2.5e-16
True	1.725658	6.667744	12.711350	19.552863	

At $n = 3$ three levels come out to 4 or 5 digits, all above the true values. At $n = 4$ the smallest eigenvalue of G reaches the size of double-precision rounding, and beyond that nothing can be distinguished.

Detour: noisy data — the way back is unstable

A subsection you may skip.

Correlations from Monte Carlo (5-07) have errors. We multiplied $C(ks)$ by normally distributed random errors of relative size ε (with a fixed seed) and ran the same procedure Computed . At $n = 4$ with $\varepsilon = 10^{-12}$, the third level is 12.70331, below the true 12.71135, and the upper-bound property is lost. At $\varepsilon = 10^{-8}$ the smallest eigenvalue of G becomes negative, -6.4×10^{-10}, and the data violate (4) (the matrix form on the left of (5)) ($E3 - E0$ is 6.66936). At $\varepsilon = 10^{-4}$ with $n = 6$, a spurious level of 1.52597, below the true gap, appears (Figure 2). Because the Laplace transform is being traced backwards, tiny errors move the upper levels a lot. The reconstruction theorem is a claim about S_n without errors, and reading levels from finite-precision data is a different, ill-conditioned problem. Glueball measurements (5-08) also use matrices of the form (9).

07

The free field — Fock space comes out

For Gaussian measures the reconstruction can be written out completely. With the one-dimensional $C(\tau) = e^{-m|\tau|}/(2m)$ (an oscillator of angular frequency m), using **Wick products**, the powers of ϕ adjusted by the variance $\sigma^2 = C(0)$ ($:\phi^2: = \phi^2 - \sigma^2$, $:\phi^3: = \phi^3 - 3\sigma^2\phi$, and in general $:\phi^k: = \sigma^k He_k(\phi/\sigma)$ with the Hermite polynomials $He_k(x) = (-1)^k e^{x^2/2}\frac{d^k}{dx^k}e^{-x^2/2}$),

$$\mathbb{E}\big[{:}\phi(\tau)^k{:}\,{:}\phi(0)^l{:}\big] = \delta_{kl}\,k!\,C(\tau)^k = \delta_{kl}\,\frac{k!}{(2m)^k}e^{-km|\tau|} \tag{10}$$

and the decay rate km is the level. For this process (the stationary Gaussian process with covariance $e^{-m|\tau|}/(2m)$ is called the **Ornstein–Uhlenbeck process**), $[{:}\phi^k{:}]$ is an eigenvector with eigenvalue km, and $H = m\,a^\dagger a$ Known . The levels $E_k - E_0 = km$ of 2-06 (2-06) come out from the correlations alone. Example. At $m = 1$, $k = 2$ it is $\frac{1}{2}e^{-2\tau}$, and the rate 2 is a two-particle state.

In d dimensions, by (3) positive exponentials line up momentum by momentum, so reflection positivity holds (problem statement § 6.2); reconstruction yields Fock space, and the spectrum is $\{0\} \cup [m, \infty)$ (2-06 § 08). Known The action of Yang–Mills has no mass term (3-02), and the gap must be an output of the interaction (4-11).

08

Regularisations that break reflection positivity, and the lattice that keeps it

Regularisations that suppress the ultraviolet (4-04) can break reflection positivity. Example: the Pauli–Villars type. Subtracting the propagator of a heavy mass M makes it decrease fast, as p^{-4}, at large p.

$$\frac{1}{p^2+m^2} - \frac{1}{p^2+M^2} = \frac{M^2-m^2}{(p^2+m^2)(p^2+M^2)} > 0, \qquad C_{\mathrm{PV}}(\tau) = \frac{e^{-m|\tau|}}{2m} - \frac{e^{-M|\tau|}}{2M} \tag{11}$$

It is legitimate as a covariance, but in the time direction the exponential weight is negative at $E = M$. At $m = 1$, $M = 3$, $CC'' - C'^2 = -\frac{1}{3}e^{-4\tau} < 0$, so $\log C$ is concave everywhere, and the test of (5) is violated at any two distinct points (the last row of Table 2) Computed . A negative weight corresponds to states with negative inner product (ghosts; a different meaning from the "ghost" of 4-08 § 03, which refers to the Faddeev–Popov anticommuting fields) Known . § 6.5 of the problem statement writes that there are few ways of recovering reflection positivity lost through regularisation, and gives dimensional regularisation and Pauli–Villars regularisation as examples (quoted in 0-03 § 03).

The lattice is the opposite. The Wilson action keeps reflection positivity (Osterwalder–Seiler 1978, problem statement § 6.5), and the transfer matrix and the Hamiltonian stand up (for an example, 0-04 § 06; for gauge theories, 5-03). In exchange, rotations are cut down to steps of 90°. Reflection positivity is an inequality, so it passes to the limit (7-01 § 07) Known .

09

What gauge theories add — quantities on curves, gauge fixing, volume

Observables. What has meaning are gauge-invariant quantities, and the local ones are composite quantities such as $\mathrm{Tr}\, F_{\mu\nu}F_{\rho\sigma}(x)$ (problem statement § 4). On Wilson loops (3-02), Chatterjee, citing Seiler, writes 'the key observables in these theories do not take values at points, but at curves'. **Gauge fixing.** Correlations of fixed fields are in general not reflection positive, and covariant gauges use spaces with indefinite inner products (4-08). Magnen–Rivasseau–Sénéor 1993 say that covariant gauges such as the Feynman gauge 'cannot be used directly at the beginning because of their lack of positivity', use a regularised axial gauge, and go no further than expecting that OS positivity 'could be shown to hold with some additional work'. **Volume.** Even on T^4, if a reflection-positive measure can be made, a finite-volume Hilbert space and transfer matrix are obtained (this is so for the periodic box of the lattice; the construction itself on the continuum T^4 is unfinished; problem statement § 6), but a Poincaré-covariant theory on $\mathbb{R}^4$ needs $T^4 \to \mathbb{R}^4$ (0-04 § 09) Known .

Example. From Driver's 1989 measure in two dimensions, the Euclidean invariance of gauge-invariant expectation values follows, but the full set of axioms is not claimed (7-02). § 6.4 of the problem statement calls the two-dimensional abelian Higgs model (Brydges–Fröhlich–Seiler) 'the only complete example of an interacting gauge theory satisfying the axioms' (an assessment from 2000), and § 6.5 continues 'Establishing a quantum mechanical Hilbert space is part of the solution to this Millennium problem.' Known . The construction published in 2026-11 built a Hilbert space from the Schwinger functions of the lattice limit by the reconstruction of this chapter, and checks the construction's own version of the Wightman axioms (the committee's assessment is that it is proved in Lean; 12-08). The volume limit is along a subsequence, and the limit is obtained as a mixture of phases (12-07).

What this chapter can say

Content	Status
Sections 01 to 03, 07 to 09	Known (the text of the OS papers not obtained)
The argument of sections 04 and 05	Proof written in one-dimensional outline Known . Real time, boosts and W_n only stated
The reading on gluons at the end of section 04	Physics (lattice numbers and their interpretation; not a theorem)
Tables 2 and 3, Figures 1 and 2, the test of (11), the detour	In double precision Computed

Content	Status
Axioms, Hilbert space and gap for four-dimensional Yang–Mills	A theory satisfying the axioms and a Hilbert space were built by the construction published in 2026-11 (at the level of the committee's assessment; 6-00). The gap has not been shown (6-04). This chapter shows nothing

Sources and reproduction

Item	Kind	Source / tools
Tables 2 and 3, the figures, the numbers in the text	Computed	python3, NumPy, SciPy (eigenvalues, diagonalisation, the generalised eigenvalue problem, random seed 1)
Quotations from the problem statement	Known	A. Jaffe, E. Witten, *Quantum Yang–Mills Theory* (Clay Mathematics Institute, 2000) § 3, § 4, § 6.2, § 6.4, § 6.5. Checked in the text
The five conditions (E0 to E4) and the linear-growth condition (E0′)	Known	K. Osterwalder, R. Schrader, Comm. Math. Phys. 31 (1973) 83–112; 42 (1975) 281–305. Bibliographic data only. The numbering and the form of the linear-growth condition follow Summers § 3 and a secondary source (Wikipedia "Osterwalder–Schrader theorem")
Schwinger's remark	Known	Discussion at the 1958 CERN conference (proceedings p.140, a reply to a question by Pauli). Via Jaffe 2018; the proceedings not read
History, $S_n = W_n(x^E)$, the role of the linear-growth condition	Known	A. Jaffe, arXiv:1802.07880; S. J. Summers, arXiv:1203.3991 (§ 3, footnote 15). Text checked
Nelson, boosts, the Laplace transform	Known	E. Nelson, J. Funct. Anal. 12 (1973) 97–112, 211–227; J. Fröhlich, K. Osterwalder, E. Seiler, Ann. of Math. 118 (1983) 461–489; A. Klein, L. Landau, J. Funct. Anal. 44 (1981) 121–137; D. V. Widder, Bull. Amer. Math. Soc. 40 (1934) 321–326. Bibliographic data only
The argument of section 05, the Wilson action	Known	J. Glimm, A. Jaffe, *Quantum Physics*, 2nd ed. (Springer, 1987), bibliographic data only (the relevant chapter not checked); K. Osterwalder, E. Seiler, Ann. Phys. 110 (1978) 440–471. Bibliographic data only
The gluon correlations at the end of section 04	Physics	A. Cucchieri, T. Mendes, A. R. Taurines, Phys. Rev. D 71 (2005) 051902; P. O. Bowman et al., Phys. Rev. D 76 (2007) 094505 (lattice QCD with quarks). Bibliographic data only (volume and page checked with Crossref)
Section 09	Known	S. Chatterjee, arXiv:1803.01950 § 1; J. Magnen, V. Rivasseau, R. Sénéor, Comm. Math. Phys. 155 (1993) 325–383 (§ I); B. K. Driver, Comm. Math. Phys. 123 (1989) 575–616 (introduction). The MRS quotations are from § I. Bochner's theorem is defined in the text

Next chapter to read: 4-11 The definition of the mass gap. For reflection positivity on the lattice, 5-03.

The definition of the mass gap — the lowest energy above the vacuum, and the uniform decay of correlations

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — The mass gap is "the infimum of the energy of states orthogonal to the vacuum", and in imaginary time this is equivalent to "all connected correlations decay at least at one rate". A single correlation, or measurements over a finite range, give only upper bounds. How particle states, clustering and the correlation length are like and unlike the gap.

Prerequisite chapters — 4-10 The Osterwalder–Schrader axioms and reconstruction, 2-08 Statistical mechanics and phase transitions. Tools from 1-10 (the spectral theorem, the variational principle). $\hbar = c = 1$, $d = 4$.

Said first — This organises definitions and known correspondences, and shows nothing about the gap of four-dimensional Yang–Mills. The Hilbert space and Hamiltonian that asking about a gap presupposes were built by the construction published in 2026-11 (the committee's assessment is that it is proved in Lean; 6-00). Whether that H has a gap has not been shown (6-04). All the computed examples are free fields.

Order of this chapter

1. Definition — the infimum of the energy of states orthogonal to the vacuum
2. The imaginary-time face — equivalent to uniform decay of all correlations
3. Decay observable by observable (Figure 1, Table 1)
4. A finite window — measurements give only upper bounds (Figure 2, Table 2)
5. How it differs from particle states
6. Clustering
7. The correlation length (Table 3)
8. Finite boxes and uniformity

9. Reading it in gauge theory
10. Appendix table: the uses of the word "gap" — equivalent, upper bound, different (appendix table)

01

Definition — the infimum of the energy of states orthogonal to the vacuum

The definition in § 4 of the problem statement is 'A quantum field theory has a mass gap Δ if H has no spectrum in the interval $(0,\ \Delta)$ for some $\Delta > 0$. The supremum of such Δ is the mass m, and we require m < ∞.' (for a word-by-word reading, 0-03 § 06). This sentence presupposes (a) a Hilbert space $\mathcal{H}$ and a self-adjoint $H \geq 0$ (what the reconstruction of 4-10 § 05 gives) and (b) a vacuum Ω unique up to a phase, with $H\Omega = 0$ (clustering in Table 1 of 4-10), and (c) asks about states orthogonal to the vacuum. In the words of § 1, 'every excitation of the vacuum has energy at least Δ'. Under (b), the definition is the same as the following Known .

$$m = \inf \sigma\big(H|_{\Omega^\perp}\big) = \inf_{\psi \perp \Omega,\ \psi \neq 0} \frac{\langle \psi, H\psi\rangle}{\langle \psi, \psi\rangle}, \qquad \text{there is a gap} \iff m > 0 \quad (1)$$

The second equality is the variational principle (1-10 § 06). That (b) is needed can also be read from (1). If there were a state Ω' of energy 0 orthogonal to the vacuum, (1) would be 0 with $\psi = \Omega'$, but the first sentence of the problem statement alone would miss it, since 0 lies outside $(0, \Delta)$. Example. For the free field of mass m, (1) is also m (we use the same letter), and for mass 0 it is 0 (4-10 § 07).

02

The imaginary-time face — the gap is equivalent to "all correlations decay uniformly"

In the notation of 4-10 § 05, a bounded quantity F at positive times gives a vector $[F]$, and the shift U_τ by τ in the time direction becomes $e^{-\tau H}$. Putting $\psi_F = [F] - \mathbb{E}[F]\,\Omega$, for the **connected correlation** (the correlation with the vacuum component subtracted),

$$\mathbb{E}\big[\Theta G\, U_\tau F\big] - \mathbb{E}[G]\,\mathbb{E}[F] = \langle \psi_G, e^{-\tau H}\psi_F\rangle, \qquad \big|\langle \psi_G, e^{-\tau H}\psi_F\rangle\big| \leq e^{-m\tau}\,\|\psi_G\|\,\|\psi_F\| \quad (2)$$

The inequality follows from the spectral theorem (1-10 § 05) Known . The converse also holds. If $\langle \psi_F, e^{-\tau H}\psi_F\rangle \leq e^{-\Delta\tau}\|\psi_F\|^2$ for every bounded F, then $m \geq \Delta$.

To see this we use the **spectral measure** μ_F of ψ_F: the distribution of weight along the energy axis E when ψ_F is split into eigencomponents of H. If there is an eigenbasis e_k ($He_k = \lambda_k e_k$), then $\mu_F = \sum_k |\langle e_k, \psi_F\rangle|^2\, \delta_{\lambda_k}$ (δ_λ is a weight 1 placed at the point λ), and even for continuous spectrum a distribution with the same meaning is fixed by the spectral theorem. Then $\langle \psi_F, e^{-\tau H}\psi_F\rangle = \int e^{-E\tau} d\mu_F(E)$. The left-hand side satisfies $\int e^{-E\tau} d\mu_F \geq e^{-\Delta'\tau}\mu_F([0,\Delta'))$ ($\Delta' < \Delta$), so as $\tau\to\infty$, $\mu_F([0,\Delta')) = 0$. The set of all ψ_F is **dense** in $\Omega^\perp$ (every vector of $\Omega^\perp$ can be approached arbitrarily closely), so $H \geq \Delta$ on $\Omega^\perp$. Since this includes $E = 0$, the uniqueness of the vacuum comes out as well Known . **The gap is equivalent to all connected correlations decaying at least at one rate that does not depend on the observable.**

A single correlation $C_F(\tau) = \int e^{-E\tau} d\mu_F(E)$ is **log-convex**, that is, $\log C_F$ is convex ($C(\tau_1+\tau_2)^2 \leq C(2\tau_1)\,C(2\tau_2)$ in (5) of 4-10 corresponds to this). So the **effective mass** decreases monotonically.

$$m_{\text{eff}}(\tau) = -\frac{d}{d\tau}\log C_F(\tau) \;\searrow\; E_F := \min \operatorname{supp}\mu_F, \qquad m = \inf_F E_F \quad (3)$$

$\operatorname{supp}\mu_F$ (the **support**) is the smallest closed set carrying the weight of μ_F, and E_F is the lowest energy that ψ_F touches. m_{eff} at any time is an upper bound on E_F: from above it is bounded by a single F, but from below every F is needed (the same shape as the box "The asymmetry of upper and lower bounds" in 1-10). On the lattice U_a is the transfer matrix $T = e^{-aH}$, $m_{\text{lat}} = \log(\lambda_0/\lambda_1)$ and $\Delta = m_{\text{lat}}/a$. λ_0 and λ_1 are the largest and second-largest eigenvalues of T (in gauge theory, on gauge-invariant states; section 08) ((1) of 2-09 is its smallest example).

03

Decay observable by observable — a single correlation sees only the part its observable touches

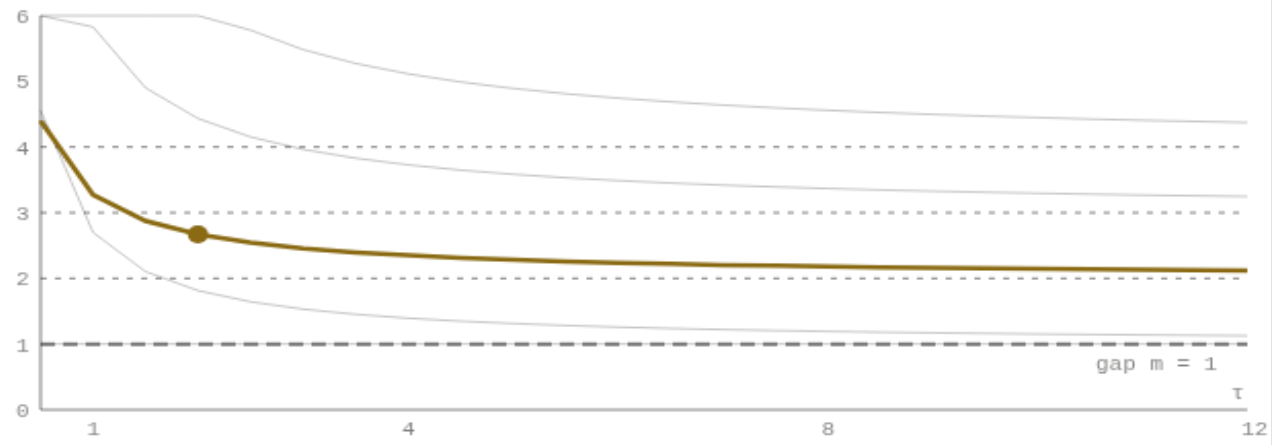
We see that E_F changes with the observable in the four-dimensional free field ($m = 1$). The correlation of O_k, the Wick products of 4-10 § 07 integrated over space to give momentum 0, is

$$O_k(\tau) = \int d^3x :\!\phi^k\!:(\tau,\vec{x}), \qquad \frac{\mathbb{E}[O_k(\tau)O_k(0)]}{\text{volume}} = k!\int d^3x\, S(\tau,\vec{x})^k, \qquad S(x) = \frac{m\,K_1(m|x|)}{4\pi^2|x|}$$

(S is the two-point function of the four-dimensional free field, K_1 a modified Bessel function). O_k creates exactly k particles with total momentum 0, so $EO_k = km$. At large τ the energy of each particle can be approximated as $m + |\vec{p}|^2/(2m)$, and the correlation of k particles with total momentum 0 is $e^{-km\tau}$ multiplied by $\tau^{-3(k-1)/2}$

from the Gaussian integral over the three-dimensional momenta of the remaining $k-1$. So near the lower end $m_{\text{eff}} \approx km + \frac{3(k-1)}{2\tau}$ Computed . Only $k=1$ sees the gap 1; however precisely $:\phi^2:$ is measured, it only approaches 2. This is because the number of particles is conserved for the free field; even with interactions, symmetries split the space into parts (for Yang–Mills, the channels of section 09). For the unprojected point-to-point $S(\tau, 0)$, even with the same ϕ, it decreases slowly, like $1 + \frac{3}{2\tau}$. The energies of moving particles form a continuum, and the integral over one three-dimensional momentum multiplies by $\tau^{-3/2}$ for the same reason; this is why lattice measurements sum over space (5-08).

Figure 1 Choosing an observable and moving τ, the effective mass (thick line and circle) comes down towards the lower end for that observable (thin dashed line). Only φ comes down to the gap m = 1 (thick dashed line)



:φ²:, τ = 2.0: m_eff = 2.671 (lower end 2, gap 1)

Computed The numbers in this figure were computed on this machine (a table of (4) at steps of 0.5 in τ is embedded).

Computed Table 1 $m_{\text{eff}}(\tau)$ of the free field ($m=1$). In brackets, the lower end E_F. The momentum-0 ϕ is 1.00000 at every τ, agreeing with the closed form $e^{-\tau}/2$ within double-precision rounding.

τ	φ point (1)	:φ²: (2)	:φ³: (3)	:φ⁴: (4)
2	1.81431	2.67113	4.43494	6.22414
4	1.39395	2.35013	3.72854	5.11460
8	1.19273	2.18027	3.36864	4.55915
16	1.09513	2.09177	3.18575	4.28029

Every column decreases monotonically (checked at all points at steps of 0.25 for $\tau \le 12$), agrees with the asymptotic form to within 0.13% at $\tau = 16$, and stays above the lower end to the last. What can be read is an upper bound on E_F Computed .

04

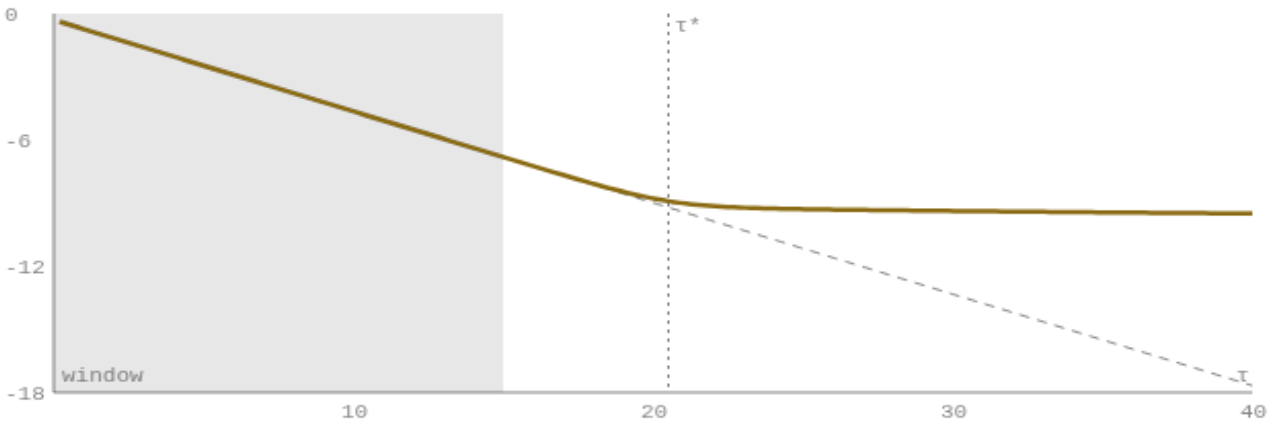
A finite window — measuring the decay gives no lower bound on the gap

We call the range of τ that a measurement can see a **window** (precisely, a finite interval $[\tau_{\min}, \tau_{\max}]$). The infimum (3) involves infinitely many F and $\tau \to \infty$, so a measurement within a window sees only part of it. We look at a theory with no gap, made by placing side by side, independently, a free field ϕ of mass 1 and a free field χ of mass 0. The cross terms vanish, and the correlation of $O = \int d^3x\, [\phi + \varepsilon :\chi^2:]$ is

$$\frac{\mathbb{E}[O(\tau)O(0)]}{\text{volume}} = \frac{e^{-\tau}}{2} + \frac{\varepsilon^2}{8\pi^2\tau}, \qquad m_{\text{eff}}(\tau) = \frac{\frac{1}{2}e^{-\tau} + \frac{\varepsilon^2}{8\pi^2\tau^2}}{\frac{1}{2}e^{-\tau} + \frac{\varepsilon^2}{8\pi^2\tau}} \quad (5)$$

(the coefficient of the second term checked to 12 digits by numerical integration Computed). Up to the τ^* at which the two terms are equal, $m_{\text{eff}} \approx 1$, and it looks like "a theory with gap 1".

Figure 2 Making ε smaller, the point τ* where the correlation (thick line, vertical axis in common logarithms) leaves $e^{-\tau}/2$ (dashed line) runs away to the right. Moving the edge of the window changes m_{eff} there



ε = 1e−3: τ* = 20.51. At the window edge τ = 15.0, m_eff = 0.9949

Computed The numbers in this figure were computed on this machine ((5) computed on the spot; τ* by bisection).

Computed Table 2 τ^* and m_{eff} of (5).

ε	τ*	m_eff(20)
10^{-1}	10.646	0.0502
10^{-2}	15.636	0.0652
10^{-3}	20.512	0.6384
10^{-4}	25.328	0.9942
10^{-6}	34.858	1.0000

Setting the two terms equal gives $\tau^* = \ln(1/\varepsilon^2) + \ln(4\pi^2\tau^*)$, so each time ε becomes an order of magnitude smaller, τ^* moves out by about 5, $2\ln 10 \approx 4.6$ plus a logarithmic correction, and for any finite window there is an ε that looks like "gap 1" (far enough out, $m_{\text{eff}} \approx 1/\tau$). Measurements in a finite window give no lower bound on the gap Known . For how to read lattice numbers, 16-08 § 07.

05

How it differs from particle states — an isolated eigenvalue is a separate requirement from the gap

The gap of the mass operator $M = \sqrt{H^2 - \vec{P}^{\,2}}$ had the same width as the gap of H (0-03 § 06). **A one-particle state of mass** μ is a state belonging to the eigenvalue μ of M (Wigner's 1939 idea of seeing particles as irreducible representations of the Poincaré group), and it is **isolated** if there are no other points of $\sigma(M)$ near μ. § 5 of the problem statement places this outside the requirements, writing 'one would like to prove the existence of an isolated one-particle state (an upper gap, in addition to the mass gap)' Known .

The difference shows in the two-point function. The Euclidean two-point function of a scalar field ϕ ($\langle\Omega, \phi\,\Omega\rangle = 0$) can be written with a positive measure ρ as

$$\widehat{S}_2(p) = \int_{[0,\infty)} \frac{d\rho(s)}{p^2 + s}, \qquad \int d^3x\, S_2(\tau, \vec{x}) = \int \frac{e^{-\sqrt{s}\,\tau}}{2\sqrt{s}}\, d\rho(s) \quad (6)$$

which is called the Källén–Lehmann representation Known (the right-hand side is (3) of 4-10 superposed over s). If there is a gap m, the support of ρ lies in $[m^2, \infty)$, and if ϕ touches a particle of mass μ, ρ has a point weight at $s = \mu^2$.

Example 1: a gap but no particle. Take ρ to be the uniform weight on $[m^2, 4m^2]$, and take the Gaussian measure with covariance $\int d\rho(s)(-\Delta + s)^{-1}$. Reflection positivity is a condition preserved by positive superpositions of covariances, so if it holds for each s (4-10 § 07) it also holds after integrating with the positive measure ρ. On reconstruction, the "one-particle" M fills $[m, 2m]$ continuously, and $\sigma(H) = \sigma(M) = \{0\} \cup [m, \infty)$. The gap is m, and M has no eigenvalue other than 0. Fields of this kind are called generalised free fields (Greenberg 1961) Known . **Example 2: a particle but no gap.** In the theory of section 04, the one-particle state of ϕ has $M = 1$, but two-particle states of χ fill $[0, \infty)$, and the eigenvalue 1 is buried in the continuous spectrum.

What physics expects for Yang–Mills is an isolated glueball at the lower end of the gap Physics . For the difficulty of telling an isolated point from the edge of a continuum using correlations, 16-08 § 05.

06

Clustering — how fast do things become independent when moved apart in a spatial direction

Clustering without regard to rate (two quantities moved apart approach independence) is Osterwalder–Schrader axiom E4, paired with the uniqueness of the vacuum (Table 1 of 4-10) Known . The massless free field also satisfies it, but its correlation decreases only as a power, $1/(4\pi^2|x|^2)$. Formula (2) of § 5 of the problem statement says, for a local field $\mathcal{O}(\underset{\sim}{x}) = e^{-i\underset{\sim}{P}\cdot\underset{\sim}{x}}\mathcal{O}e^{i\underset{\sim}{P}\cdot\underset{\sim}{x}}$ with $\langle\Omega, \mathcal{O}\Omega\rangle = 0$,

$$|\langle\Omega, \mathcal{O}(\underset{\sim}{x})\,\mathcal{O}(\underset{\sim}{y})\,\Omega\rangle| \le e^{-C|\underset{\sim}{x}-\underset{\sim}{y}|} \qquad (0 < C < \Delta,\ \ |\underset{\sim}{x}-\underset{\sim}{y}| \text{ large enough}) \quad (7)$$

The outline on the Euclidean side is (2) and rotation invariance (Euclidean covariance; Osterwalder–Schrader axiom E1). If two quantities localised within radius δ are a distance R apart, rotating the direction joining them onto the τ axis, (2) gives a connected correlation at most (a constant independent of R) $\times e^{-m(R-2\delta)}$. At equal times and spacelike-separated points, the Wightman function (the vacuum expectation value of a product of fields; 4-09 § 04) equals the value of the Schwinger function at time 0 (the analytic continuation of 4-10 § 05), so this decay of the Euclidean correlation becomes (7) directly. The point field $\mathcal{O}(\underset{\sim}{x})$ is not a bounded quantity, so precisely one considers the field slightly smeared over radius δ. The "$C < \Delta$" and "large enough" in (7) account for the constant in front and the width 2δ. The theorem on the Minkowski side is introduced in the introduction of Nachtergaele–Sims 2006: 'in [1] Fredenhagen proves a general theorem applicable to arbitrary strictly local theories' ([1] is Fredenhagen 1985) Known .

The converse does not hold in general; the same introduction writes that there are quantum lattice systems with a unique ground state and exponentially decaying correlations but no gap, and gives Example 2 of Nachtergaele 1996 as an example Known . In a rotation-invariant Euclidean theory, time is one of the directions. If "it decays at a uniform rate in every direction", it decays at the same rate in the time direction too, so the converse argument of section 02 (where the constant in front may depend on F) applies directly, and this is equivalent to the gap. A state mixing two phases (2-09 § 08) breaks clustering itself, and assumption (b) of (1) fails.

07

The correlation length — the inverse of the gap, said as a length

(6) of 2-08 defined a correlation length from a single correlation. The correlation length of the theory is the supremum over observables.

$$\frac{1}{\xi_F} = -\lim_{\tau\to\infty}\frac{1}{\tau}\log C_F(\tau) = E_F, \qquad \xi := \sup_F \xi_F = \frac{1}{m} \tag{8}$$

Both are restatements of (3) Known (restoring units, $\xi = \hbar c/\Delta$; on the lattice $\xi_{\text{phys}} = a\,\xi_{\text{lat}} = 1/\Delta$). The lattice correlation-length problem of Chatterjee 2018 (Chatterjee's Problem 5.1) defines the lattice correlation length by plaquette correlations, and writes 'The number ξ represents the reciprocal of the mass of the lightest glueball' Known .

$\xi_{2\text{nd}}^2 = \sum_x |x|^2 G(x) / \big(2d\sum_x G(x)\big)$ from second moments is also often used, and on the lattice the rate of decay also depends on the direction (4-02). What equals the inverse of the gap is the length along the direction of the transfer matrix (an axis). For the free field on the four-dimensional lattice (covariance $(-\Delta_{\text{lat}} + m_0^2)^{-1}$) there are closed forms.

$$\cosh\frac{1}{\xi_{\text{axis}}} = 1 + \frac{m_0^2}{2}, \qquad \xi_{2\text{nd}} = \frac{1}{m_0}, \qquad \text{decay rate along } e = \max\Big\{q\cdot e : \sum_\mu (2\cosh q_\mu -$$

The e in the third is a vector of length 1: on an axis $e = (1,0,0,0)$, on the diagonal $e = (1,1,1,1)/2$.

Computed Table 3 The free field on the four-dimensional lattice. The two columns on the right are ratios to ξ_{axis} (diagonal means the direction $e = (1,1,1,1)/2$). The third formula of (9) was checked by computing the two-point function on a two-dimensional periodic lattice with the fast Fourier transform, agreeing to within 0.11% on the axis and the diagonal.

m_0	ξ_{axis}	$\xi_{\text{diag}}/\xi_{\text{axis}}$	$\xi_{2\text{nd}}/\xi_{\text{axis}}$
2	0.5673	0.91579	0.88137
1	1.0390	0.97228	0.96242
0.2	5.0083	0.99876	0.99834
0.05	20.0021	0.99992	0.99990

When the correlation length is of the order of the lattice spacing ($m_0 = 1, 2$), the three lengths disagree by 3 to 12%; at $\xi_{\text{axis}} = 20$ the differences are of order 10^{-4}, and rotational symmetry returns (the continuum limit; 2-08 § 08). Outside free fields there is no guarantee that the lengths agree, and what is used for the gap is ξ_{axis}.

08

Finite boxes and uniformity — the gap is asked about on the side of the limit

On a finite lattice, a positive gap is almost trivial. In the temporal gauge, the transfer matrix T (5-03 § 05) is an integral operator with a positive continuous kernel acting on the space $L^2(G^n)$ of values of the links at one time (G a compact group, n the number of links). A continuous kernel on a compact space makes T a compact operator (precisely, a norm limit of finite-dimensional matrices), and its eigenvalues form a discrete sequence accumulating only at 0. Further, since the kernel is positive, Jentzsch's theorem (the infinite-dimensional version of the Perron–Frobenius theorem) makes the largest eigenvalue λ_0 simple, and $\lambda_1 < \lambda_0$ Known . If the group is finite, T is a finite matrix with positive entries, and it is literally the Perron–Frobenius theorem. Simplicity of the largest eigenvalue alone leaves open that the other eigenvalues crowd up to λ_0, so discreteness is needed. What is asked is whether $m_{\mathrm{lat}}(a, L)/a$ stays positive in the limits of volume $L \to \infty$ and lattice spacing $a \to 0$. § 6.5 of the problem statement writes 'New ideas are needed to prove the existence of a mass gap that is uniform in the volume of space-time.', and § 5 also says that a uniform gap can play a role in the proof of the infinite-volume limit Known . (2) is an inequality between correlations (the $\|\psi_F\|^2 = \mathbb{E}[\Theta F F] - |\mathbb{E}[F]|^2$ on the right-hand side is also a correlation). So if it holds along a sequence of approximations with the same Δ and the correlations converge, it holds in the limit too. Using the converse of section 02 in the limiting theory gives $m \geq \Delta$ (that the set of all ψ_F is dense in $\Omega^\perp$ comes from the way the reconstruction is built; 7-01).

The ways a finite-box gap closes as $L \to \infty$ are massless particles ($2\pi/L$ in a periodic box; 2-05 § 06) and symmetry breaking (the tunnelling splitting between phases; 2-09 § 08). In the latter each phase in infinite volume has a gap, and § 6.3 of the problem statement writes that each phase of two-dimensional ϕ^4 has 'a non-zero mass gap and a unique, Poincaré-invariant vacuum state' Known . As $a \to 0$, if Δ is held fixed then $m_{\mathrm{lat}} = a\Delta \to 0$, and a gap in lattice units that is uniformly positive means $\Delta = \infty$ in the continuum limit (5-09).

09

Reading it in gauge theory — the gap on gauge-invariant states

For Yang–Mills, the $\mathcal{H}$ of (1) is the space of gauge-invariant states (4-01), and the observables of (2) and (3) are local quantities built from the field strength such as $\mathrm{Tr}\, F_{ij} F_{kl}(x)$ (problem statement § 4) and Wilson loops (4-10 § 09). States split by the combination J^{PC} of spin, parity and charge conjugation (channels) (J is the spin, and P and C are the signs $\pm$ under spatial reflection and under the operation that replaces the field by its complex conjugate [the antiparticle]), and, as with k in section 03, the minimum over channels of the lowest energy in each is m. In lattice numbers 0^{++} is

the lowest (about 0.7 times 2^{++} for pure $SU(3)$; Chen et al. 2006; 5-08), and physics reads it as the lower end of the gap Physics .

Correlations of gauge-fixed fields are in general not guaranteed to be reflection positive (4-10 § 04, § 09), and for the gluon propagator it is read as violated Physics . The decay of a correlation that is not reflection positive does not give points of $\sigma(H)$. The string tension is fixed by the energy of a pair of static charges, and is a different quantity from excitations above the vacuum (16-08 § 06). In mathematics there are long-standing results on the gap in lattice units at strong lattice coupling (small β) (5-04). Shen–Zhu–Zhu 2023, in 't Hooft's convention (writing the coupling of the action in the form $N\beta$; for conversion to other conventions, Table 4 of 5-01), showed that for $SU(N)$ with $|\beta| < 1/(16(d-1))$ the correlations of a certain family of observables decay exponentially. The abstract calls this 'strictly positive mass gap', and also writes 'Our method improves earlier results' Known . To use it in the equivalence of section 02, that family must be dense. This is a result short of the limits of section 08. For the four-dimensional continuum theory, the $\mathcal{H}$ and H of (1) were built by the construction published in 2026-11 (at the level of the committee's assessment; 12-08), but their gap has not been shown (6-04; for where things stood before that, 7-10).

10

Appendix table: the uses of the word "gap" — the same word, equivalent, upper bound, or different

More than ten quantities called "gap" appear in this book. Some are **equivalent** to the Δ of the problem statement, some are only **upper bounds** on Δ, and some are **different** (the gap of another operator, or an additional requirement); confusing them changes the direction of a claim. The table below lists, one row each, the definition, the main chapters and the relation to Δ (the assignment of relations is this book's organisation; for the direction of upper and lower bounds, the box "The asymmetry of upper and lower bounds" in 1-10).

Appendix table The word "gap". "Equivalent" in the relation column means the same number under the stated assumptions.

#	Name	Definition (one line)	Chapter	Relation to Δ
1	Gap of the Hamiltonian (the mass gap)	The infimum Δ of the spectrum of H on $\Omega^{\perp}$	§ 01, 0-03 § 06	Reference (the Δ of the problem statement)
2	Gap of the mass operator	The next point after 0 for $M = \sqrt{H^2 - P^2}$	§ 05, 0-03 § 06	Equivalent (same width)

#	Name	Definition (one line)	Chapter	Relation to Δ
3	Uniform decay rate	The largest m such that all connected correlations decay at least as fast as $e^{-m\tau}$	§ 02	Equivalent ((2), (3); under reflection positivity)
4	Inverse correlation length	$1/\xi$, $\xi = \sup_F \xi_F$ (along the direction of the transfer matrix)	§ 07	Equivalent ((8)). $\xi_{2\mathrm{nd}}$ and the length along the diagonal are different (close for free fields)
5	Effective mass	$m_{\mathrm{eff}}(\tau) = -\frac{d}{d\tau}\log C_F(\tau)$	§ 02, § 04, 18-01 § 08	Upper bound ($m_{\mathrm{eff}}(\tau) \geq E_F \geq \Delta$)
6	Gap of the transfer matrix (lattice units)	$m_{\mathrm{lat}} = \log(\lambda_0/\lambda_1)$	§ 02, 5-03	Equivalent on the lattice ($T = e^{-aH}$)
7	Gap in physical units	$\Delta = m_{\mathrm{lat}}/a$	§ 08, 5-09	Related to 6 by a. This is what is asked about in the continuum limit
8	Finite-volume gap	$\lambda_1 < \lambda_0$ of T in a box of side L	§ 08	Always positive (Jentzsch's theorem). Says nothing about Δ
9	Uniform gap	$\inf m_{\mathrm{lat}} > 0$ over the volume L (and β in a fixed range)	§ 08, 5-04, 16-05	A lower bound on 6 in infinite volume. A gap in lattice units uniform in β is, in physical units, on the side of $\Delta = \infty$ (a false friend; 18-01)
10	Upper gap	The width by which the one-particle eigenvalue μ is separated from the rest of $\sigma(M)$	§ 05	Different (an additional requirement in § 5 of the problem statement)
11	Langevin gap	The gap $1/C_P$ of the generator L of the path measure (the lattice Wilson measure)	1-10 § 09, 18-03 § 05	Different (the gap of an artificial time). For free fields $1/C_P = m^2$ (the square of the gap). Assuming they line up is 23-01 The line of construction and the diffusive-relaxation-time assumption
12	Poincaré constant of the time-zero measure	C_P of the ground-state distribution $\phi_0^2 dx$	2-05 § 08	Equivalent (ground-state transformation; with the normalisation $C_P = \hbar^2/(2m\Delta)$)

11 and 12 are both Poincaré constants, but 11 concerns the measure on whole paths in imaginary time and 12 the measure at a single time, and only 12 is equivalent (2-05 § 08). 5 and 9 point in opposite directions: 5 is an upper bound obtained from a single observable, 9 a lower bound over all observables and volumes.

What this chapter can say

Content	Status
(1) to (3) and (8), the examples of section 05, the finite-lattice gap of section 08	Known . Outlines written in the text
Tables 1 to 3, Figures 1 and 2, the checks of (5) and (9), the comparison with the asymptotic forms of section 03	For free fields, in double precision Computed
The Euclidean side of (7)	Outline only. The Minkowski side is Fredenhagen 1985 (bibliographic data only)
"Equivalent, upper bound, different" in the appendix table of section 10	This book's organisation. The grounds for each row are the text and the chapters cited (restatements of Known results)
The glueballs of section 09	Physics (lattice numbers and extrapolation)
The gap of four-dimensional Yang–Mills	$\mathcal{H}$ and H were built by the construction published in 2026-11 (at the level of the committee's assessment; 12-08). The gap has not been shown (6-04). This chapter shows nothing

Sources and reproduction

Item	Kind	Source / tools
Tables, figures, numbers in the text	Computed	python3, NumPy, SciPy (numerical integration, bisection, FFT)
Problem statement §1, §4, §5, §6.3, §6.5, the lattice correlation-length problem	Known	A. Jaffe, E. Witten, *Quantum Yang–Mills Theory* (Clay, 2000); S. Chatterjee, arXiv:1803.01950 §5. Checked in the text
Section 06	Known	B. Nachtergaele, R. Sims, Comm. Math. Phys. 265 (2006) 119–130 (arXiv:math-ph/0506030), introduction checked; K. Fredenhagen, 'A Remark on the Cluster Theorem', Comm. Math. Phys. 97 (1985) 461–463, bibliographic data only; B. Nachtergaele, Comm. Math. Phys. 175 (1996) 565–606, Example 2, bibliographic data only
Section 05	Known	H. Lehmann, Nuovo Cimento 11 (1954) 342–357; O. W. Greenberg, Ann. Phys. 16 (1961) 158–176; E. Wigner, Ann. of Math. 40 (1939) 149; G. Källén 1952 'On the Definition of the Renormalization Constants in Quantum Electrodynamics' (journal and volume of first publication not checked). Bibliographic data only
Section 08	Known	K. Osterwalder, E. Seiler, Ann. Phys. 110 (1978) 440–471, bibliographic data only (for the construction of the transfer matrix, 5-03)

Item	Kind	Source / tools
Section 09	Physics, Known	Y. Chen et al., Phys. Rev. D 73 (2006) 014516, Table VIII (text); H. Shen, R. Zhu, X. Zhu, Comm. Math. Phys. 400 (2023) 805–851 (abstract of arXiv:2204.12737)

Next chapter to read: 4-12 Detour: the perturbation series diverges. For what to hold fixed in the limit, 5-09.

Detour: the perturbation series diverges — the factorial in the coefficients is often the shadow of an e^{-c/g^2} that the series does not show

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — Where coefficients growing like $n!$ come from (negative coupling, saddle points, the ends of momentum integrals). Optimal truncation and the Borel sum, and how their ambiguity pairs with terms of size e^{-c/g^2}. The singularities expected in the Borel plane of Yang–Mills.

Prerequisite chapters — 4-04 Divergences and renormalisation (the running coupling). We use the asymptotic series and the zero-dimensional model (2) of 4-03 § 08. From complex analysis, only Cauchy's integral formula is used, stated on the spot.

Said first — The formulas for the models can be derived, and the numbers were checked on this machine. The divergence of the series of four-dimensional Yang–Mills, renormalons, and the relation between the series and the theory are expectations at the level of physics, not theorems. This chapter does not show anything about Yang–Mills.

Order of this chapter

1. Asymptotic series — approaching without converging: the coefficients are fixed by the function, not the other way round
2. Dyson's argument — flipping the sign of the coupling: negative coupling gives an imaginary part $e^{-3/(2s)}$
3. The coefficients are fixed on the flipped side — dispersion relations and saddle points: from the imaginary part, $n!\,(2/3)^n$
4. Optimal truncation — stopping where the terms are smallest: the floor of the error is $e^{-3/(2u)}$
5. The Borel sum — divide by $n!$, restore with an integral: in the model it returns the exact value

6. When there is a singularity on the positive axis — ambiguity and exponentially small terms: the sum splits
7. Renormalons — $n!$ from the ends of momentum integrals: it stops at the order at which the peak reaches Λ
8. The Borel plane of Yang–Mills — a map of the expected singularities: renormalons are the closest
9. Seen from mathematics — what is a theorem and what is an expectation: in four dimensions asymptoticity has not been shown

---

01

Asymptotic series — approaching without converging

The problem of this chapter is that the series of coefficients rendered finite by renormalisation **does not itself converge** (a different matter from the divergences of 4-04). Even so it has uses. If, fixing the order N and letting $u \to 0$, the truncation error vanishes as fast as u^{N+1}, we say that the series is **asymptotic** to f, and write $f(u) \sim \sum a_n u^n$.

$$\left| f(u) - \sum_{n=0}^{N} a_n u^n \right| \le K_{N+1}\, u^{N+1} \qquad (0 < u < u_0,\ N = 0, 1, 2, \dots) \quad (1)$$

The order of the limits is the reverse of convergence. The coefficients are determined uniquely by f Known , but the converse does not hold. $e^{-1/u}$ approaches 0 faster than any u^N, so f and $f + e^{-1/u}$ have the same series (7-12 § 07). The theme of this chapter is that this "part not shown in the coefficients" is recorded in the way the coefficients grow.

Example. That all the coefficients of $e^{-1/u}$ are 0 can be checked in the form of (1) divided by u^N. Putting $v = 1/u$, $e^{-1/u}/u^N = v^N e^{-v}$, and as $u \to 0$ ($v \to \infty$) the exponential beats the power and it approaches 0. This is so for every N, so $a_0 = a_1 = \cdots = 0$. An example in the other direction is the model (2) of this chapter. (14) of 4-03 § 08 shows $\left| Z(u) - \sum_{n<N} a_n u^n \right| \le |a_N|\, u^N$ for $u \ge 0$, and $|a_N|$ can be taken as the K_N of (1). K_N grows factorially with N, but (1) needs only an estimate for each fixed N, so the series is asymptotic to Z.

02

Dyson's argument — flipping the sign of the coupling

The abstract of Dyson (1952) says it presents an argument leading 'tentatively' to the conclusion that the power series of quantum electrodynamics all diverge even after renormalisation Known . Its core is usually presented in the form: "for $e^2 < 0$, like charges attract and the vacuum collapses, so the series cannot continue smoothly to both sides of $e^2 = 0$". It is a heuristic argument. For a small model everything can be computed.

$$Z(u) = \int_{-\infty}^{\infty} \frac{dx}{\sqrt{2\pi}}\, e^{-S_u(x)}, \qquad S_u(x) = \frac{1}{2}x^2 + \frac{u}{4!}x^4 \tag{2}$$

For $u = -s < 0$ the integral does not exist (the "collapse of the vacuum" in the model). Even so, turning the variable as $x \to e^{-i\theta/4}x$ brings the quartic term of $u = s\,e^{i\theta}$ back to $s\,x^4$, so Z extends to complex u. There are two reasons why the contour may be turned. One is that the integral of a function complex-differentiable on and inside a closed path, taken along that path, is 0 (Cauchy's theorem Known ; we use it in the form of the integral formula in § 03). The integrand is the exponential of a polynomial in x, so it is complex-differentiable everywhere. The other is that on the far arc of the sector between the original real axis and the turned line, the real part of the quartic term stays positive and acts at the rate $|x|^4$, so the contribution of the arc vanishes as $|x| \to \infty$. Approaching the negative axis from above,

$$Z(-s + i0) = e^{-i\pi/4} \int_{-\infty}^{\infty} \frac{dx}{\sqrt{2\pi}}\, e^{\,\frac{i}{2}x^2 - \frac{s}{24}x^4} \tag{3}$$

takes a non-real value, and from below its complex conjugate. **The values above and below the negative axis disagree: this is the source of the divergence.** The imaginary part is fixed by the height of the wall. $S_{-s}(x) = \frac{1}{2}x^2 - \frac{s}{24}x^4$ takes its maximum $\frac{3}{2s}$ at $x^2 = 6/s$, and the rough probability of leaking from the valley over the wall is $e^{-3/(2s)}$. Strictly, there is no tunnelling in a zero-dimensional integral. What is meant is that the value $\frac{3}{2s}$ of the action at the saddle point $x^2 = 6/s$ of the integrand of (3) fixes the exponent of the imaginary part, and the leading term as $s \to 0$ is $-e^{-3/(2s)}/\sqrt{2}$ (Table 1).

Computed Table 1 Numerical integration of (3). The imaginary part approaches $-e^{-3/(2s)}/\sqrt{2}$ as $s \to 0$

s	$\operatorname{Im} Z$	$e^{-3/(2s)}/\sqrt{2}$	Ratio
0.50	-3.351e-2	3.521e-2	0.952
0.20	-3.825e-4	3.911e-4	0.978
0.10	-2.138e-7	2.163e-7	0.988

03

The coefficients are fixed on the flipped side — dispersion relations and saddle points

The tool we use is **Cauchy's integral formula**: a function complex-differentiable on and inside a closed path satisfies $f(u) = \frac{1}{2\pi i}\oint \frac{f(w)}{w-u}\,dw$ (anticlockwise) at points inside Known . Z is complex-differentiable on the plane minus the negative axis and decreases as $|w|^{-1/4}$ at infinity, so taking the path to be a large circle and the upper and lower sides of the negative axis, only the difference $2i\,\mathrm{Im}\,Z$ between the sides remains.

$$Z(u) = -\frac{1}{\pi}\int_0^\infty \frac{\mathrm{Im}\,Z(-s+i0)}{s+u}\,ds \quad (4)$$

This is the **dispersion relation** (numerically integrating the right-hand side with the imaginary part computed from (3), it agrees with the left-hand side to order 10^{-6} at $u = 0.5, 1, 2$ Computed). Expanding $\frac{1}{s+u} = \sum_n (-u)^n/s^{n+1}$ and integrating term by term,

$$a_n = \frac{(-1)^{n+1}}{\pi}\int_0^\infty \frac{\mathrm{Im}\,Z(-s+i0)}{s^{\,n+1}}\,ds \quad (5)$$

(since it is not complex-differentiable at $u = 0$, what is obtained is an asymptotic series), and for large n only the imaginary part $\approx -e^{-3/(2s)}/\sqrt{2}$ near $s \to 0$ matters. From $\int_0^\infty e^{-c/s} s^{-n-1}\,ds = \Gamma(n)/c^n$,

$$a_n \approx (-1)^n \left(\frac{2}{3}\right)^n \frac{\Gamma(n)}{\pi\sqrt{2}} \quad (6)$$

The factorial comes from $\Gamma(n) = (n-1)!$, and the base $2/3$ is the inverse of the coefficient of the wall height $3/(2s)$; the alternation of signs comes from the singularity lying on the negative axis.

Computed Table 2 The exact coefficients $a_n = \frac{(4n-1)!!}{n!}\left(-\frac{1}{24}\right)^n$ (rationals) and their ratio to (6)

n	a_n	a_n/(6)
1	-0.1250	0.833
10	1.3901e+3	0.981
20	8.1571e+12	0.991
40	4.1327e+38	0.995

Bringing this line of reading large orders from the action at a saddle point into field theory is attributed to Lipatov (1977), and the review of Beneke (1999) credits Lipatov 1977 with factorial divergence due to instantons (classical solutions in imaginary time; defined in § 06) Known .

The large orders of the ground-state energy $\frac{1}{2}+\frac{3}{4}g-\frac{21}{8}g^2+\cdots$ of $H=\frac{1}{2}p^2+\frac{1}{2}x^2+gx^4$ in quantum mechanics were studied by Bender–Wu (1969, 1973) (abstracts) Known . The ratio of the known form $E_n \approx (-1)^{n+1}\frac{\sqrt{6}}{\pi^{3/2}}3^n\Gamma(n+\frac{1}{2})$ (the text of the source not checked) to the coefficients computed with rationals up to order 150 is 0.832 at $n=10$ and 0.991 at 150 Computed . The base 3 is the inverse of the coefficient of the action $\frac{1}{3|g|}$ for passing through the wall of $V=\frac{1}{2}x^2-|g|x^4$ in imaginary time.

04

Optimal truncation — stopping where the terms are smallest

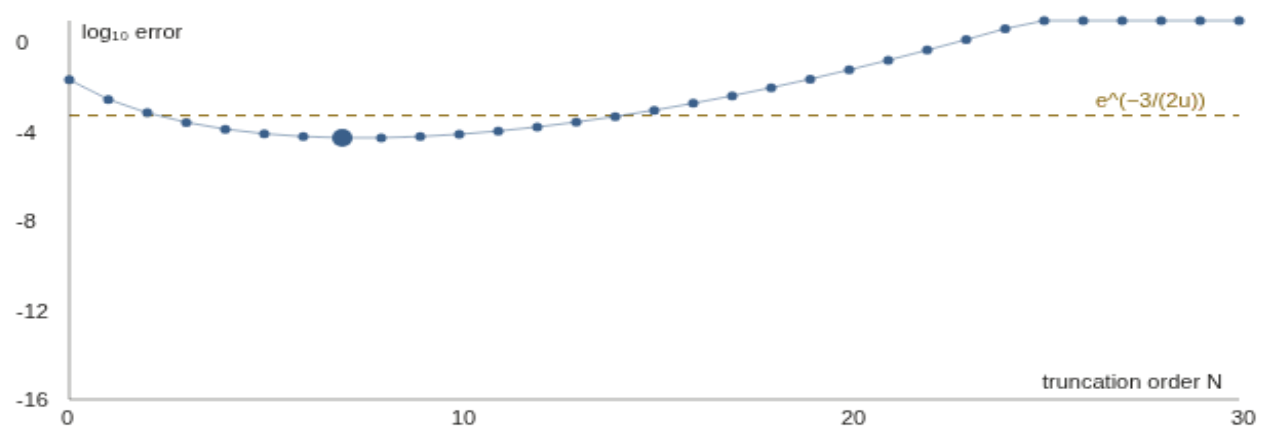
The naive method is to **stop where the terms are smallest**. The terms $\Gamma(n)(2u/3)^n$ of (6) decrease up to $n \approx 3/(2u)$ and grow beyond it. Putting Stirling's formula $\Gamma(n) \approx \sqrt{2\pi/n}\,(n/e)^n$ in there gives

$$\min_n |a_n u^n| \;\approx\; \frac{1}{\pi\sqrt{2}}\sqrt{\frac{2\pi}{N^*}}\,e^{-N^*} \;\propto\; \sqrt{u}\,e^{-3/(2u)}, \qquad N^* \approx \frac{3}{2u} \quad (7)$$

The floor of the error is the size of **an exponentially small quantity fixed by the height of the wall**, and the series can say nothing finer than that. The "part not shown in the coefficients" of section 01 appears as a limit on precision (the ratio on the right moves because of $\sqrt{u}$).

Why the error is about the smallest term. In this model, by (14) of 4-03 § 08, the error of stopping after N terms does not exceed the first discarded term $|a_N|u^N$ Known . Stopping where the terms are smallest, the error is at most that smallest term. Since the signs of the coefficients alternate, the actual error settles at about half the smallest term (0.47 to 0.50 times in the 6 rows of Table 3 Computed). Thus "the best error ≈ the smallest term ≈ (7)". **How to read Table 3.** Each row fixes one u, compares the partial sums from order 0 to order N ($N<60$) with direct numerical integration $Z(u)$, and shows the N with the smallest error and that error. The best N comes around $3/(2u)$. The rightmost column is the error divided by $e^{-3/(2u)}$, and stays within 0.07 to 0.26 as u varies from 0.1 to 2. That is, the error becomes small at the same rate as $e^{-3/(2u)}$.

Figure 1 Changing u moves the bottom of the valley of the truncation error $|S_N(u)-Z(u)|$ (logarithmic scale) parallel to the dashed line $e^{-3/(2u)}$ (about one order of magnitude below it)



u = 0.20, Z(u) = 0.97791537, best N = 7, smallest error 5.52e−5, e^(−3/(2u)) = 5.53e−4

Computed The numbers in this figure were computed on this machine (Simpson's rule and the recurrence for the coefficients; agreeing with Table 3).

Computed Table 3 Optimal truncation ($Z(u)$ is direct numerical integration. The Borel sum (8) of section 05 agrees in all 6 rows to within 5×10^{-16}, within double-precision rounding)

u	Best N	Smallest error	Error $/e^{-3/(2u)}$
0.1	15	2.18e−8	0.071
0.2	7	5.52e−5	0.100
0.3	5	7.96e−4	0.118
0.5	3	7.31e−3	0.147
1.0	1	4.39e−2	0.197
2.0	1	1.22e−1	0.259

05

The Borel sum — divide by $n!$, restore with an integral

The **Borel sum** first divides by $n!$ to make the series converge, then multiplies back with $\int_0^\infty e^{-t} t^n \, dt = n!$.

$$B(t) = \sum_{n \geq 0} \frac{a_n}{n!} t^n, \qquad \tilde{Z}(u) = \frac{1}{u} \int_0^\infty e^{-t/u} \, B(t) \, dt \tag{8}$$

B is called the **Borel transform**, and the complex plane of t the **Borel plane**. Term by term $\frac{1}{u} \int_0^\infty e^{-t/u} \frac{t^n}{n!} dt = u^n$, so the asymptotic series of $\tilde{Z}$ is the original series. For the zero-dimensional model, from $(4n-1)!! = 16^n (\frac{1}{4})_n (\frac{3}{4})_n$ ($(a)_n = a(a+1)\cdots(a+n-1)$),

$$B(t) = \sum_{n \geq 0} \frac{(\frac{1}{4})_n (\frac{3}{4})_n}{(n!)^2} \left(-\frac{2t}{3}\right)^n = {}_2F_1\left(\tfrac{1}{4}, \tfrac{3}{4}; 1; -\tfrac{2t}{3}\right) \tag{9}$$

The hypergeometric function ${}_2F_1(a,b;c;z)$ extends to the plane minus a cut running from $z=1$ Known . **The only singularity in the Borel plane is the single point $t=-3/2$, at the same position as the coefficient of the wall height**. The path of integration $t\geq 0$ does not touch the singularity, and (8) returns to $Z(u)$ as Table 3 shows Computed .

The theorem guaranteeing that the Borel sum agrees with the original function requires analyticity of the original function in a disc tangent at the origin, and a uniform estimate bounding the K_N of (1) by $A\sigma^N N!$ (A,σ constants independent of N). (1) alone is not enough. Beneke cites this as the Watson–Nevanlinna–Sokal theorem (Sokal 1980), and writes 'These requirements are too strong for renormalizable theories' Known .

06

When there is a singularity on the positive axis — ambiguity and exponentially small terms

For a series whose terms have the same sign and grow as $a_n\approx K\,a^n n!$ ($a>0$), the singularity comes at $t=1/a$ on the positive axis, on the path of (8). Going round above or below gives two sums **whose imaginary parts disagree by an amount of size** $e^{-1/(au)}$. We look at the simplest case, $a_n=n!\,a^n$. The Borel transform is $B(t)=\sum_n (at)^n=\frac{1}{1-at}$, with a pole at $t=1/a$. Subtracting the integral going round below from the one going round above gives an integral once clockwise round the pole, and from the residue

$$\tilde{Z}_{\text{above}}(u)-\tilde{Z}_{\text{below}}(u)=\frac{2\pi i}{au}\,e^{-1/(au)}$$

(the imaginary part of each is $\pm\pi e^{-1/(au)}/(au)$). This difference is of a size that shows up in no power of u. Tables and figures are in 7-12 § 08. Beneke shows, in examples where the exact function is known, that the two ambiguities, of the Borel integral and of the exponentially small terms, cancel each other, and sets up the following chain Known .

factorial divergence with equal signs $\Longrightarrow$ ambiguity of the Borel integral $\Longrightarrow$ add exponen

He then warns 'it is not rigorous'. It is not a theorem Physics .

Example. The gap of the double well of 2-05 § 07 has size e^{-S_0} and appears at no order of perturbation theory. A classical solution that moves in imaginary time from one valley to the other is called an **instanton**, and its action is the S_0 of (9) in 2-05. A solution moving the other way is an **anti-instanton** (for the counterpart in Yang–Mills,

see 3-05). The ambiguity of the Borel sum of the series around a valley has size e^{-2S_0} (the weight of an instanton–anti-instanton pair), and is said to cancel against the ambiguity of the pair's contribution (Bogomolny 1980, Zinn-Justin 1981) Physics . The way the series diverges announces the existence of terms the series does not show.

07

Renormalons — $n!$ from the ends of momentum integrals

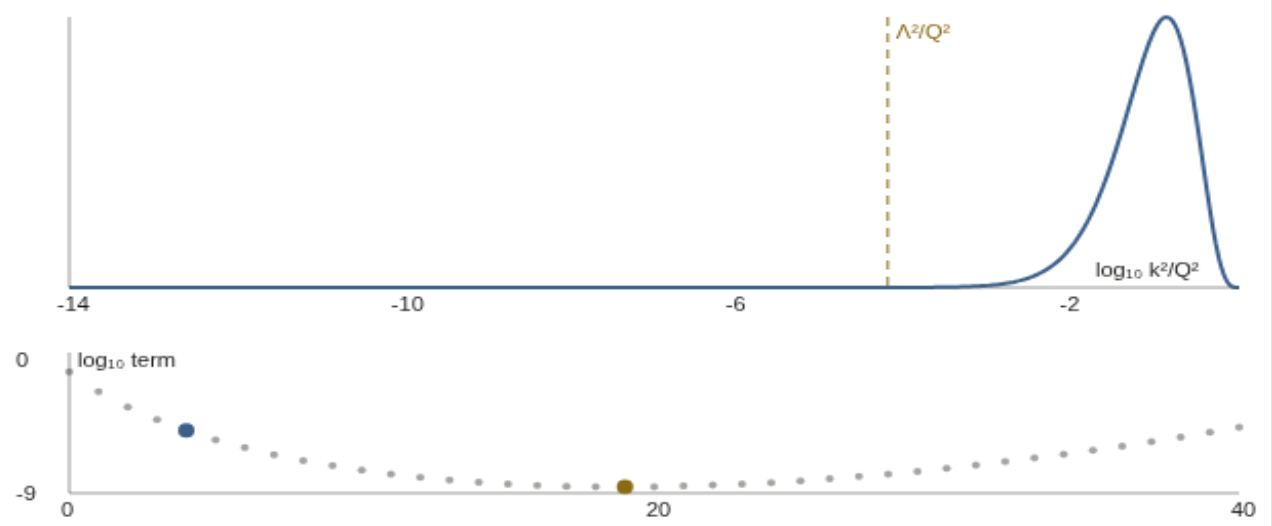
Renormalisable theories have a source of $n!$ other than saddle points: **renormalons** (according to Beneke, found by Gross–Neveu 1974, Lautrup 1977 and 't Hooft 1977, with the name coming from 't Hooft's Erice lectures Known). The source is the **bubble chain** diagram: a propagator with n one-loop bubbles (loops in which a pair of particles is created and annihilated) strung on it; summed over n, it becomes the one-loop running coupling $\alpha(k)$ inserted into a single line (Beneke § 2.2) Known . It is a different thing from the vacuum bubbles of 4-03 § 03.

We use a model that shrinks this to one variable. Suppose there is a quantity measured by injecting a momentum Q from outside, and the integral over the loop momentum k inside it carries the running coupling $\alpha(k) = \alpha/(1 - b\alpha \ln(Q^2/k^2))$ ($\alpha = \alpha(Q)$, $b = \beta_0/4\pi$, $\beta_0 = 11N/3$ for pure $SU(N)$; for conversion between the four ways of writing the coefficient, Table 1 of 4-06; 4-04 § 05, 4-06). Expanding $\alpha(k)$ in α, the order $n+1$ contains the n-th power of $\ln(Q^2/k^2)$. About the rest of the integrand we assume only that, measured with $d\ln k^2$, it vanishes in proportion to k^4 at small k (x^2) and decreases as $1/k^2$ at large k (x^{-1}). With $x = k^2/Q^2$,

$$r_n^{\mathrm{IR}} = \int_0^1 x\,\big(b\ln\tfrac{1}{x}\big)^n\,dx = \frac{b^n\,n!}{2^{n+1}}, \qquad r_n^{\mathrm{UV}} = \int_1^\infty \frac{dx}{x^2}\big(-b\ln x\big)^n = (-b)^n\,n! \tag{11}$$

each term is finite, yet the coefficients grow as $n!$ (the closed form agrees with numerical integration to 6 digits Computed). The reason is the position of the peak. The infrared integrand (measured in $\ln x$, $x^2(\ln\frac{1}{x})^n$) peaks at $x = e^{-n/2}$, and **as the order rises the peak runs away to smaller momenta**. The ultraviolet peak is at $x = e^n$. (2.24) and (2.25) of Beneke give the same positions for the bubble chain diagram Known . On the infrared side the signs are the same (positive axis, with ambiguity), and on the ultraviolet side they alternate. Changing the power at small k to k^a, the infrared renormalon sits at $u = a/2$ on the scale of § 08, and the size of the ambiguity is $(\Lambda/Q)^a$ (Beneke § 2.2) Known . This model has $a = 4$.

Figure 2 Raising n moves the infrared peak of (11) (top panel, horizontal axis $\log_{10} k^2/Q^2$) to the left, and the term $r_n\alpha^{n+1}$ (bottom panel, logarithmic) becomes smallest about where it crosses the dashed line Λ^2/Q^2. Λ moves with α




```
n = 4, peak k²/Q² = 1.35e−1, Λ²/Q² = 6.24e−5 (N = 3), term = 1.01e−5, smallest term at n* = 19
```

Computed The numbers in this figure were computed on this machine (the formula of (11) evaluated inside the figure; agreeing with Table 4).

Where does the peak stop? At one loop $\Lambda^2/Q^2 = e^{-1/(b\alpha)}$ (4-07). The term $r_n^{\mathrm{IR}}\alpha^{n+1}$ is smallest at $n^* \approx 2/(b\alpha)$, and then the peak is at $e^{-n^*/2} \approx \Lambda^2/Q^2$. **The order at which the series stops being useful is the order at which the loop momentum reaches Λ,** and the smallest term has the size of $(\Lambda/Q)^4$ times a factor of the order of a power of α.

Computed Table 4 The order of the smallest term and the position of the peak in the model for pure SU(3) ($b = 11/4\pi$)

$\alpha(Q)$	Λ^2/Q^2	n^*	Peak $/(\Lambda^2/Q^2)$	Smallest term	Smallest term $/(\Lambda/Q)^4$
0.300	2.22e−2	7	1.36	5.09e−4	1.03
0.200	3.31e−3	11	1.24	9.23e−6	0.84
0.118	6.24e−5	19	1.20	2.53e−9	0.65
0.080	6.29e−7	28	1.32	2.11e−13	0.54

Beneke notes that the positions of renormalons do not depend on the higher coefficients of the β function, and are in general logically separate from the Landau pole Known .

08

The Borel plane of Yang–Mills — a map of the expected singularities

Transforming the series in α in the manner of (8) and measuring positions by $u = \beta_0 t/4\pi$, $e^{-t/\alpha} = (\Lambda^2/Q^2)^u$, so **a singularity at position u corresponds to an ambiguity of size $(\Lambda/Q)^{2u}$**. We prepare two terms. The **operator product expansion** is a physical prescription that expands a quantity measured at short distance $1/Q$ as a sum of "coefficient $\times$ $(\Lambda/Q)^d$ $\times$ the vacuum expectation value of a combination of fields of dimension d" Physics . In pure Yang–Mills, the gauge-invariant combination of fields of

lowest dimension is $F^a_{\mu\nu}F^a_{\mu\nu}$, of dimension 4, and its expectation value $\langle F^2\rangle$ is called the **gluon condensate**. There is no gauge-invariant combination of dimension 2 ($A_\mu A_\mu$ has dimension 2 but changes under gauge transformations).

For the two-point correlation functions of QCD that Beneke summarises (quantities measured by injecting a momentum Q from outside; an example is the correlation of quark currents) Physics , the infrared renormalons are at $u = 2, 3, \ldots$ (there is none at $u = 1$ because the operator product expansion has no term of dimension 2, Parisi 1979; $u = 2$ corresponds to the $(\Lambda/Q)^4$ of the gluon condensate, and the weight k^4 of the model of § 07 is this $u = 2$), the ultraviolet renormalons at $u = -1, -2, \ldots$, and the instanton–anti-instanton pairs at $t = 4\pi n$ (Bogomolny–Fateyev 1977). At which u they are present depends on the observable. Converting to $t = 4\pi u/\beta_0$ with the β_0 of pure $SU(N)$, the renormalons move towards the origin in proportion to $1/N$, and the pairs stay at 4π. For $N = 3$, $u = 2$ is at $t = 2.285$ and the pair 5.5 times farther; for $N = 10$, 18 times Computed .

The weight of a pair $e^{-4\pi/\alpha} = e^{-16\pi^2/g^2}$ is $(\Lambda/Q)^{22N/3}$, a far smaller quantity than the $(\Lambda/Q)^4$ of $u = 2$ (for a single one, (14) of 4-07 § 08). Beneke also expects that in QCD the instanton singularities are far away and do not determine the large orders Physics .

Observation on the lattice. Bali–Bauer–Pineda (2014) computed numerically the perturbative coefficients of the average plaquette of pure SU(3) up to order 35 in α, and write 'The on-set of the dominance of the dimension four renormalon associated to the gluon condensate is clearly observed' (abstract) Physics . The PDG review also says that the renormalon divergence is 'not usually visible' at the orders available, but is 'closely connected' with non-perturbative contributions.

09

Seen from mathematics — what is a theorem and what is an expectation

On the side of theorems. (4) to (6) and (9) for the zero-dimensional model are checkable mathematics Known . There are papers with Borel summability in their titles, for the anharmonic oscillator (Graffi–Grecchi–Simon 1970) and for $P(\phi)_2$ and ϕ^4_3 (Eckmann–Magnen–Sénéor 1975, Magnen–Sénéor 1977; 7-03) (bibliographic data only). According to their titles, they show Borel summability for the eigenvalues of the oscillator and for the Schwinger functions (correlation functions) of $P(\phi)_2$ and ϕ^4_3 respectively. In those cases, by the Watson–Nevanlinna–Sokal theorem of § 05, the series determines the value uniquely.

On the side of expectations. For four-dimensional Yang–Mills the coefficients are well defined and finite at each order (4-05), but the relation between the series and the

theory has not been shown. The non-perturbative theory that the question presupposes was built by the construction published in 2026-11 (at the level of the committee's assessment; 6-00), but neither agreement with perturbation theory at all orders nor that the series is an asymptotic expansion of that theory has been shown (6-04). Before the published construction, Beneke (1999) wrote 'Since there is no rigorous non-perturbative definition of R in theories such as QED and QCD, we cannot even ask whether series expansions are asymptotic' Known . The problem statement (Jaffe–Witten § 6.1) also writes that in most known examples perturbation series are 'divergent expansions' and 'even Borel and other resummation methods have limited applicability' Known .

To sum up: (i) the $n!$ in the coefficients is the shadow of quantities of size e^{-c/g^2} that the series does not show Physics . Precisely, in the model, in the sense of (4) to (6), the exponent c of the imaginary part at negative coupling fixes the base $1/c$ of the coefficients. For Yang–Mills it is an expectation at the level of the chain (10), and ultraviolet renormalons (with alternating signs) carry no ambiguity of exponentially small terms. (ii) If the gap is proportional to $\Lambda \propto e^{-8\pi^2/(\beta_0 g^2)}$, it appears at no order of the series (4-07 § 05), and the only clue left is on the side of how the series diverges Physics . (iii) The formalism that assembles the unseen terms from the way the series diverges (transseries) is in 7-12 § 09, and for pure Yang–Mills on $\mathbb{R}^4$ it is not established even at the level of physics.

Outside the scope of this chapter

For obstacles other than divergence, see 7-11; for the planar theory, in which the number of diagrams grows only exponentially, 17-04.

What this chapter can say

	Content
What is said	(4) to (6) and (9) for the zero-dimensional model and the agreement of the Borel sum (Tables 1 to 3). The coefficients of the oscillator to order 150. The closed form of (11), and that the smallest term comes at the order at which the peak reaches Λ (Table 4) Computed . The conversion of positions in the Borel plane
Conditional	The chain (10) Physics . The cancellation in the double well is taken from the literature (text not checked). The positions of renormalons are expectations based on the bubble-chain approximation and the operator product expansion. $u = 2$ for the plaquette is a numerical observation

	Content
What is not said	The divergence, asymptoticity and the reality of renormalons for the series of four-dimensional Yang–Mills are none of them theorems. Dyson's argument is heuristic. This site shows neither the mass gap nor the construction of the continuum theory

Sources and reproduction

Item	Kind	Source / tools
Beneke's statements and quotations in §03 and §05 to §09	Known	M. Beneke, Phys. Rept. 317 (1999) 1, hep-ph/9807443 (text §1, §2)
Quotations from the problem statement and the PDG	Known	Jaffe–Witten (the Clay problem statement) §6.1; J. Huston, K. Rabbertz, G. Zanderighi, '9. Quantum Chromodynamics' (revised August 2023), in S. Navas et al. (Particle Data Group), Phys. Rev. D 110 (2024) 030001, footnote 1 (text of both)
Dyson, the anharmonic oscillator	Known	Dyson, Phys. Rev. 85 (1952) 631; Bender–Wu, Phys. Rev. 184 (1969) 1231, PRD 7 (1973) 1620 (abstracts of all the above)
The lattice plaquette	Physics	Bali–Bauer–Pineda, PRD 89 (2014) 054505, arXiv:1401.7999 (abstract)
Others	Bibliographic data only	Lipatov, JETP 45 (1977) 216; 't Hooft, Erice 1977 lectures (Plenum, 1978); Gross–Neveu, PRD 10 (1974) 3235; Lautrup, PLB 69 (1977) 109; Parisi, NPB 150 (1979) 163; Bogomolny–Fateyev, PLB 71 (1977) 93; Bogomolny, PLB 91 (1980) 431; Zinn-Justin, NPB 192 (1981) 125; Sokal, JMP 21 (1980) 261; Graffi–Grecchi–Simon, PLB 32 (1970) 631; Eckmann–Magnen–Sénéor, CMP 39 (1975) 251; Magnen–Sénéor, CMP 56 (1977) 237
Tables 1 to 3, Figure 1, the check of the dispersion relation, the ratio to the smallest term in §04	Computed	`zero_dim.py` (SciPy, rationals), `min_term.py`. Figure 1 uses Simpson's rule in JavaScript
The coefficients of the anharmonic oscillator to order 150	Computed	`oscillator.py` (recurrence with rationals; the first 4 terms agree with the known values)
(11), Table 4, Figure 2, the positions in §08	Computed	`renormalon.py`. The figure evaluates the same formula in JavaScript

Next chapter to read: 5-01 Wilson's lattice. The detour continues in 7-12.

Wilson's lattice — put group elements on the edges, and gauge invariance survives without approximation

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — Group elements (link variables) are placed on the edges of the lattice, and the action is built from products around small squares (plaquettes). That action is exactly gauge invariant whatever the lattice spacing, approaches the continuum action at the rate a^2 for smooth fields, and this fixes $\beta = 2N/g^2$. The measure is defined without gauge fixing.

Prerequisite chapters — 3-03 (the action) and 1-04 (SU(2), SU(3)). Conventions follow 3-02, parallel transport 1-07, Haar measure 1-08 § 03.

Said first — This is a chapter of definitions and classical properties. The continuum limit of § 05 is how the action approaches its limit for smooth fields, a different thing from the continuum limit of the quantum theory (5-05). This chapter shows nothing about the mass gap.

Order of this chapter

1. Why a lattice — to a finite-dimensional integral. Detour: the summer of 1973
2. Link variables — group elements on the edges (Table 1)
3. Plaquettes and Wilson loops — only closed paths survive (Figure 1, Table 2)
4. The Wilson action — five properties
5. The classical continuum limit — where β comes from (Figure 2, Table 3), converting conventions (Table 4)
6. The measure on the lattice — no gauge fixing needed
7. What survives, what breaks — choosing a regularisation
8. The map of Part 5 — two limits

01

Why a lattice — making the integral finite-dimensional

Euclideanised quantum Yang–Mills theory (4-02) formally integrates $e^{-S[A]}$ over all connections A. But there is no uniform measure on an infinite-dimensional space (1-09 § 02), and the integral counts multiple times the A related by gauge transformations (3-04, 4-08). Wilson replaced spacetime by a lattice. This is the abstract of the 1974 paper.

> "It is shown how to quantize a gauge field theory on a discrete lattice in Euclidean space-time, preserving exact gauge invariance and treating the gauge fields as angular variables (which makes a gauge-fixing term unnecessary)."

"Angular variables" means holding the field as group elements rather than as elements of the Lie algebra. The problem statement (Jaffe–Witten § 6.5) writes that, for compact groups and compactified spacetime, "the lattice approximation reduces the functional integration to a finite-dimensional integral", and continues that one must check that the expectation values of gauge-invariant observables exist in the limit of lattice spacing 0 and infinite volume Known . We look at the limits in § 08.

Conventions. Euclidean $\mathbb{R}^4$, indices 1 to 4, $\hbar = c = 1$. As in 3-02, $A_\mu = A_\mu^a T^a$ (Hermitian, $\mathrm{tr}\, T^a T^b = \frac{1}{2}\delta^{ab}$), $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - ig[A_\mu, A_\nu]$, and the continuum action is $S = \frac{1}{2}\int \mathrm{tr}\, F_{\mu\nu} F_{\mu\nu}\, d^4x$ (summed over all indices). β is the inverse lattice coupling (a different quantity from the β function $\beta(g)$ of 4-06), and θ is just an angle (not the θ term).

Detour: the summer of 1973

A subsection you may skip.

In Wilson's recollection (published 2005), after the asymptotic freedom papers (published June 1973) he came to QCD from the lattice familiar from statistical mechanics, and in the summer of 1973 he gave a talk at Orsay (he does not remember the talk himself, and writes that he knows of it from a reference in a paper of Balian–Drouffe–Itzykson). He writes that by then he knew that the gauge variables "must be unitary matrices from the gauge group itself, rather than being variables from the infinitesimal Lie algebra of the group", that there is a separate gauge invariance at each point, and that finite results come out without gauge fixing. He also mentions that, since the coefficient is $1/g^2$, the strong-coupling expansion (5-02) was immediately visible, that J. Smit was developing a formulation of lattice gauge theory, that Polyakov had obtained the same thing independently, and the model of Wegner (1971) (2-09) Known .

02

Link variables — group elements on the edges

On the lattice $a\mathbb{Z}^4$ of spacing a (in finite volume, a periodic lattice with L points per side), the edge joining x and $x + a\hat{\mu}$ is called a link. In the continuum theory the gauge field was a tool for comparing frames at neighbouring points (3-01). On the lattice the tool itself

becomes the variable: on each link we put $U_{x,\mu} \in G$ and read it as "carrying the value at $x + a\hat{\mu}$ into the frame of x". A gauge transformation, by $\Omega(x) \in G$ at each point, acts as follows.

$$U_{x+a\hat{\mu},\,-\mu} = U_{x,\mu}^{-1}, \qquad U'_{x,\mu} = \Omega(x)\, U_{x,\mu}\, \Omega(x + a\hat{\mu})^{-1} \quad (1)$$

If a matter field transforms as $\psi \mapsto \Omega\psi$, the difference $U_{x,\mu}\psi(x + a\hat{\mu}) - \psi(x)$ changes only by $\Omega(x)$ (the four-direction version of the chain of 3-02). From a smooth A, the link is the inverse of parallel transport (1-07) along the edge. Divide the edge into n equal parts and multiply the small rotations from the left, starting from the initial end.

$$U_{x,\mu} = \lim_{n\to\infty} \prod_{k=0}^{n-1} e^{-igh A_\mu(x+(k+\frac{1}{2})h\hat{\mu})} = \exp\bigl(-iga\, A_\mu(x + \tfrac{a}{2}\hat{\mu}) + O(a^3)\bigr), \qquad h = \tfrac{a}{n} \quad (2)$$

(1) is the transformation law $A' = \Omega A \Omega^{-1} - \frac{i}{g}(\partial\Omega)\Omega^{-1}$ of 3-02 accumulated along the edge Known . Example: for U(1), $U = e^{-i\theta}$, and if $\Omega(x) = e^{i\alpha(x)}$, (1) is $\theta \mapsto \theta + \alpha(x + a\hat{\mu}) - \alpha(x)$ (2-03 § 03). For SU(2), $U = u_0 + i\,\mathbf{u}\cdot\boldsymbol{\sigma}$ ($u_0^2 + |\mathbf{u}|^2 = 1$), each link is a point of S^3, and $\frac{1}{2}\mathrm{tr}\, U = u_0$.

A periodic lattice has $4L^4$ links, and the configuration space G^{4L^4} is a compact manifold of real dimension $4(N^2-1)L^4$ (1-05). The gauge freedom G^{L^4} is a quarter of it, the same proportion as removing 1 gauge component from the 4 components in the continuum.

Table 1 Sizes of periodic lattices Computed

L	Links	Plaquettes	Real variables for SU(3)
4	1,024	1,536	8,192
16	262,144	393,216	2,097,152
32	4,194,304	6,291,456	33,554,432

03

Plaquettes and Wilson loops — only closed paths survive

A single link changes sandwiched between the Ω at its two ends, but when two links are joined end to end, the $\Omega^{-1}\Omega$ at the joining point cancels. Going round a closed path, only the Ω at the starting point remains, on both sides. The smallest closed path is a square of side a: the plaquette.

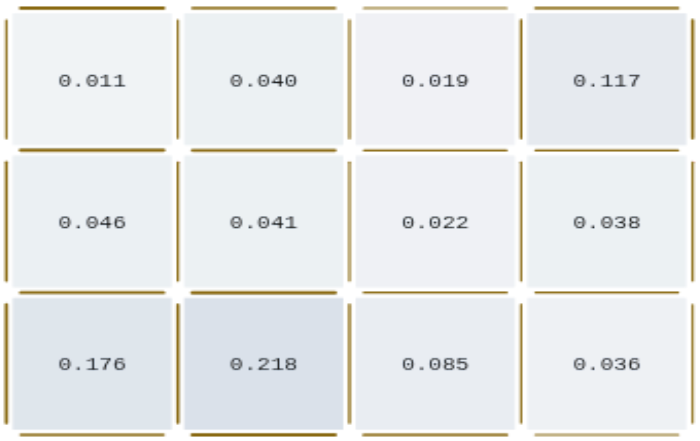
$$U_P = U_{x,\mu}\, U_{x+a\hat{\mu},\nu}\, U_{x+a\hat{\nu},\mu}^{-1}\, U_{x,\nu}^{-1}, \qquad U'_P = \Omega(x)\, U_P\, \Omega(x)^{-1} \quad (3)$$

So $\mathrm{tr}\, U_P$ is invariant without approximation, whatever the lattice spacing and however rough the configuration. For a general closed path C, the trace of the product taken in the order of the path,

$$W_C = \frac{1}{N}\,\mathrm{tr} \prod_{\ell\in C} U_\ell \quad (4)$$

is called a Wilson loop. Made from a smooth field via (2), it approaches the continuum Wilson loop (the trace of the holonomy of 1-07) as the lattice path approaches a closed curve (Chatterjee 2018 § 4) Known . Rectangular loops are used to test for confinement (5-02, 17-02). The order of multiplication matters. Multiplying the same four links in the order $U_{x,\mu}U_{x+a\hat{\nu},\mu}U_{x+a\hat{\mu},\nu}^{-1}U_{x,\nu}^{-1}$, factors such as $\Omega(x+a\hat{\nu})^{-1}\Omega(x+a\hat{\mu})$ get stuck between them and do not cancel. For U(1) the order does not matter (2-03).

Figure 1 A 5×4-point patch of SU(2). Moving the slider s (the strength of the gauge transformation) changes the thickness of the edges ($1-\frac{1}{2}\mathrm{tr}\,U_\ell$) but not the numbers on the faces ($1-\frac{1}{2}\mathrm{tr}\,U_P$). "Break the order" makes the face numbers move too.




```
s = 0.00, Σ_P(1−½trU_P) = 0.849977, Σ_ℓ ½trU_ℓ = 27.0024
```

Computed The numbers in this figure were computed on this machine. The browser computes the same configuration with the same formulas as Table 2 (if JavaScript is disabled, Table 2 gives the content).

Table 2 Sums over the configuration of Figure 1 (31 edges, 12 faces)

s	$\sum_\ell \frac{1}{2}\mathrm{tr}\,U_\ell$	$\sum_P(1-\frac{1}{2}\mathrm{tr}\,U_P)$	$\sum_P(1-\frac{1}{2}\mathrm{tr})$ **in broken order**
0	27.0024	0.849977	0.5719
0.5	17.5284	0.849977	3.0662
1	0.6552	0.849977	5.1220

Computed Applying a random gauge transformation to a random configuration on a four-dimensional periodic lattice with 4 points per side, the change in the action of § 04 was 2×10^{-13} for SU(2) and 0 for SU(3). Building the sum of squares of the differences $\partial_\mu A_\nu-\partial_\nu A_\mu$ from the logarithms of the links (elements of the Lie algebra), for SU(2) 34,707 changes to 35,659. Even with the commutator added, for finite gauge transformations the transformation law of A is inhomogeneous (the term $-\frac{i}{g}(\partial\Omega)\Omega^{-1}$ below (1)), so the differences are not exactly invariant. This is the meaning of "group elements rather than Lie algebra variables" in the recollection in the detour.

> 04

The Wilson action — five properties

$$S_W(U) = \beta \sum_P \Big(1 - \frac{1}{N} \operatorname{Re}\operatorname{tr} U_P\Big), \qquad \beta = \frac{2N}{g^2} \quad (G = SU(N)) \tag{5}$$

The sum takes every plaquette once ($6L^4$ of them on a periodic lattice). There are five properties Known .

1. **Gauge invariant**: each term is unchanged by (3).
2. **Real**: reversing the orientation leaves $\operatorname{Re}\operatorname{tr} U_P^\dagger$ the same.
3. **Bounded**: the eigenvalues have absolute value 1, so each term is between 0 and 2. The maximum is 2 for SU(2) ($U_P = -1$) and $3/2$ for SU(3) (all three eigenvalues $e^{2\pi i/3}$, or all three $e^{-2\pi i/3}$) Computed .
4. **Minimum**: $S_W = 0$ only when all $U_P = 1$. A pure gauge $\Omega(x)\Omega(x+a\hat{\mu})^{-1}$ is an example. On a periodic lattice, so is a configuration with commuting elements V_μ placed uniformly in each direction. If its product once round, V_μ^L, is not 1, it is not a pure gauge, because for a pure-gauge configuration the product once round is $\Omega(x)\Omega(x)^{-1} = 1$.
5. **Local**: each term depends on four links only.

Chatterjee (2018) writes $\beta' \sum_P \operatorname{Re}\operatorname{tr}(I - U_P)$, so $\beta' = \beta/N$ (the same holds for the β of SU(2) in 7-02).

05

The classical continuum limit — where $\beta = 2N/g^2$ comes from

Map a smooth A to links via (2) and look at $a \to 0$. Expanding (3) with the Baker–Campbell–Hausdorff formula $e^X e^Y = e^{X+Y+\frac{1}{2}[X,Y]+\cdots}$, which combines a product of exponentials into one, gives at the centre c of the plaquette

$$U_P = \exp\big(-iga^2 F_{\mu\nu}(c) + O(a^3)\big) \tag{6}$$

The difference of A on opposite edges gives $-iga^2(\partial_\mu A_\nu - \partial_\nu A_\mu)$, and the commutator gives $(-iga)^2[A_\mu, A_\nu]$; together they make the field strength (curvature = the shift after going once round a small loop; 1-07). Correspondence with the notation of 1-07: the anti-Hermitian connection and curvature of 1-07 correspond to $-igA$ and $-igF$ of this chapter, and parallel transport in 1-07 carries from the start to the end, so the links and plaquettes of this chapter are its inverses. This is why the sign in the exponent of (6) is opposite to 1-07's $U_p \approx \exp(-a^2 F)$; the value of the trace does not change. For Hermitian X, $\operatorname{Re}\operatorname{tr} e^{-iX} = \operatorname{tr}\cos X = N - \frac{1}{2}\operatorname{tr} X^2 + \cdots$, so

$$1 - \frac{1}{N}\operatorname{Re}\operatorname{tr} U_P = \frac{g^2 a^4}{2N} \operatorname{tr} F_{\mu\nu}(c)^2 + O(a^5) \tag{7}$$

Taking the centre of the expansion at the centre of the plaquette, the a^5 terms cancel by the symmetry of orientations, and the remaining error is a^6 per plaquette Known . Summing a^{-4} plaquettes, the relative error is a^2, and this is the term that Symanzik improvement (5-05) removes. Counting plaquettes by a point x and a pair $\mu < \nu$ and replacing $\sum_x a^4$ by an integral,

$$SW \approx \frac{\beta g^2}{2N} \sum_{\mu<\nu} \int \operatorname{tr} F_{\mu\nu}^2 \, d^4x = \frac{\beta g^2}{2N} \, S \tag{8}$$

It agrees with the continuum action S when $\beta = 2N/g^2$, and this is where the β in (5) comes from Known . Chatterjee (2018) § 3 does the same calculation and writes that in four dimensions β is classically independent of a, that in the quantum theory physicists believe β should grow like $\log(1/a)$ though there are also doubts, and that "the question remains an open mathematical problem" Physics (5-05).

Converting conventions — five ways of writing one coupling

The same coupling g^2 is written with different letters and different normalisations in different references. This is why the same condition looks like different numbers in different chapters. The standard of this book is Wilson's $\beta = 2N/g^2$ in (5) (the table of notation A-01 § 06, where Table 7 lists six ways of writing it, adding g^2, with the same numerical examples). The other ways of writing are fixed by comparing term by term with the action (5).

$$\beta = \frac{2N}{g^2}, \qquad \beta_{\mathrm{Ch}} = \frac{\beta}{N} = \frac{2}{g^2}, \qquad \beta_{\mathrm{SZZ}} = \frac{\beta}{N^2}, \qquad \lambda = g^2 N = \frac{2N^2}{\beta} = \frac{2}{\beta_{\mathrm{SZZ}}}, \qquad \alpha = \frac{g^2}{4\pi} \tag{8a}$$

β_{Ch} is the form writing the action as $\beta_{\mathrm{Ch}} \sum_P \operatorname{Re}\operatorname{tr}(I - U_P)$ (without $1/N$; (3) of 7-10), β_{SZZ} is 't Hooft's convention writing the inverse coupling in front of the action as $N\beta_{\mathrm{SZZ}}$ (Shen–Zhu–Zhu; 5-04), λ is the 't Hooft coupling of large N (17-04), and α is the coupling strength used in physics. Differences in the constant term of the action disappear in the normalisation Z and do not change the measure.

Table 4 Converting conventions, with numerical examples (the Shen–Zhu–Zhu condition is in four dimensions, $|\beta$ $1/(16(d-1)) = 1/48$) Computed

Form	Conversion from β	SU(2), $\beta = 2.5$	SU(3), $\beta = 6.0$	SZZ condition: SU(2)	SZZ condition
Wilson's β (this book)	$2N/g^2$	2.5	6.0	$\beta < 0.0833$	β
β_{Ch} (the form without $1/N$)	β/N	1.25	2.0	< 0.0417	
β_{SZZ} ('t Hooft's convention)	β/N^2	0.625	0.6667	$< 1/48 = 0.0208$	

Form	Conversion from β	SU(2), $\beta = 2.5$	SU(3), $\beta = 6.0$	SZZ condition: SU(2)	SZZ condition
't Hooft coupling λ	$2N^2/\beta$	3.2	3.0	$\lambda > 96$	
g^2	$2N/\beta$	1.6	1.0	$g^2 > 48$	
$\alpha = g^2/4\pi$	$N/(2\pi\beta)$	0.1273	0.0796	$\alpha > 3.820$	

Example: the Shen–Zhu–Zhu condition is $1/48$ independent of N when written in β_{SZZ}, $\lambda > 96$ when written in λ, and $N^2/48$ when converted to Wilson's β. That N disappears in β_{SZZ} and λ is the advantage of 't Hooft's convention. The SU(2) value $\beta = 2.5$ often used in numerical work is $\alpha \approx 0.13$, and the edge of the condition is $\alpha \approx 3.8$; the two are about 30 times apart. The conversion is this book's arithmetic, and the conditions themselves are to be read in the normalisation of each reference (5-04, 5-09, 7-10).

Looking at a single plaquette

For the constant SU(2) field $A_1 = cT^1$, $A_2 = cT^2$ there are no derivative terms, and $F_{12} = gc^2T^3$ comes only from the commutator (Example 2 in 3-02 § 04). The links are $e^{-i\theta\sigma 1/2}$ and $e^{-i\theta\sigma 2/2}$ ($\theta = gac$), and the plaquette is the group commutator $U_1U_2U_1^{-1}U_2^{-1}$. We multiply in the form $u_0 + i\,\mathbf{u}\cdot\boldsymbol{\sigma}$ of § 02. From the product of Pauli matrices $\sigma_j\sigma_k = \delta_{jk} + i\epsilon_{jkl}\sigma_l$,

$$(a_0 + i\,\mathbf{a}\cdot\boldsymbol{\sigma})(b_0 + i\,\mathbf{b}\cdot\boldsymbol{\sigma}) = a_0b_0 - \mathbf{a}\cdot\mathbf{b} + i\,(a_0\mathbf{b} + b_0\mathbf{a} - \mathbf{a}\times\mathbf{b})\cdot\boldsymbol{\sigma}$$

Putting $c = \cos\frac{\theta}{2}$ and $s = \sin\frac{\theta}{2}$, $U_1 = (c;\,-s,0,0)$ and $U_2 = (c;\,0,-s,0)$ (writing u_0 and $\mathbf{u}$ side by side), $U_1U_2 = (c^2;\,-cs,-cs,-s^2)$ and $U_1^{-1}U_2^{-1} = (U_2U_1)^{-1} = (c^2;\,cs,cs,-s^2)$. The u_0 of the product of the two is $\frac{1}{2}\mathrm{tr}\,U_P$, so

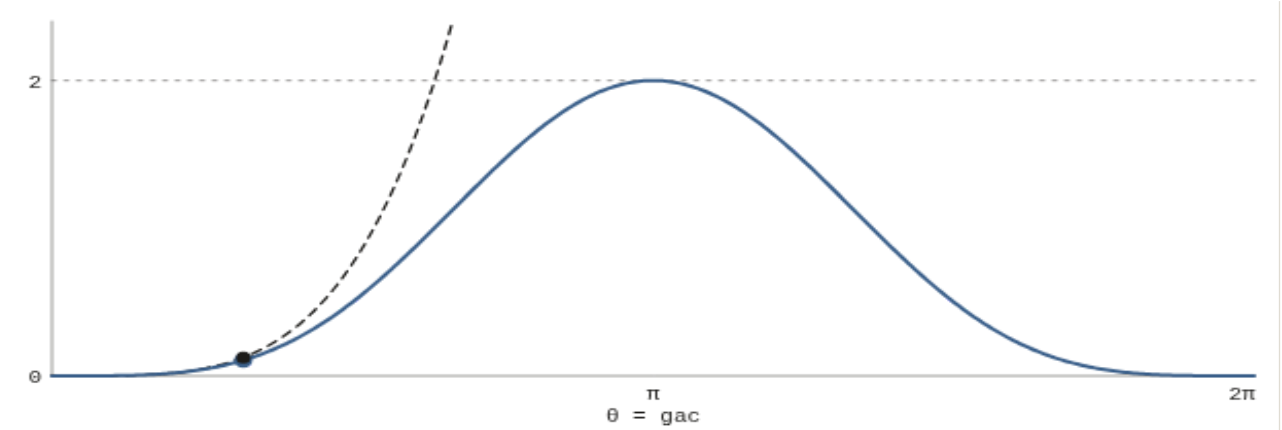
$$\tfrac{1}{2}\mathrm{tr}\,U_P = c^4 - \left(-c^2s^2 - c^2s^2 + s^4\right) = (c^2+s^2)^2 - 2s^4 = 1 - 2\sin^4\frac{\theta}{2}$$

and the following closed form comes out.

$$1 - \tfrac{1}{2}\mathrm{tr}\,U_P = 2\sin^4\frac{\theta}{2}, \qquad \frac{g^2a^4}{2N}\,\mathrm{tr}\,F_{12}^2 = \frac{\theta^4}{8} \quad (9)$$

For small θ the ratio is $1 - \theta^2/6 + \cdots$, and the relative error is proportional to a^2. For large θ the lattice value does not exceed 2, and returns to 0 at $\theta = 2\pi$. The link sees the field only in the form e^{-igaA}, so it does not distinguish a field that is too strong from a weak one. This boundedness is one reason why the integral in § 06 is finite.

Figure 2 A single plaquette. The solid line is the lattice $2\sin^4(\theta/2)$, the dashed line the continuum $\theta^4/8$, the dotted line the upper bound 2. Moving the slider for $\theta = gac$ (proportional to a for a fixed field) changes the two points and the ratio.



θ = 1.00, lattice 0.10566, continuum 0.12500, ratio 0.8453

Computed The numbers in this figure were computed on this machine. The curves are (9) (differing from the matrix calculation by at most 2×10^{-16}). If JavaScript is disabled: the ratio (lattice / continuum) is 0.998, 0.959, 0.845 and 0.501 at $\theta = 0.1,\ 0.5,\ 1,\ 2$, and the lattice value is at its maximum 2 at $\theta = \pi$ and 0 at 2π.

Looking at a smooth four-dimensional field

We compare using a smooth SU(2) field on the torus $[0,1)^4$ (a sum of trigonometric functions with 8 wave numbers, $g = 1$, $\beta = 4$). The links truncate (2) at $n = 16$, and since the continuum action is a trigonometric polynomial, it was integrated exactly as a sum over finitely many points ($S = 184.2275$).

Table 3 Lattice spacing $a = 1/L$ and the Wilson action Computed

L	S_W	$(S_W - S)/S$	The same value $/a^2$
4	134.502	−0.2699	−4.319
8	170.242	−0.07592	−4.859
16	180.618	−0.01959	−5.016
32	183.318	−0.004938	−5.057

The difference is proportional to a^2, with coefficient near −5.06 for this field (the value depends on the field). Changing n to 8, 16, 32 changes S_W by 0.03 and 0.008 at $L = 8$. The $\|U_P - e^{-iga^2F(c)}\|/a^2$ of (6) is 0.190, 0.119, 0.065 and 0.034 from $a = 1/8$ to $1/64$, nearly halving each time a is halved (ratios from 1.6 to 1.9), shrinking in proportion to a. With the sign of the commutator reversed it stayed at 5 to 6.5 Computed .

06

The measure on the lattice — no gauge fixing needed

The quantum theory on a finite lattice is defined by the following probability measure (1-08 § 01).

$$d\mu_\beta(U) = \frac{1}{Z}\, e^{-S_W(U)} \prod_\ell dU_\ell, \qquad Z = \int e^{-S_W(U)} \prod_\ell dU_\ell \tag{10}$$

dU_ℓ is the Haar measure of G (the measure of total weight 1 unchanged by multiplication from either left or right; for SU(2), the uniform distribution on S^3). **It is finite**: $0 < e^{-S_W} \leq 1$ is continuous and the domain is compact, so $0 < Z \leq 1$, and the boundedness of § 04 also gives $Z \geq e^{-12\beta L^4}$ ((3) of 7-10) Known . For SU(3) with 4 points per side it is an ordinary multiple integral over 8,192 real dimensions.

No gauge fixing needed: Haar measure is unchanged by left and right multiplication, so moving a configuration by (1) changes neither $\prod dU_\ell$ nor S_W, and $\langle f \rangle = \langle f \circ \Omega \rangle$. In the continuum, gauge fixing was needed to choose representatives from orbits of infinite volume (4-08, 3-04), but on the lattice the whole of the gauge transformations G^{L^4} is compact with weight 1, so the multiple counting is merely averaged. This is what "making the gauge-fixing term unnecessary" in the abstract means Known .

There are two consequences. (i) For general f, $\langle f \rangle$ equals the expectation value of f averaged over gauge transformations. In particular, quantities that change non-trivially under gauge transformations (a single link U_ℓ or its components, $\mathrm{tr}\, U_\ell$, and so on) are cancelled by the Haar average and have expectation value 0 (the detour of 17-02; that this survives in infinite volume is Elitzur's 1975 theorem). Not every gauge-non-invariant quantity vanishes: the expectation value of $|\mathrm{tr}\, U_\ell|^2$ is positive. (ii) The links on a tree (a set of links containing no cycles) may be set to 1 (7-02 § 02). Typical observables are Wilson loops and the correlation between two distant loops. The rate at which the correlation decreases is the mass on the lattice (4-11, 5-09), and the lattice correlation-length problem in § 5 of Chatterjee (2018) (Chatterjee's Problem 5.1) asks about it.

07

What survives, what breaks — choosing a regularisation

A device for avoiding infinities (a regularisation) changes part of the theory, and what it keeps decides the later work. On Wilson's lattice, gauge invariance survives exactly at each a, and of rotations and translations only rotations by 90° and shifts by the lattice spacing survive (the recovery of the continuum symmetries is to be shown in the limit; 7-10; for the recovery in the construction published in 2026-11, 12-06). Reflection positivity holds (Osterwalder–Seiler 1978; 5-03), and topological charge (3-05) is not defined as it stands.

The problem statement (JW § 6.5) calls the fact that reflection positivity holds in Wilson's approximation “a major advantage” Known , because it is the key to building the Hamiltonian and the Hilbert space (4-10). By contrast, a continuum momentum cutoff breaks gauge invariance, which is restored with counterterms (7-07 § 06). JW § 6.5 also writes that when reflection positivity is lost with regularisations such as dimensional regularisation or Pauli–Villars, there are few ways of recovering it Known . On this basis, a regularisation that keeps all three (gauge invariance, reflection positivity, continuum rotations) has not been found as far as we have looked (this book's organisation; for the version that counts four by adding ultraviolet control, 7-10 § 03).

The abstract of Lüscher (1982) states that non-abelian gauge fields on a four-dimensional hypercubic lattice with small action density have "an integer topological charge Q, which is invariant under continuous deformations of the field" Known . This Q is said to return to the Chern number (3-05) in the classical continuum limit. (A note of this chapter: without the condition, $SU(N)^{\text{number of links}}$ is connected, so the only continuous integer-valued quantities are constants.) Nor is the action limited to Wilson's form: there are heat-kernel actions (Villain type; 7-02), and improved actions that add 1×2 rectangles to remove the a^2 difference of Table 3 (Symanzik improvement; 5-05). The view that all of them reach the same continuum limit is universality Physics .

08

The map of Part 5 — two limits

The theory on a finite lattice was defined by (10). The theory of the problem statement needs two limits, $L \to \infty$ and $a \to 0$ (in four dimensions thought to be paired with $\beta \to \infty$ Physics). The order is first to fix β (that is, a) and take $L \to \infty$, then $a \to 0$. The lattice correlation-length problem in § 5 of Chatterjee (2018) also takes the infinite-volume limit at fixed β as its starting point Known . Even for the first, Chatterjee (2018) § 3 writes that the uniqueness of the infinite-volume limit is not known in general for large β above two dimensions Known . The construction published in 2026-11 takes the opposite order: it holds the volume fixed and takes a $\to 0$, and then takes the limit along a subsequence of volumes (the committee's assessment is that it is proved in Lean; 8-05). This construction is read in Volume 2: the family of lattices and the order of limits in 8-01 to 8-05, and one step of renormalisation in 9-01 to 9-07. The rest of Part 5 shares the work. The expansion in small β is 5-02, the Hamiltonian 5-03, infinite volume on the strong-coupling side 5-04, $\beta \to \infty$ and $a(\beta)$ 5-05, numerical integration 5-07 and 5-08, and what to hold fixed in the limit 5-09. This chapter shows nothing about the existence of the two limits or their mass gap.

What this chapter can say

Item	What is said	What is not said
The definition and gauge invariance, finiteness of the measure	Exact at each lattice spacing (algebraic identities, compactness) Known . Also checked numerically Computed	Uniqueness of the infinite-volume limit (unknown in general at large β)
The classical continuum limit (6) to (8)	The formal expansion and $\beta = 2N/g^2$ Known . For one field, the difference proportional to a^2 Computed	Error estimates for general fields. The continuum limit of the quantum theory (built by the construction published in 2026-11, at the level of the committee's assessment; 6-00, 8-05)
Reflection positivity, topological charge	Presentation of claims in the literature	Proofs (5-03, 3-05)

Item	What is said	What is not said
The mass gap	—	This chapter shows nothing about it, on the lattice or in the continuum

Sources and reproduction

Item	Kind	Source / tools
The abstract of § 01 / the recollection in the detour	Primary source (1974: abstract only checked; 2005: text checked)	K. G. Wilson, “Confinement of Quarks”, Phys. Rev. D 10 (1974) 2445–2459 / “The Origins of Lattice Gauge Theory”, Nucl. Phys. B Proc. Suppl. 140 (2005) 3–19, arXiv:hep-lat/0412043
Quotations on finite-dimensional integrals, the two limits and reflection positivity (including the case where it is lost by regularisation)	Primary source (text checked)	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory” (the official Clay problem description) § 6.5
β', the calculation of the expansion, $\log(1/a)$, the approximation of Wilson loops, uniqueness, the lattice correlation-length problem	Primary source (text checked)	S. Chatterjee, “Yang–Mills for probabilists”, arXiv:1803.01950 (2018) § 3– § 5
Topological charge	Primary source (abstract checked; agreement with the Chern number not read in the text)	M. Lüscher, “Topology of lattice gauge fields”, Commun. Math. Phys. 85 (1982) 39–48
Reflection positivity, Wegner's model, Elitzur's theorem	Bibliographic data only (checked with Crossref)	K. Osterwalder, E. Seiler, Ann. Phys. 110 (1978) 440–471 / F. J. Wegner, J. Math. Phys. 12 (1971) 2259–2272 / S. Elitzur, Phys. Rev. D 12 (1975) 3978–3982
Table 4, (8a)	This book's arithmetic (conditions in the normalisation of each reference)	H. Shen, R. Zhu, X. Zhu, Comm. Math. Phys. 400 (2023) 805–851 (abstract of arXiv:2204.12737; for the reading of the normalisation, 5-04)
Tables 1 to 3, Figure 1, Figure 2, numbers in the text	Computed on this machine	Python 3.12, NumPy 1.26, SciPy 1.11. `gauge.py` (Table 1, Table 2, § 03), `plaquette.py` ((9), the 3/2 of SU(3)), `continuum.py` (Table 3, (6); about 110 seconds). The figures compute the same formulas in plain JavaScript

Next chapter to read: 5-02 The strong-coupling expansion (expanding (10) at small β and seeing the area law).

The strong-coupling expansion — at small β the Wilson loop becomes a sum over surfaces, and the area law comes out

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — At small β the expectation value of a Wilson loop becomes a sum over surfaces bounded by the loop (the character expansion); from the smallest surface come the area law $\langle W \rangle \approx u^{RT}$ and the leading term of the string tension $\sigma_0 = -\ln u$, from bumps the first correction, and from tubes the leading term of the mass $m_0 = -4 \ln u$. We set them beside SU(2) numbers and see how far they work.

Prerequisite chapters — 5-01. Representations in 1-04, Haar measure in 1-08, mass in 4-11.

Said first — The strong-coupling area law is a theorem at small β with the lattice spacing fixed, and it holds for $U(1)$ too. It concerns the side opposite to the continuum limit ($\beta \to \infty$), and is not a proof of confinement or of the mass gap. The proof of convergence is left to 5-04.

Order of this chapter

1. The world of small β — the sum over surfaces seen in $\mathbb{Z}_2$
2. The character expansion — splitting the weight of one plaquette by representation (Table 1). Detour: the derivation of (4)
3. Tiling a surface — the leading term and the area law (Figure 1). Detour: sums over surfaces and strings
4. Bumps — the first correction and the number of dimensions
5. Tubes — the leading term of the mass and the correlation length
6. Comparing with numbers — how far it works (Figure 2, Table 2)
7. Convergence — finite boxes and infinite boxes. Where it stops
8. Is the area law a proof of confinement? — it comes out for abelian groups too

01

The world of small β — the sum over surfaces seen in $\mathbb{Z}_2$

A small coefficient $\beta = 2N/g^2$ in the Wilson action (5-01 § 04) means a large coupling g, so the expansion on this side is called the **strong-coupling expansion**. β plays the role of inverse temperature, so it is also called the **high-temperature expansion** (2-08). At $\beta = 0$ the links scatter independently with Haar measure, and since $\int_G U\, dU = 0$, the expectation value of a simple loop is 0. The abstract of Wilson 1974 writes "The lattice gauge theory has a computable strong-coupling limit; in this limit the binding mechanism applies and there are no free quarks. There is unfortunately no Lorentz (or Euclidean) invariance in the strong-coupling limit." Known .

Example: $\mathbb{Z}_2$. For the group of links ± 1 (2-09) with action $\beta \sum_P (1 - U_P)$, since $U_P = \pm 1$, the weight of one plaquette splits into two terms, up to a constant.

$$e^{\beta U_P} = \cosh\beta \left(1 + t\, U_P\right), \qquad t = \tanh\beta \quad (1)$$

Opening the product gives a term $t^{|X|} \prod_{P \in X} U_P$ for each set X of plaquettes. The average of a uniform ± 1 is 1 for even powers and 0 for odd ones, so in the partition function only sets in which every link is surrounded by an even number of plaquettes (closed surfaces) survive, and in the numerator of the Wilson loop $\prod_{\ell \in C} U_\ell$ only surfaces with boundary C. The smallest surface is a flat rectangle, so

$$\langle W_{R\times T}\rangle = t^{RT}(1 + \cdots), \qquad -\ln\langle W_{R\times T}\rangle \approx RT\,(-\ln\tanh\beta) \quad (2)$$

and the expectation value decreases exponentially in the area (the **area law**). That the gap for small β in the $\mathbb{Z}_2$ of 5-03 § 08 is close to $-4\log\tanh\beta$ is the same count.

02

The character expansion — splitting the weight of one plaquette by representation

For a general group too, the weight of one plaquette $e^{\frac{\beta}{N}\mathrm{Re}\,\mathrm{tr}\,U}$ is a function unchanged by conjugation VUV^{-1} (a **class function**). To split it by representation, we prepare just two tools.

Characters. For a representation r of the group (1-04), $\chi_r(U) = \mathrm{tr}\, r(U)$ is called its **character**. The trace is unchanged by conjugation, so characters are class functions. The characters of irreducible representations are orthonormal with respect to Haar

measure dU (the uniform probability measure on the group; 1-08 § 03) (Schur's orthogonality relations; proof omitted) Known .

$$\int_G \chi_r(U)\,\chi_s(U)\,dU = \delta_{rs}$$

Class functions can be expanded in these characters. It is the generalisation of Fourier series to non-commutative groups Known .

Integrals over classes of $SU(2)$. Elements of $SU(2)$ can be written as $U = V\,\mathrm{diag}(e^{i\theta}, e^{-i\theta})\,V^{-1}$ ($0 \le \theta \le \pi$), so a class function is a function of θ alone. Its integral takes the following form (the $SU(2)$ version of Weyl's integration formula) Known .

$$\int_{SU(2)} f(U)\,dU = \frac{2}{\pi}\int_0^\pi f(\theta)\,\sin^2\theta\,d\theta$$

Then $\mathrm{tr}\,U = 2\cos\theta$, the character of the spin-j representation is $\chi_j = \sin((2j+1)\theta)/\sin\theta$, and its dimension is $2j+1$.

Character coefficients

For a general group, the weight is expanded in characters as follows.

$$e^{\frac{\beta}{N}\mathrm{Re\,tr}\,U} = c_0(\beta)\Big[1 + \sum_{r\,/=0} d_r\,u_r(\beta)\,\chi_r(U)\Big], \qquad u_r = \frac{\int e^{\frac{\beta}{N}\mathrm{Re\,tr}\,U}\,\chi_r(U)\,dU}{d_r\int e^{\frac{\beta}{N}\mathrm{Re\,tr}\,U}\,dU} \quad (3)$$

d_r is the dimension. We call u_r the **character coefficients**, and write the one for the fundamental representation simply as u (for $\mathbb{Z}_2$, $u = t$). For $SU(2)$, the class integral above can be written with modified Bessel functions I_n (Detour: the derivation of (4)):

$$u_j = \frac{I_{2j+1}(\beta)}{I_1(\beta)}, \qquad u = \frac{I_2(\beta)}{I_1(\beta)} = \frac{\beta}{4} - \frac{\beta^3}{96} + \cdots \quad (4)$$

Known (agreeing with numerical integration to 2×10^{-13} Computed). It is the same number as the ratio λ_j/λ_0 of eigenvalues of the transfer matrix of a single link in 5-03 § 06.

If $\beta > 0$, then $0 < u_r < 1$. They are positive because each term of the opened exponential is a non-negative sum of characters of tensor products (5-03 § 06). They are below 1 because $\mathrm{Re}\,\chi_r(U) \le d_r$, with equality only when $r(U)$ is the identity matrix.

At small β, u approaches β for $\mathbb{Z}_2$, $\beta/2$ for $U(1)$, $\beta/4$ for $SU(2)$, and $\beta/(2N^2)$ for $SU(N)$ ($N \ge 3$) Computed . The reason is the first-order term. Opening the exponential of (3) to first order gives $u \approx \frac{\beta}{N^2}\int(\mathrm{Re\,tr}\,U)\,\chi_F(U)\,dU$ (F is the fundamental representation),

and this integral is $1/2$ for $N \geq 3$, since $\int (\operatorname{tr} U)^2 dU = 0$, and 1 for $\mathbb{Z}_2$ and $SU(2)$, since $\operatorname{tr} U$ is real Known .

For $SU(2)$, the β of 7-02 is $\beta/2$ of this chapter (compare the w of (3.3) in 7-02 with the u of (4) in this chapter; for $U(1)$ it is the same β).

Table 1 The character coefficient $u(\beta)$ of the fundamental representation. The action is $\beta \sum_P (1 - \frac{1}{N} \operatorname{Re} \operatorname{tr} U_P)$ in every case Computed

β	$\mathbb{Z}_2$	$U(1)$	$SU(2)$	$SU(3)$
0.5	0.4621	0.2425	0.1237	0.0289
1	0.7616	0.4464	0.2402	0.0601
2	0.9640	0.6978	0.4331	0.1286

Detour: the derivation of (4)

A subsection you may skip.

The weight of one plaquette for $SU(2)$ is $e^{\frac{\beta}{2} \cdot 2 \cos\theta} = e^{\beta \cos\theta}$. The formulas used are two standard properties of the modified Bessel functions I_n Known .

$$\int_0^\pi e^{\beta \cos\theta} \cos n\theta \, d\theta = \pi I_n(\beta), \qquad I_{n-1}(\beta) - I_{n+1}(\beta) = \frac{2n}{\beta} I_n(\beta)$$

Denominator. Since $\sin^2\theta = \frac{1}{2}(1 - \cos 2\theta)$,

$$\frac{2}{\pi} \int_0^\pi e^{\beta \cos\theta} \sin^2\theta \, d\theta = I_0 - I_2 = \frac{2}{\beta} I_1.$$

Numerator. By the product-to-sum formula $\sin^2\theta \cdot \frac{\sin((2j+1)\theta)}{\sin\theta} = \sin\theta \, \sin((2j+1)\theta) = \frac{1}{2}\big[\cos 2j\theta - \cos(2j+2)\theta\big]$, so

$$\frac{2}{\pi} \int_0^\pi e^{\beta \cos\theta} \cdot \tfrac{1}{2}\big[\cos 2j\theta - \cos(2j+2)\theta\big] \, d\theta = I_{2j} - I_{2j+2} = \frac{2(2j+1)}{\beta} I_{2j+1}.$$

The u_j of (3) is the numerator divided by $d_j = 2j+1$ and by the denominator, so $u_j = I_{2j+1}/I_1$. At small β, $I_n(\beta) \approx (\beta/2)^n/n!$ gives $u = I_2/I_1 \approx \beta/4$.

03 Tiling a surface — the leading term and the area law

Multiplying (3) over all plaquettes and opening the product, each term is an assignment of "which representation to put on which plaquette". There are two rules for integrating out the links one at a time (Schur's orthogonality relations Known ; (3.1) of 7-02 is a lemma of the same idea).

$$\int_G \chi_r(U)\,dU = \delta_{r0}, \qquad \int_G \chi_r(AU)\,\chi_s(U^{-1}B)\,dU = \delta_{rs}\,\frac{\chi_r(AB)}{d_r} \quad (5)$$

The left says "if an edge lies on only one plaquette, the result is 0", and the right says "gluing two plaquettes with the same representation along a common edge makes a larger loop, with a factor $1/d_r$". What survives are terms in which the plaquettes carrying representations form a closed surface, or a surface whose boundary is the observed loop.

Example: the 1×2 loop. The observable is $\frac{1}{d}\chi(U_C)$ (d is the dimension of the fundamental representation). The leading term is the one with representations on the two plaquettes inside the loop, $(du)^2\chi(U_{P_1})\chi(U_{P_2})$. For $SU(N)$, the surface carries the conjugate representation $\bar r$ so as to pair with the χ of the observable (for $SU(2)$, $\bar r = r$). We integrate in three steps.

1. Integrate the edge shared by the two plaquettes with (5) right. The two loops join into the boundary loop $\chi(U_\partial)$ of the 1×2, with a factor $1/d$.
2. Integrate one of the six boundary edges with (5) right. The loop of the observable and the boundary loop are glued with opposite orientations, the remaining five links cancel in the form VV^{-1}, and $\chi(1)/d = d/d = 1$ remains.
3. Normalisation: the $\frac{1}{d}$ of the observable remains (the partition function is 1 in the leading term apart from powers of the constant c_0, so dividing by it changes nothing).

Altogether $(du)^2\cdot\frac{1}{d}\cdot 1\cdot\frac{1}{d} = u^2$, that is, $\langle W_{1\times 2}\rangle = u^2$. Tiling an $R\times T$ too, the powers of d cancel, and what remains is

$$\langle W_{R\times T}\rangle = u^{RT}\big(1+O(u^4)\big), \qquad \sigma_0 = -\ln u \quad (6)$$

The energy of a static pair of charges $V(R) = -\lim_{T\to\infty}\frac{1}{T}\ln\langle W_{R\times T}\rangle$ (17-02 § 01) is $\sigma_0 R$ in the leading term: the confining form that grows in proportion to the distance Known . Example: for $SU(2)$ at $\beta = 1$, $\sigma_0 = 1.426$, and the 2×2 loop is $u^4 = 0.0033$ Computed .

Figure 1 The smallest surface bounded by the loop (thick line), and bumps. The sliders R, T, β change the number of plaquettes and the weight, and "Play" changes the position and orientation of the bump (the remaining five faces of a cube) (the two along the fourth axis enter only the count).



β=1.5 u=0.3441 u^6=0.00166 bumps 24 ways ×u⁴=0.0140 with correction 0.00233

Computed The numbers in this figure were computed on this machine (u from the series of (4), computed on the spot). If JavaScript is disabled: the surface has RT plaquettes with weight u^{RT}, and in four dimensions there are $4RT$ bumps, each with u^4.

Detour: sums over surfaces and strings

A subsection you may skip.

The abstract of Wilson 1974 states that the strong-coupling expansion becomes a sum over quark paths and the lattice surfaces connecting them, and concludes "This structure is reminiscent of relativistic string models of hadrons." A sum over surfaces paying $u = e^{-\sigma 0}$ per plaquette looks like a sum over world sheets of a string with an action proportional to area Known . Speaking in metaphor, it is a membrane stretched on the lattice. Precisely, physics considers that the strong-coupling surfaces are stiff along the lattice and in a different state from a fluctuating continuum string (17-03 § 07) Physics .

04

Bumps — the first correction and the number of dimensions

The next surfaces that matter replace one plaquette of the flat surface by the remaining five faces of the unit cube with that plaquette as its base (Figure 1). Paying five plaquettes instead of one multiplies the weight by u^4, and there are $2(d-2)$ such cubes, on both sides along the $d-2$ axes perpendicular to the surface: 4 in four dimensions. Separated bumps may be counted independently at this order, so together they make an exponential (the count for $SU(2)$).

$$\langle W_{R\times T}\rangle \approx u^{RT}\, e^{2(d-2)RT\,u^4}, \qquad \sigma 1 = -\ln u - 4u^4, \qquad \left\langle \tfrac{1}{2}\mathrm{tr}\, UP\right\rangle \approx u + 4u^5 \quad (d=4)$$

The next corrections come from overlapping bumps, two-storey bumps and surfaces carrying higher representations (higher coefficients are in the review of Drouffe–Zuber 1983, and the higher series for the string tension of $SU(2)$ in four dimensions in

Münster 1981; see the table of sources). Justifying the exponential form uniformly in the volume is the cluster expansion (5-04, 7-04) Known .

Two dimensions have no bumps. For $d = 2$ the surface cannot leave the plane, and the leading term u^{RT} is exact for every β (7-02 § 03) Known . The difficulty in three or more dimensions begins with the surface being able to fluctuate out of the plane. Example: for $SU(2)$ at $\beta = 2$, $4u^4 = 0.141$ reaches 17% of $\sigma_0 = 0.837$, and the truncation becomes doubtful Computed .

05

Tubes — the leading term of the mass and the correlation length

Mass is the rate at which correlations decrease (4-11). The smallest closed surface joining a spatial plaquette $P = \frac{1}{2}\mathrm{tr}\, UP$ and the same shifted by t in the time direction, $P(t)$, is the $1 \times 1 \times t$ box with the two as lids. The lids are observables and carry no weight, and the $4t$ side plaquettes pay u^{4t}.

$$\langle P(t)P(0)\rangle - \langle P\rangle^2 \approx C\, u^{4t}, \qquad m_0 = -4\ln u = 4\sigma_0, \qquad \xi_0 = 1/m_0 \ (8)$$

Computed (lowest-order counting). The lowest-order form is said to be standard (Drouffe–Zuber 1983, bibliographic data only). That the gap in the $\mathbb{Z}_2$ of 5-03 § 08 is close to $-4\log\tanh\beta$ has the same form. The correlation length ξ_0 is 0.175 at $\beta = 1$ and still 0.38 lattice units at $\beta = 2.6$: the correlation vanishes within one lattice unit Computed . The continuum limit needs ξ much longer than the lattice spacing ($\xi(\beta) \to \infty$ in the lattice correlation-length problem of Chatterjee 2018 (Chatterjee's Problem 5.1); 5-09), and the leading term says the opposite.

Extending the ratio $m_0/\sqrt{\sigma_0} = 4\sqrt{\sigma_0}$ of (8) outside the region of convergence, the ratio of leading terms falls to 0 as β grows. The ratio from lattice numbers rises from 3.3 ($\beta = 2.25$) to 3.8 ($\beta = 2.6$) and approaches a constant value (17-12 § 08) Physics . This concerns the form of the leading terms, not a statement about the theory. The leading terms at $\beta = 2.6$, $a\sqrt{\sigma_0} = 0.808$ and $am_0 = 2.61$, are 6.0 and 5.1 times the $0.13395(62)$ and $0.514(8)$ of Table 1 of Lucini–Teper (2001) Computed (the literature values are Physics). From $\beta = 2.4 \to 2.6$ the literature $a\sqrt{\sigma}$ halves, while the leading term decreases by only 4%.

06

Comparing with numbers — how far it works

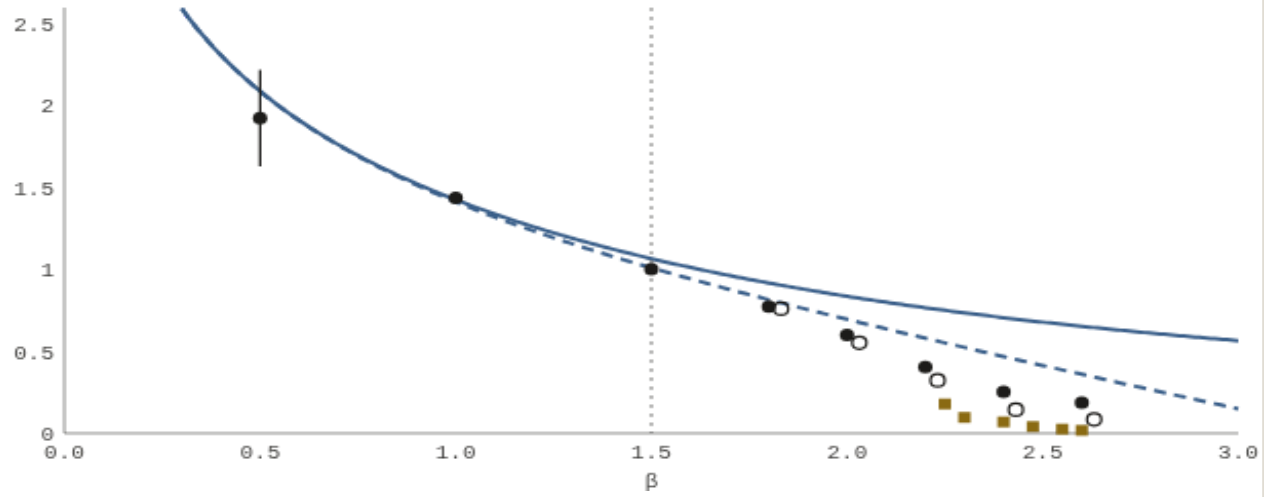
We integrated $SU(2)$ on four-dimensional periodic lattices (6 and 8 points per side) with the heat-bath method (an update that redraws the links one at a time from their

conditional distributions; 5-07), and measured the string tension with Creutz ratios ((3) of 17-02).

$$\chi(R) = -\ln \frac{W(R,R)\,W(R-1,R-1)}{W(R,R-1)^2} \tag{9}$$

The perimeter and corner terms cancel in the ratio, and with the form (7), $\chi(R) = \sigma_1$.

Figure 2 The string tension ($SU(2)$, lattice units). Solid line σ_0, dashed line σ_1, ● $\chi(2)$, ○ $\chi(3)$, ■ literature $(a\sqrt{\sigma})^2$. The slider moves the vertical line at β, and "Log scale" changes the vertical axis.



β=1.50 σ₀=1.067 σ₁=1.011 χ(2)=1.0026±0.0059

Computed The numbers in this figure were computed on this machine (curves computed on the spot, points the values of Table 2, ■ literature values). If JavaScript is disabled, Table 2 gives the content.

Table 2 Heat-bath results for four-dimensional $SU(2)$ (6 per side: $\beta \le 2$; 8: above that; statistical errors in brac and the strong-coupling formulas Computed . The last column is Lucini–Teper Physics

β	Plaquette	$u + 4u^5$	$\chi(2)$	$\chi(3)$	σ_1	Literature: plaquette / (a
0.5	0.12375(13)	0.1238	1.93(30)	–	2.089	
1.0	0.24335(14)	0.2434	1.438(32)	–	1.413	
1.5	0.36249(15)	0.3635	1.0026(59)	–	1.011	
1.8	0.44074(16)	0.4400	0.7730(21)	0.761(79)	0.816	
2.0	0.50104(22)	0.4941	0.5991(17)	0.553(33)	0.696	
2.2	0.56925(22)	0.5510	0.4054(12)	0.3231(55)	0.581	
2.4	0.63031(16)	0.6106	0.2525(11)	0.1445(23)	0.469	0.629995(17) / 0.
2.6	0.670275(68)	0.6729	0.18618(38)	0.08514(90)	0.360	0.670009(3) / 0.

There are three readings Computed . (i) For $\beta \le 1$ the plaquette agrees with $u + 4u^5$, and $\chi(2)$ with σ_1, within errors. (ii) From 1.8, $\chi(2)$ falls below σ_1, and at 2.4 it is nearly half (that the plaquette is close to $u + 4u^5$ at 2.6 is a coincidence of crossing curves). (iii) $\chi(3)$ is smaller still, yet 2 to 5 times the literature $(a\sqrt{\sigma})^2$ (small loops mix in short-

distance contributions Physics). The plaquette differs from the literature values by about 0.0003, four times the statistical error at $\beta = 2.6$. The difference may include finite-volume differences between the lattice of 8 per side and the literature's 12^4 lattice ($\beta = 2.4$) and 24^4 lattice ($\beta = 2.6$) Computed .

The abstract of Creutz (1980), who did the same kind of calculation, reports that when the lattice is made finer with the string tension held fixed, the bare coupling decreases (logarithmically) as predicted by the perturbative renormalisation group, and concludes “This supports the coexistence of confinement and asymptotic freedom for quantized non-Abelian gauge fields.” It can be read as the area law on the strong-coupling side and asymptotic freedom on the weak-coupling side appearing connected without being cut by a phase transition. That $\beta_c = \infty$ (the continuum limit only on the $\beta \to \infty$ side) for many non-abelian lattice models of interest in four dimensions is, in the words of Chatterjee (2018) § 5, at the level of “It is believed” Physics .

07

Convergence — finite boxes and infinite boxes

On a lattice with L points per side, $Z_L(\beta) = \int e^{-S_W} \prod dU_\ell$ is an integral over a finite-dimensional compact space and an entire function of β (a function expressible as a power series converging on the whole complex plane). Z_L itself is entire, but expectation values and $\ln Z_L/|\Lambda_L|$ ($|\Lambda_L|$ is the number of lattice points; the free energy) involve dividing by Z_L, so zeros of Z_L in the complex β plane shrink the radius of convergence of their expansions. The question is whether that radius of convergence survives independently of L (phase transitions in 2-08) Known . The number of connected surfaces grows with the number n of plaquettes as K^n (K fixed by the dimension), and each plaquette weighs u, so the sum converges for roughly $Ku < 1$ (a crude estimate; the precise condition is in 5-04): that is the core. The introduction of Shen–Zhu–Zhu (2023) summarises Osterwalder–Seiler (1978) as “when the coupling is sufficiently strong, the cluster expansion … for the expectation values of local observables … is convergent, uniformly in volume”, and writes that the mass gap (exponential clustering: correlations of distant observables decreasing exponentially in the distance; the mass gap of 4-11) and the existence and analyticity of the infinite-volume limit follow from it Known . SZZ themselves showed “a strictly positive mass gap” with Langevin dynamics (7-09) for $SU(N)$ with $|\beta_{\mathrm{SZZ}}| < 1/(16(d-1))$. From their formula (1.2), $\beta = N^2 \beta_{\mathrm{SZZ}}$, so for four-dimensional $SU(2)$, $\beta < 1/12$ (the conversion is this book's arithmetic). It is a sufficient condition, not the place where the expansion stops.

Where it stops

(a) Roughening: physics considers that at some β the fluctuations of the surface win, and the series for σ cannot be extended beyond it (17-03 § 07; about 2 for SU(2), with a large uncertainty in the estimate) Physics . (b) For four-dimensional $U(1)$ the area law breaks down at large β (§ 08). (c) Formulas truncated at finite order cannot express exponential smallness like $e^{-c\beta}$ as $\beta \to \infty$ (the leading term $-\ln u$ decreases only as a power, like $3/(2\beta)$). On the other hand, in the continuum limit of asymptotic freedom (4-06, 4-07), at one loop (the lowest order of the β function of 4-06) $a^2\sigma \propto e^{-(6\pi^2/11)\beta}$ should become exponentially small Physics . Taking β from 2 to 4 merely halves $-\ln u$, from 0.837 to 0.418, while $e^{-(6\pi^2/11)\beta}$ becomes about 5×10^4 times smaller Computed . The continuum-limit side cannot be reached with finite-order expansions (this book's organisation; 5-05).

The forms machine-checked on this book's definitions are in 20-01 to 20-04, and the strong-coupling area law is in 20-06. All are statements at fixed lattice spacing inside the strong-coupling window, at the opposite end from the questions on the side of the continuum limit.

08

Is the area law a proof of confinement? — it comes out for abelian groups too

The u of Table 1 is below 1 for every group, so the area law (6) comes out for $\mathbb{Z}_2$ and for $U(1)$ too. Chatterjee (2018) § 4 writes “the area law holds for any lattice gauge theory at sufficiently small β”, and goes on to say that therefore “the area law at small β cannot be evidence for quark confinement”. The proof is Osterwalder–Seiler (1978), whose abstract writes “We then prove existence and analyticity of the infinite volume limit of strongly coupled Yang-Mills theories on the lattice and we verify Wilson's confinement bound.” Known . The four-dimensional $U(1)$ lattice does not satisfy the area law at large β (Guth 1980; Guth showed it with the Villain form of the action, different from the Wilson form; a fully rigorous proof is Fröhlich–Spencer 1982; 5-06) Known . The strong-coupling area law is a property of the lattice being coarse, and does not use non-commutativity. The lattice area-law problem of the same paper (Chatterjee's Problem 4.1) asks to show, in any infinite-volume limit of a four-dimensional non-abelian compact group, for all $\beta > 0$,

$$|\langle W_{R\times T}\rangle| \le C(\beta)\, e^{-c(\beta)RT} \quad (10)$$

and writes that at large β it “remains elusive”. On the strong-coupling side, Cao–Nissim–Sheffield extended the range of the area law for $U(N)$ at large N (arXiv:2505.16585, a preprint, publication not checked; abstract: “This improves on a

classical result of Osterwalder-Seiler from 1978”), and in the 't Hooft regime (letting N grow with β proportional to N; for converting conventions, Table 4 of 5-01) they checked, with the SZZ method, a mass gap condition that the abstract cites as [DF80], and derived the area law from it (arXiv:2509.04688, a preprint, publication not checked) Known . Both are results at fixed lattice spacing. The mass gap and confinement are different statements (17-02 § 02; finite groups in 5-10), and this chapter shows neither for the continuum theory.

What this chapter can say

Item	What is said	What is not said
The character expansion, the leading terms (3) to (6)	Standard tools Known . Agreeing with numbers Computed	—
The correction (7)	Lowest-order counting for SU(2) in four dimensions. Agrees with the plaquette and $\chi(2)$ for $\beta \leq 1$ at $L = 6$ Computed	Higher coefficients, error estimates for truncation (5-04), infinite-volume values
The leading term of the mass (8)	Lowest-order counting Computed . Not compared with numbers on this machine. 5.1 times the literature value at $\beta = 2.6$	Numerical check of the mass, higher corrections
Theorems on convergence and the area law	Uniform convergence, the area law and exponential decay at small β (literature) Known	The content of the proofs. The area law at large β (the lattice area-law problem)
Confinement, the mass gap	Presentation of results on the strong-coupling side at fixed lattice spacing	Both in the continuum limit. This chapter shows nothing about them

Sources and reproduction

Item	Kind	Source / tools
Quotations in § 01 and the detour	Abstract	K. G. Wilson, Phys. Rev. D 10 (1974) 2445–2459
The results of Osterwalder–Seiler in § 07 and § 08, the quotation in § 08	Abstract (INSPIRE, Elsevier version; text not obtained)	K. Osterwalder, E. Seiler, Ann. Phys. 110 (1978) 440–471
The introduction in § 07, formula (1.2), the condition	Text § 1	H. Shen, R. Zhu, X. Zhu, Commun. Math. Phys. 400 (2023) 805–851, arXiv:2204.12737

Item	Kind	Source / tools
Quotations in § 05, § 06 and § 08, the lattice area-law problem and the lattice correlation-length problem (Chatterjee's Problems 4.1 and 5.1)	Text	S. Chatterjee, arXiv:1803.01950 (2018) § 4– § 6
Quotation in § 06	Abstract	M. Creutz, Phys. Rev. D 21 (1980) 2308–2315
Literature values in § 05, Table 2 and Figure 2	Tables 1 and 7 of the text Physics	B. Lucini, M. Teper, JHEP 0106 (2001) 050, arXiv:hep-lat/0103027
Extension of the area law in § 08	Abstracts (preprints; publication of neither checked)	S. Cao, R. Nissim, S. Sheffield, arXiv:2505.16585, arXiv:2509.04688 (2025). The [DF80] cited in the abstract of 2509.04688 is said to be Durhuus–Fröhlich 1980 (bibliographic data not checked)
The non-confining phase of $U(1)$ (Guth)	Abstract (INSPIRE)	A. H. Guth, Phys. Rev. D 21 (1980) 2291–2307
The non-confining phase of $U(1)$ (Fröhlich–Spencer)	Bibliographic data only (from Chatterjee's account)	J. Fröhlich, T. Spencer, Commun. Math. Phys. 83 (1982) 411–454
Review of higher coefficients, the lowest-order form of (8)	Bibliographic data only	J.-M. Drouffe, J.-B. Zuber, Phys. Rep. 102 (1983) 1–119
Higher series for the string tension of four-dimensional $SU(2)$	Abstract only	G. Münster, Nucl. Phys. B180 (1981) 23–60
The roughening transition in § 07	From the sources of 17-03	Itzykson, Peskin, Zuber, Phys. Lett. B 95 (1980) 259–264
Tables 1 and 2, the figures, numbers in the text	Computed on this machine	Python 3.12, NumPy, SciPy. `chars.py` (Table 1, (4)), `sc_tables.py` (§ 05, § 07), `mc_su2.py`, `run_mc.sh` (Table 2; heat bath, jackknife with 20 bins, 2 to 4 minutes per run). The figures are plain JavaScript

Next chapter to read: 5-03 Reflection positivity and the transfer matrix (extracting the Hamiltonian from the same lattice).

Reflection positivity and the transfer matrix — if the integral reflected in a mirror is positive, a Hamiltonian stands up on the lattice

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — The Wilson action is reflection positive for site reflection at every β, and for link reflection at $\beta \geq 0$. From this a positive transfer matrix and a lattice Hamiltonian stand up, and taking the time step to 0 gives the Kogut–Susskind form.

Prerequisite chapters — 5-01, 4-10. See also 2-09 (transfer matrices) and 7-02 § 03 (characters). β is the inverse lattice coupling of 5-01, a different quantity from the β function $\beta(g)$ of 4-06.

Said first — This is about finite lattices. On a finite box a positive gap is automatic, and this chapter shows nothing about whether the gap survives when the lattice spacing goes to 0 and the volume to infinity. The numbers come from calculations such as $\mathbb{Z}_2$ with space up to 4×4.

Order of this chapter

1. Where this chapter stands — from integrals to operators
2. Time reflection on the lattice — two kinds of plane (Figure 1)
3. A one-dimensional chain — two kinds of "positive"
4. Reflection positivity of the Wilson action — two proofs
5. The transfer matrix — the lattice Hamiltonian
6. A single link — the energy of the electric field
7. The Kogut–Susskind Hamiltonian — taking the time step to 0
8. Z_2 gauge theory — a transfer matrix that is not positive (Figure 2, Table 1)
9. What is kept, and what remains — limits, improved actions, what is not said

01

Where this chapter stands — with one mirror, quantum mechanics comes back

The theory on a lattice (5-01) is written as the following probability measure.

$$d\mu_\beta(U) = \frac{1}{Z}\, e^{-S_W(U)} \prod_\ell dU_\ell, \qquad S_W = \beta \sum_P \Big(1 - \tfrac{1}{N} \operatorname{Re}\operatorname{tr} U_P\Big), \qquad \beta = \frac{2N}{g^2} \quad (1)$$

dU is the Haar measure of total weight 1, and β the coefficient of the action (for example, $\beta = 4$ for $SU(2)$ with $g^2 = 1$). The last coordinate t is called imaginary time. Here there are only integrals and expectation values, with no Hilbert space and no Hamiltonian. The mass gap is a gap in the spectrum of the Hamiltonian (4-11), so we first recover the operators. The tool is the reflection positivity of 4-10, which on the lattice can be proved from the action. § 6.5 of the problem statement writes 'Reflection positivity holds for the Wilson approximation [36], a major advantage' ([36] is Osterwalder–Seiler 1978) Known .

02

Time reflection on the lattice — two kinds of plane

For the lattice reflected in the mirror to coincide with the original, the $2t_0$ of the plane $t = t_0$ must be an integer. If t_0 is an integer, the plane passes through lattice points (**site reflection**); if it is a half-integer, it cuts the midpoints of the vertical (time-direction) links (**link reflection**).

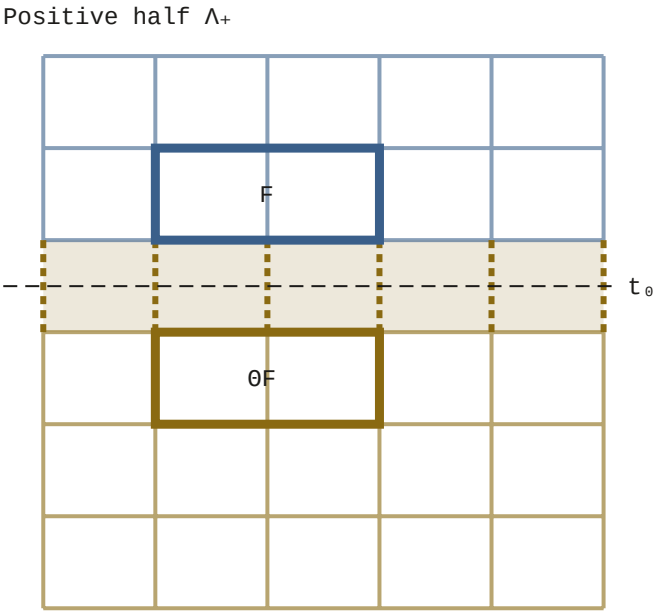
$$\theta(\boldsymbol{x}, t) = (\boldsymbol{x}, 2t_0 - t), \qquad (\theta U)_{x \to y} = U_{\theta x \to \theta y}, \qquad (\Theta F)(U) = F(\theta U) \quad (2)$$

A link variable traversed backwards is the inverse ($U_{y\to x} = U^{-1}_{x\to y}$). Let Λ_+ be the set of links whose two ends both have $t \geq t_0$ (for link reflection, $t > t_0$), and $\mathcal{A}_+$ the set of all functions determined by those variables alone. Reflection positivity is the condition of the same form as (4) of 4-10,

$$\langle\, \Theta F \; F \,\rangle \geq 0 \qquad (F \in \mathcal{A}_+) \quad (3)$$

θ maps plaquettes to plaquettes (reversing the orientation of those containing the time direction), and $\operatorname{Re}\operatorname{tr}$ is unchanged by inversion, so $S_W(\theta U) = S_W(U)$.

Figure 1 Two kinds of mirror — Moving the plane t_0 by half a lattice unit at a time with the slider, site reflection and link reflection appear alternately. Blue is the positive half and F (a rectangular Wilson loop), brown its image.



t₀ = 1/2 (link reflection): the dashed links straddle the plane, and the shaded plaquettes join the two halves.

Example (Figure 1). At $t_0 = \frac{1}{2}$ a vertical link crossing the plane is mapped to itself with orientation reversed ($(\theta U)_b = U_b^{-1}$), belongs to neither half, and straddles the plane. The plaquettes having it as an edge join the upper and lower halves. At $t_0 = 0$ the horizontal links at $t = 0$ are mapped to themselves, and nothing straddles.

03

Seen in a one-dimensional chain — two kinds of "positive"

Put the weight $\prod_t e^{K s_t s_{t+1}}$ on a chain of spins $s_t = \pm 1$ (the one-dimensional Ising model of 2-09 with zero external field; $K < 0$ is also allowed). The transfer matrix is

$$T = \begin{pmatrix} e^{K} & e^{-K} \\ e^{-K} & e^{K} \end{pmatrix}, \qquad \lambda_0 = 2\cosh K, \quad \lambda_1 = 2\sinh K \tag{4}$$

For **site reflection** ($t_0 = 0$), fixing s_0, the sum $\varphi(s_0)$ over the upper half and the sum over the lower half are complex conjugates of each other, and whatever the sign of K, $\langle \Theta F F \rangle \propto \sum_{s_0} |\varphi(s_0)|^2 \geq 0$. For **link reflection** ($t_0 = \frac{1}{2}$) one straddling coupling remains, and

$$\langle \Theta F F \rangle \;\propto\; \sum_{s_0, s_1} \varphi(s_0)\; T_{s_0 s_1}\; \varphi(s_1) \tag{5}$$

is a quadratic form in T. It is non-negative for every φ only when $T \geq 0$, that is, $K \geq 0$. Example. For $F = s_1$, with link reflection $\langle s_0 s_1 \rangle = \tanh K$ becomes -0.4621 at $K = -\frac{1}{2}$ and (3) is violated, while with site reflection $\langle s_{-1} s_1 \rangle = \tanh^2 K = 0.2136 \geq 0$ Computed .

There are two kinds of "positive". One is that the entries are positive. For a matrix with all entries positive, the largest eigenvalue is not repeated and is larger than the absolute values of the other eigenvalues (the Perron–Frobenius theorem) Known . The entries of the T in (4) are positive whatever the sign of K, so $\lambda_0 > |\lambda_1|$, and the correlation $(\tanh K)^t$ decreases exponentially. The other is being positive as an operator. $T \geq 0$ only when $K \geq 0$, and if $K > 0$ one can write $T = e^{-H}$ with a real H. For $K < 0$ the correlation decreases while changing sign, and $-\log \lambda_1$ is not real. Site-reflection positivity corresponds to $T^2 \geq 0$, and link-reflection positivity to $T \geq 0$ Known .

04

Reflection positivity of the Wilson action — two proofs

Claim: on a lattice with the Wilson action, (3) holds for all real β with site reflection, and for $\beta \geq 0$ with link reflection. Reflection positivity of lattice gauge theory and the existence of a positive self-adjoint transfer matrix are due to Osterwalder–Seiler 1978. Their abstract states that physical positivity was verified, and continues 'This implies the existence of a positive self-adjoint transfer matrix' Known . The conditions on β for each kind of reflection rest on the proofs (6) to (8) below Paper . An example where link reflection actually fails at $\beta < 0$ is in section 08 Computed . Below are outlines of the proofs (with periodic time there are two planes, but it is the same).

Site reflection — fixing the variables on the plane

No plaquette straddles. Distributing the action of the plaquettes on the plane half up and half down gives $S_W = S_+ + S_-$ and $S_-(U) = S_+(\theta U)$. Fixing the variables U_0 on the plane, the lower integral becomes the complex conjugate of the upper one by the substitution $U_< = \theta U_>$ (Haar measure is invariant under inversion), so

$$\langle \Theta F F \rangle = \frac{1}{Z} \int dU_0 \left| \int dU_> \, F \, e^{-S_+} \right|^2 \geq 0 \quad (6)$$

The sign of β is used nowhere.

Link reflection — opening the straddling plaquettes into sums of squares

The straddling vertical links have distinct upper endpoints, so gauge transformations at those points can set them to 1 (7-02 § 02). For this, a general F is first replaced by $\widetilde{F}$, averaged over gauge transformations at the points of the layer just above the plane ($t = 1$). ΘF is determined by the variables with $t \leq 0$ only, and does not touch the points of that layer. So changing the integration variables $U \to U^g$ by a gauge transformation g of that layer (neither the action nor Haar measure changes), $\langle \Theta F F \rangle = \langle \Theta F F^g \rangle$, and

averaging over g gives $\langle \Theta F \widetilde{F} \rangle$. Doing the same at the mirror layer ($t = 0$), the left-hand side of (3) equals $\langle \Theta \widetilde{F}\, \widetilde{F} \rangle$. A straddling plaquette becomes $\mathrm{Re}\,\mathrm{tr}(U_{\theta\ell} U_\ell^\dagger)$, involving only the upper horizontal link ℓ and its image, and with the entries $h_{ij} = (U_\ell)_{ij} \in \mathcal{A}_+$ ($\Theta h_{ij} = (U_{\theta\ell})_{ij}$) it can be written

$$\mathrm{Re}\,\mathrm{tr}\big(U_{\theta\ell} U_\ell^\dagger\big) = \frac{1}{2} \sum_{i,j} \Big(\Theta h_{ij}\, h_{ij} + \Theta h_{ij}\, h_{ij} \Big) \tag{7}$$

Expanding the weight of the straddling plaquettes $\exp\big(\frac{\beta}{2N} \sum_k \Theta h_k h_k\big)$ (h_k ranges over all h_{ij} and h_{ij}; constant factors omitted) as a series, it takes the form $\sum_\alpha c_\alpha \Theta H_\alpha H_\alpha$ with products H_α of the h_k, and if $\beta \geq 0$ then $c_\alpha \geq 0$. The upper and lower halves no longer share variables, so

$$\langle \Theta F F \rangle \propto \sum_\alpha c_\alpha \Big| \int dU_> \, F\, H_\alpha \, e^{-S_+} \Big|^2 \geq 0 \tag{8}$$

Just as $T \geq 0$ was needed in (5) for the chain, $\beta \geq 0$ is needed here. An example where it actually fails is in section 08.

The forms machine-checked on this book's definitions are in 20-01 to 20-04.

05

The transfer matrix — the lattice Hamiltonian

Given reflection positivity, the reconstruction of 4-10 makes a Hilbert space and an operator that advances time by one lattice unit; on the lattice this operator (the transfer matrix) can be written down directly. Make time periodic (length L_t) and set as many vertical links as possible to 1 by gauge transformations (temporal gauge); then one link V_x remains on the time loop at each spatial point x. Integrating over the remaining V_x is the same as applying the gauge transformation $g_x = V_x$ to all links at one time and averaging, and that average is the projection P_0 onto gauge-invariant states. The state space is the space $\mathcal{H} = L^2(G^{E_s}, dU)$ of functions of the spatial links E_s at one time, and T is the integral operator with kernel

$$T(U, U') = e^{-\frac{1}{2} S_s(U)} \prod_{\ell \in E_s} e^{-\beta(1 - \frac{1}{N} \mathrm{Re}\,\mathrm{tr}\, U_\ell U_\ell'^{\dagger})}\, e^{-\frac{1}{2} S_s(U')} \tag{9}$$

(S_s is the action of the spatial plaquettes at one time, and the product in the middle comes from the vertical plaquettes). Then

$$Z = \mathrm{Tr}\big(P_0 T^{L_t}\big), \qquad \langle O(t) O(0) \rangle - \langle O \rangle^2 \xrightarrow{L_t \to \infty} \sum_{n \geq 1} |\langle \Omega | O | n \rangle|^2 \Big(\frac{\lambda_n}{\lambda_0} \Big)^t \tag{10}$$

(O is a gauge-invariant real observable, λ_n and $|n\rangle$ the eigenvalues and eigenvectors of $P_0 T$, $\Omega = |0\rangle$; the same calculation as (8) of 2-09).

(a) The kernel is real symmetric, so T is self-adjoint (the content of site reflection). (b) $T = VKV$ (V the operator multiplying by $e^{-S_s/2}$, K the product in the middle), and if $\beta \geq 0$ then $K \geq 0$ (section 06), so $T \geq 0$ (the content of link reflection). (c) For $\beta > 0$, V multiplies by a positive function and so has an inverse, and the λ_r of section 06 are all positive, so K does not have 0 as an eigenvalue either, and T does not have 0 as an eigenvalue Paper . Lüscher 1977 calls this property 'strictly positive' in its title. (d) On a finite box the kernel of $P_0 T$ is continuous and positive and the space of gauge orbits is compact, so by a generalisation of Jentzsch's theorem on integral operators with continuous positive kernels on compact spaces (of Krein–Rutman type; the integral-operator version of the Perron–Frobenius theorem), λ_0 is not repeated Known (the generalisation is Schaefer 1974; see the table of sources).

From here on we take $\beta > 0$. By (b)(c) all eigenvalues are positive, so since λ_0 is not repeated, $\lambda_1 < \lambda_0$. Thus on $P_0 \mathcal{H}$

$$H_a = -\frac{1}{a} \log \frac{T}{\lambda_0} \;\geq\; 0, \qquad m_L(\beta) = \log \frac{\lambda_0}{\lambda_1} \;>\; 0 \quad (11)$$

stands up (a is the lattice spacing). Example. The gap in physical units is $\Delta = m_L / a$. That $m_L > 0$ for each box ($\beta > 0$) is automatic from (d), and what is asked is whether Δ remains positive as $L \to \infty$ and $a \to 0$ (4-11 § 08, 5-09).

For how this book uses this, see 26-04 § 01 (positivity of the transfer matrix and reflection positivity applied to the bias of the twisted sectors).

06

The transfer matrix of a single link — the energy of the electric field appears

K is a product of operators k, one per link. For a single link $(kf)(U) = \int_G e^{\frac{\beta}{N} \mathrm{Re}\,\mathrm{tr}\, UU'^{-1} - \beta} f(U')\, dU'$, and the kernel is a function unchanged by conjugation $W \to hWh^{-1}$ of $W = UU'^{-1}$ (a class function). Diagonalising it needs two words from representation theory.

- We write the **character** of a representation r (of dimension d_r) as $\chi_r(W) = \mathrm{tr}\, r(W)$. It is a class function, and between irreducible representations $\int_G \chi_r(W) \chi_s(W)\, dW = \delta_{rs}$ (orthogonality) Known .
- Convolution with a class function acts as a scalar multiple on the space spanned by the matrix elements $r(U)_{ij}$ of each irreducible representation (from Schur's lemma

in 1-04). $L^2(G)$ is spanned by these matrix elements (the Peter–Weyl theorem), so k is determined by its eigenvalue on each irreducible representation Known .

For details, (3.1) in 7-02 § 03. The eigenvalue on the representation r is

$$\lambda_r = \frac{1}{d_r}\int_G e^{\frac{\beta}{N}\operatorname{Re}\operatorname{tr} W - \beta}\,\chi_r(W)\,dW \tag{12}$$

Expanding the exponential in powers of $\operatorname{tr} W$ and $\operatorname{tr} W$, each power is a character of a tensor-product representation, and the multiplicities of irreducible representations are non-negative, so if $\beta \geq 0$ then $\lambda_r \geq 0$, that is, $K \geq 0$. If $\beta > 0$ all the coefficients of the expansion are positive, and every irreducible representation appears in some tensor product of the fundamental representation and its complex conjugate Known , so $\lambda_r > 0$. For $SU(2)$, with the eigenvalues of W written $e^{\pm i\theta}$ ($0 \leq \theta \leq \pi$), $\operatorname{tr} W = 2\cos\theta$ and $\chi_j = \sin((2j+1)\theta)/\sin\theta$. The Haar measure for integrating class functions is $\frac{2}{\pi}\sin^2\theta\,d\theta$, which is the density of the uniform distribution on S^3 of 5-01 seen through the angle θ from the north pole (the simplest case of Weyl's integration formula). From these Known

$$\lambda_j = e^{-\beta}\,\frac{2}{\beta}\,I_{2j+1}(\beta), \qquad j = 0, \tfrac{1}{2}, 1, \ldots \tag{13}$$

($I_n(\beta) = \frac{1}{\pi}\int_0^\pi e^{\beta\cos\theta}\cos n\theta\,d\theta$ is the modified Bessel function; agreeing with numerical integration to relative error below 10^{-10} Computed ; the ratio λ_j/λ_0 is the u_j of (4) in 5-02). For $\beta < 0$ it is negative for half-integer j. From the asymptotic form at large β, $I_\nu(\beta) \approx \frac{e^\beta}{\sqrt{2\pi\beta}}\left(1 - \frac{4\nu^2-1}{8\beta}\right)$ (DLMF 10.40.1), the energy per step is

$$E_j = -\log\frac{\lambda_j}{\lambda_0} \;\approx\; \frac{2j(j+1)}{\beta} = \frac{g^2}{2}\,j(j+1) \tag{14}$$

$j(j+1)$ is the value of the Casimir E^aE^a (the J^2 of 1-04). Speaking in metaphor, a single link is a top spinning on $SU(2)$ (a three-dimensional sphere), and E^a is its angular momentum. Precisely, at large β, k approaches the heat kernel $e^{-\frac{g^2}{2}E^aE^a}$ on the group (7-02 § 04).

Example Computed . $E_{1/2}$ is 0.0967 at $\beta = 16$ ($g^2 = \frac{1}{4}$) ((14) gives 0.0938), and 0.0236 at $\beta = 64$ (0.0234). The deviation of the ratio $E_{1/2}/(14)$ (the relative error) shrinks from 1.032 to 1.008, roughly as $1/\beta$ (the difference itself shrinks from 0.00295 to 0.00018, as $1/\beta^2$). At strong coupling, $\beta = 1$, $E_1 = 3.24$ departs from the approximation 4. For the ratio $E_1/E_{1/2} \to 8/3$, see 7-02 § 07.

07

The Kogut–Susskind Hamiltonian — taking the time step to 0

The H_a of (11) is the logarithm of an integral operator, and its form is not visible. A local form becomes visible if only the spacing a_t in the time direction is made finer than the spatial spacing a. In four dimensions, we put separate coefficients on the plaquettes containing the time direction and those that are purely spatial,

$$\beta_t = \frac{2N}{g^2}\,\frac{a}{a_t}, \qquad \beta_s = \frac{2N}{g^2}\,\frac{a_t}{a} \tag{15}$$

From the number of plaquettes per volume $1/(a^3 a_t)$ and (7) of 5-01, the classical continuum limit is the action (8) of 5-01 in either direction. As $a_t \to 0$, $\beta_t \to \infty$, so for a step in the time direction the general form of (14), $-\log(\lambda_r/\lambda_0) \approx \frac{N}{\beta_t} C_2(r) = a_t \frac{g^2}{2a} C_2(r)$, can be used ($C_2$ is the value of the Casimir, normalised by $\operatorname{tr} T^a T^b = \frac{1}{2}\delta^{ab}$). This is because near $W = e^{iX^a T^a}$, $\frac{\beta}{N}\operatorname{Re}\operatorname{tr} W \approx \beta - \frac{\beta}{4N} X^a X^a$, and k is close to the heat kernel at time N/β (for the fundamental representation of $SU(3)$ at $\beta = 640$, the ratio is 1.003 Computed). The spatial part $V^2 = e^{-S_s}$ has $\beta_s \propto a_t$, so it has the form $e^{-a_t(\cdots)}$ from the start. Together $T = e^{-a_t H_{\mathrm{KS}} + O(a_t^2)}$, with

$$H_{\mathrm{KS}} = \frac{g^2}{2a}\sum_{\ell} E_\ell^a E_\ell^a + \frac{2N}{g^2 a}\sum_{p}\Big(1 - \frac{1}{N}\operatorname{Re}\operatorname{tr} U_p\Big) \tag{16}$$

The first term is the energy of the electric field (the spinning of the top on each link), the second that of the magnetic field, and the states are restricted to those satisfying the Gauss law ($P_0\mathcal{H}$). This form was given by Kogut–Susskind 1975, and the abstract of Creutz 1977 writes 'Taking the timelike lattice spacing to zero, we extract the Kogut-Susskind Hamiltonian from the transfer matrix in the gauge A0 = 0' Known . The coefficients of (15)(16) are calculations in the conventions of this chapter, and the limit is an approximation eigenvalue by eigenvalue. From canonical quantisation, see 4-01.

Example. When g^2 is large the first term dominates, and the first excitation is a state in which an electric field in the fundamental representation goes once round the boundary of a single plaquette, with energy $\frac{g^2}{2a}\cdot 4 \cdot C_2(\mathrm{fund})$. For $SU(N)$, $C_2(\mathrm{fund}) = \frac{N^2-1}{2N}$, and for $SU(2)$ ($j = \frac{1}{2}$, $C_2 = \frac{3}{4}$) it is $\frac{3g^2}{2a}$. The energy of a line of electric field is proportional to its length (5-02).

For how this book uses this, see 24-05 § 04 (reading the gap as the Poincaré constant of the vacuum measure on the Hamiltonian lattice).

08

Seeing everything in $\mathbb{Z}_2$ gauge theory — a transfer matrix that is not positive

Take the group to be $\mathbb{Z}_2 = \{\pm 1\}$ (the detour of 2-09, 5-10), spacetime three-dimensional (space a periodic $L \times L$ lattice), and the action $\beta \sum_P (1 - U_P)$. The factor for a single link is $e^{-\beta}$ times the matrix (4) with K replaced by β, and the ratio of its eigenvalues is $\tanh \beta$. For $\beta < 0$ it is negative.

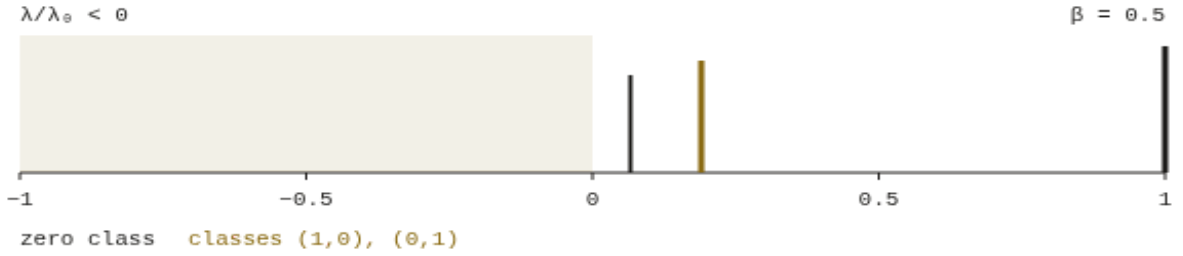
In the eigenvectors of the factor of each link (the electric-field basis), it is decided for each link whether a line of electric field passes through it. The Gauss law says that an even number of lines meet at each point, gauge-invariant states are labelled by collections e of closed lines, and K is a diagonal matrix determined only by the total length $|e|$ of the lines.

$$K(e) \;\propto\; (\tanh \beta)^{|e|} \qquad (17)$$

Closed lines fall into four classes by the parities of their winding numbers in the x and y directions, and V (which adds or removes the line round a plaquette's boundary) preserves the class. Adding a boundary does not change the parity of the length, so lines in class $(1,0)$ have length of the same parity as L. Hence for $\beta < 0$ with L odd, $K \leq 0$ in classes $(1,0)$ and $(0,1)$, and $T = VKV \leq 0$. If L is even, the ratios λ/λ_0 are the same for β and $-\beta$. Flipping the signs of the x-direction links on every other row gives $U_p \to -U_p$, and $V(-\beta)$ becomes similar to a constant multiple of $V(\beta)$. If L is even, all closed lines have even length, so the K of (17) also changes only by a constant factor Paper (agreeing for all eigenvalues at $L = 2$ and the top ones at $L = 4$ Computed).

Each class has dimension 2^{L^2-1}; $L \leq 3$ is diagonalised directly and $L = 4$ with the Lanczos method (a method that finds only the extreme eigenvalues of a large symmetric matrix by iteration) (at $L = 3$ the two agree to the 8 to 11 digits printed Computed). Let $F = W(1)$ be the Wilson line going once round in the x direction at time 1. The left-hand side of (3) for link reflection ($t_0 = \frac{1}{2}$) is $\langle W(0)W(1)\rangle = \sum_n |\langle 0|W|n\rangle|^2 \lambda_n/\lambda_0$ (n in class $(1,0)$), and at $L = 3$, $\beta = -0.5$ it is -0.1306, negative. For site reflection ($t_0 = 1$, $F = W(2)$), $\langle W(0)W(2)\rangle = +0.0183$. Monte Carlo without the transfer matrix ($3 \times 3 \times 18$) also gives $-0.1305(18)$ and $+0.0189(11)$, agreeing within errors Computed .

Figure 2 Eigenvalues of the $\mathbb{Z}_2$ transfer matrix (space 3×3) — Moving β (−1 to 2) with the slider moves the vertical lines to the positions of the eigenvalue ratios λ/λ_0 (black: λ_0 and λ_1 of the zero class; brown: the one of largest absolute value in classes $(1,0)$ and $(0,1)$).



β = 0.5: zero class λ_1/λ_0 = 0.0663 (m = 2.71), largest λ/λ_0 in class (1,0) = 0.190. All eigenvalues positive.

Computed The numbers in this figure were computed on this machine.

Computed Table 1 The gap $m = \log(\lambda_0/\lambda_1)$ of the zero class (states with no winding lines), and the lowest energy $E_w = \log(\lambda_0/\lambda_0^{(1,0)})$ of class $(1,0)$ (lattice units).

β	m (L=3)	m (L=4)	−4 log tanh β	E_w (L=3)	E_w (L=4)
0.10	9.233	9.233	9.224	6.915	9.223
0.75	1.587	1.153	1.816	0.1860	0.1443
1.00	2.991	2.840	1.089	0.0200	0.00459

How to read it. (i) For $\beta > 0$ all eigenvalues are positive and $m > 0$ ((d) of section 05). (ii) At small β, $m \approx -4 \log \tanh \beta$ (a line of length 4; the m_0 of 5-02), and $E_w \approx -L \log \tanh \beta$ (proportional to the length of the winding line: the confining form). (iii) At large β, E_w becomes rapidly smaller with L. (iv) Near $\beta = 0.75$, m moves a lot with L. Three-dimensional $\mathbb{Z}_2$ gauge theory is dual to the three-dimensional Ising model and is said to have a phase transition (Wegner 1971) Known , and the m of finite boxes alone does not determine the infinite-volume gap. (v) For $\beta < 0$ with $L = 3$, all 512 eigenvalues of classes $(1,0)$ and $(0,1)$ are at most 0 (the smallest is $-0.788\lambda_0$ at $\beta = -1$), and $\langle W(0)W(t)\rangle$ changes sign with the parity of t (-0.0027 at $\beta = -0.5$, $t = 3$) Computed .

09

What is kept, and what remains

Kept in the limit. (3) is an inequality, so if expectation values converge along a sequence of lattices and the mirror plane lies on a lattice plane at every stage, the limiting measure is also reflection positive (7-01 § 07) Known . In the four-dimensional non-abelian case, the existence of the continuum limit was shown by the construction published in 2026-11, and the reflection positivity of the limiting correlations is also stated there (the committee's assessment is that it is proved in Lean; 12-01, 12-05). The limit, however, is at the level of correlation functions; a limit as a measure has not been constructed (6-04).

Can be lost when the action is changed. For improvements adding dimension-6 terms (the Symanzik improvement of 5-05), the abstract of Lüscher–Weisz 1984 writes

'physical positivity is lost in general', and states that the transfer matrix can have complex eigenvalues (correlations with damped oscillation) and that negative weights necessarily appear in the spectral decomposition of the two-point function Known . The $\beta < 0$ of section 08 is the simplest example of a transfer matrix that is no longer positive (the eigenvalues are real and negative). For continuum regularisations, see 4-10 § 08.

What is not said. The tools of this chapter do not answer whether $mL > 0$ survives as $L \to \infty$ (the strong-coupling side is 5-04) and $a \to 0$. The recovery of rotational symmetry is also a task on the side of the limit.

What this chapter can say

Content	Status
Reflection positivity of lattice gauge theory and a positive self-adjoint transfer matrix	Known (abstract of Osterwalder–Seiler 1978)
The conditions on β in (3): all β for site reflection, $\beta \geq 0$ for link reflection	Paper ((6) to (8) of this chapter)
An example where link reflection fails at $\beta < 0$	Computed (the $\mathbb{Z}_2$ of section 08, $L = 3$)
T self-adjoint, positive for $\beta \geq 0$	Known (Osterwalder–Seiler 1978, Lüscher 1977)
For $\beta > 0$, T does not have 0 as an eigenvalue	Paper ((c) of section 05, $\lambda_r > 0$ of section 06)
Simplicity of λ_0, $mL > 0$ for each box ($\beta > 0$)	Known (a generalisation of Jentzsch's theorem; Schaefer 1974)
(12) to (16)	The formulas are Known ; the numbers of section 06 and the check of the general form for $SU(3)$ are Computed . The coefficients of (15)(16) are in this chapter's conventions, and $a_t \to 0$ is an approximation eigenvalue by eigenvalue
Table 1 and Figure 2 for $\mathbb{Z}_2$ ($L \leq 4$)	Computed (diagonalisation and independent Monte Carlo)
The mass gap, the continuum limit	Nothing shown

Sources and reproduction

Item	Kind	Source / tools
Table 1, Figure 2, numbers in the text	Computed	python3 (NumPy, SciPy): `table1.py`, `one_link.py`, `z2_transfer.py`, `check_lanczos.py`, `even_L_check.py`, `winding_corr.py`, `mc_check.py`, `fig_data.py`, `su3_casimir.py`. $\langle W(0)W(t)\rangle$ in two ways, by spectral representation and by Monte Carlo
Problem statement § 6.5	Known	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory" (2000). Text checked
Reflection positivity, the transfer matrix	Known	K. Osterwalder, E. Seiler, "Gauge field theories on a lattice", Ann. Phys. 110 (1978) 440–471 (abstract checked, text not checked) / M. Lüscher, Commun. Math. Phys. 54 (1977) 283–292 (title and abstract checked)
(16), improved actions	Known	J. Kogut, L. Susskind, Phys. Rev. D 11 (1975) 395–408 / M. Creutz, Phys. Rev. D 15 (1977) 1128–1136 / M. Lüscher, P. Weisz, Nucl. Phys. B 240 (1984) 349–361. Abstracts of all checked
Perron–Frobenius, Jentzsch	Known	O. Perron, Math. Ann. 64 (1907) 248–263 / R. Jentzsch, "Über Integralgleichungen mit positivem Kern", J. reine angew. Math. 141 (1912) 235–244. Bibliographic data of both checked with Crossref; text bibliographic data only (Jentzsch's claim [for a continuous positive kernel on a square, the spectral radius is a simple eigenvalue with a positive eigenfunction] checked in the secondary source arXiv:1910.01120). Generalisation to kernels on compact spaces: the theorem on p. 337 of H. H. Schaefer, *Banach Lattices and Positive Operators*, Springer (1974) (for an integral operator with kernel $N \geq 0$, some power of which is compact, if the integral of the kernel linking any set S of positive measure with its complement is positive, then the spectral radius is a simple eigenvalue with an eigenfunction positive almost everywhere; if $N > 0$, the other eigenvalues are smaller in absolute value). Bibliographic data only (the claim checked through the citations in arXiv:1910.01120 and arXiv:1501.06011; a continuous positive kernel on a finite box is Hilbert–Schmidt and so satisfies the assumptions). M. G. Krein, M. A. Rutman, Uspekhi Mat. Nauk 3:1(23) (1948) 3–95 (English translation Amer. Math. Soc. Translation 26, 1950). Bibliographic data checked with Math-Net.Ru; text bibliographic data only
The asymptotic form of I_ν	Known	NIST DLMF 10.40.1, 10.17.1. Formulas checked
The duality of $\mathbb{Z}_2$ and Ising	Known	F. J. Wegner, J. Math. Phys. 12 (1971) 2259–2272. Bibliographic data only

Next chapter to read: 5-04 The infinite-volume limit and the cluster expansion.

The infinite-volume limit and the cluster expansion — at small β the expansion in lumps converges independently of the box, and the limit and a lattice gap follow

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — An infinite-volume limit can always be taken along a subsequence, but it need not be unique. At small β the expansion in lumps of plaquettes converges independently of the box, and the existence of the limit, analyticity and exponential decay of correlations (a gap in lattice units) follow. Uniqueness independent of the boundary condition was shown by Shen–Zhu–Zhu along the road of Langevin dynamics (§ 08). Both guarantees cover a much narrower range.

Prerequisite chapters — 5-02. See also 4-11, 5-03 and 7-04. β is the inverse lattice coupling of 5-01, a different quantity from the β function $\beta(g)$ of 4-06.

Said first — This is about the side of small β at fixed lattice spacing, the shore opposite the continuum limit $\beta \to \infty$. This chapter shows nothing about the infinite-volume limit or the mass gap of the continuum theory. Table 3 is a fit of a crude estimate, not the best value in the literature.

Order of this chapter

1. From the box to all of space — what the infinite-volume limit is
2. A limit can always be taken, but it need not be unique (Figure 1, Table 1)
3. Detour: Elitzur's theorem — quantities that change under gauge transformations are 0 in every limit
4. Splitting into lumps — only closed lumps survive
5. Counting lumps — their number grows exponentially, and their weight falls exponentially (Table 2)

6. The condition for convergence — Kotecký–Preiss (Figure 2, Table 3)
7. What comes out of the expansion — the limit, analyticity, the lattice gap
8. Another road — Langevin dynamics and the Poincaré inequality
9. Where it stops — the strong-coupling shore and the continuum-limit shore (Table 4)

01

From the box to all of space — what the infinite-volume limit is

Lattice gauge theory is first defined on a box Λ_L of side L. The weight of the Wilson action (5-01) together with the Haar measure is an ordinary probability measure on a compact space.

$$\langle F\rangle_L = \frac{1}{Z_L}\int F(U)\, e^{-S_W(U)} \prod_{\ell\in\Lambda_L} dU_\ell, \qquad S_W = \beta\sum_P \Big(1 - \tfrac{1}{N}\operatorname{Re}\operatorname{tr} U_P\Big) \quad (1)$$

The mass gap is a statement about distances stretched to infinity (4-11 § 02), so we consider the limit of larger and larger boxes. We call a continuous function determined by finitely many links a **local observable**. When a probability measure μ on the whole lattice satisfies

$$L_k \to \infty \text{ along a sequence, and for every local observable } F \quad \langle F\rangle_{L_k} \longrightarrow \int F\, d\mu \quad (2)$$

we call μ an **infinite-volume limit**. It is weak convergence of measures. Chatterjee (2018) § 3 writes "a weak limit of the above probability measures as $\Lambda \uparrow \mathbb{Z}^n$" Known .

Example: in two dimensions with free boundary, fixing the gauge along a tree makes the plaquette variables independent, so in every box $\langle W_{R\times T}\rangle = u^{RT}$ (5-02 § 04), and the limit is trivial Known .

02

A limit can always be taken, but it need not be unique

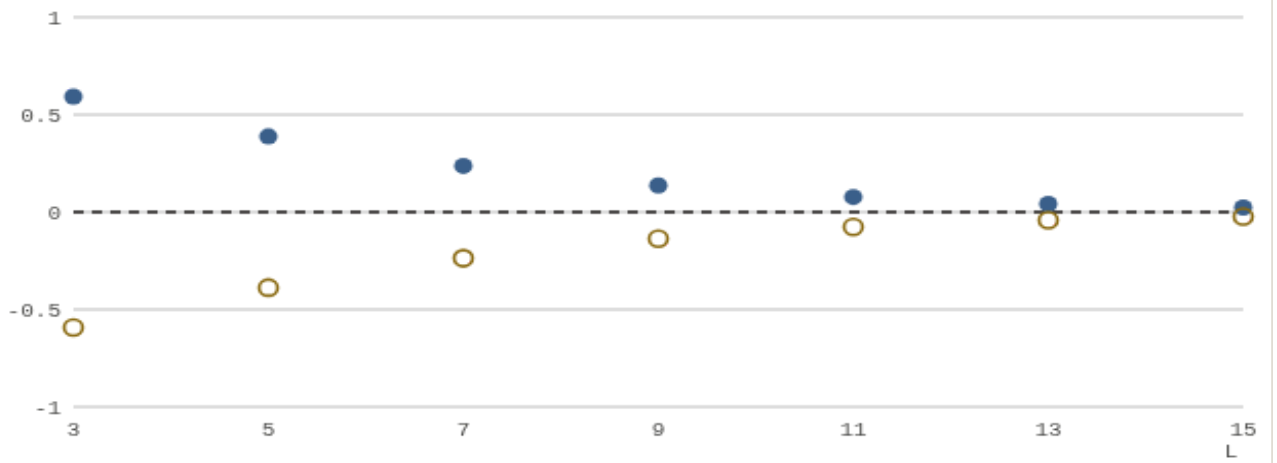
Existence comes almost free from compactness. $\langle F\rangle_L$ is a bounded sequence, so it has a convergent subsequence, and by a diagonal argument it can be made to converge simultaneously for countably many functions that approximate the local observables Known .

The hard part is **uniqueness**. Chatterjee (2018) § 3 writes “The infinite volume limit may or may not be unique. Indeed, the uniqueness (or non-uniqueness) is in general unknown for lattice gauge theories in dimensions higher than two when β is large.” Known . We look at an example where it fails in the two-dimensional Ising model (2-09). We fix the outer boundary of an $L \times L$ box to ± 1 and measure the mean of the central spin.

$$\langle \sigma 0 \rangle_L^{\pm} \xrightarrow{L\to\infty} \pm m^*, \qquad m^* = \begin{cases} 0 & \beta \le \beta_c \\ \left(1 - \sinh^{-4} 2\beta\right)^{1/8} & \beta > \beta_c \end{cases}, \qquad \beta_c = \tfrac{1}{2}\ln(1+\sqrt{2})$$

The right-hand side is the exact formula for the spontaneous magnetisation Known . At $\beta > \beta_c$ a different limit appears for each boundary condition (a phase transition; 2-08).

Figure 1 The central spin of a box whose outer boundary is + (●) or − (○) ($L = 3, \dots, 15$); the dashed lines are $\pm m^*$. The knob, "Play" and the rows of Table 1 change β.



β = 0.30, L=15: ±0.023, m* = 0.000

Computed The numbers in this figure were computed on this machine (column transfer matrix). If JavaScript is disabled, Table 1 holds the content.

Table 1 $\langle \sigma 0 \rangle_L^{\pm}$ (with the opposite sign for the − boundary). Selecting a row changes β in Figure 1 Computed

β	$L = 5$	$L = 15$	m^*	**Reading**
0.30	**0.3882**	**0.0231**	**0**	**The influence of the boundary dies away exponentially**
0.44	0.8473	0.7544	0	Near the critical point. The boxes are small
0.50	0.9241	0.9119	0.9113	Splits into $\pm m^*$

The distance between the centre and the outer boundary is $(L+1)/2$, and at $\beta = 0.2$, $\ln(\langle\sigma 0\rangle 13 / \langle\sigma 0\rangle 15) = 1.219$ is almost the same as the decay rate of correlations $2(K^* - K) = 1.223$ (2-09 § 06) Computed . **The influence of the boundary dies away as fast as correlations do.** The tool that shows decay of correlations in gauge theory is the expansion in lumps from § 04 on.

03

Detour: Elitzur's theorem — quantities that change under gauge transformations are 0 in every limit

A section you may skip.

In gauge theory the boundary cannot pick a direction. If we make a gauge transformation g at a single point x inside the box (outgoing links $gU\ell$, incoming links $U\ell g^{-1}$), neither S_W nor the Haar measure nor the boundary changes, so we can average over g, and for a link ℓ leaving x

$$\langle \operatorname{tr} U\ell \rangle_L = \left\langle \operatorname{tr}\left(\textstyle\int_G g\, dg\, U\ell \right) \right\rangle_L = 0 \quad (4)$$

holds in every box, for every boundary condition, and hence in every limit. $\int_G g\, dg = 0$ follows from the fact that multiplying by a central element ωI ($\omega = e^{2\pi i/N}$ for $SU(N)$, -1 for $U(1)$ and $\mathbb{Z}_2$, and so on) does not change the Haar measure, so $\int g\, dg = \omega \int g\, dg$. The abstract of Elitzur (1975) is "It is argued that a spontaneous breaking of local symmetry for a symmetrical gauge theory without gauge fixing is impossible." Known . Precisely, what can be said is that "quantities that change under gauge transformations are not markers of order", not that there is no phase transition. Differences of phase can appear in gauge-invariant quantities (the area law, the Polyakov loop; 18-03) (5-06).

04

Splitting into lumps — a gas of lumps of plaquettes

At small β the weight of each plaquette is almost constant. We split the weight as "constant $\times$ (1 + small deviation)" and expand the product.

$$e^{\frac{\beta}{N}\operatorname{Re}\operatorname{tr} U} = c_0(\beta)\big(1 + f(U)\big), \qquad c_0 = \int_G e^{\frac{\beta}{N}\operatorname{Re}\operatorname{tr} U}\, dU, \qquad \int_G f\, dU = 0 \quad (5)$$

We split the set X of plaquettes that pick up the deviation into connected parts by the rule **they are connected if they share a link**. The integrals over parts that share no link are independent, so they factorise.

$$\frac{Z_L}{(e^{-\beta} c_0)^{|\mathcal{P}|}} = \sum_{\substack{\{X_1,\dots,X_k\} \\ \text{sharing no link pairwise}}} \prod_{i=1}^{k} z(X_i), \qquad z(X) = \int \prod_{P \in X} f(U_P) \prod_{\ell \in X} dU_\ell \quad (6)$$

$|\mathcal{P}|$ is the number of plaquettes in the box. This is exactly the polymer model of 7-04 § 03 Known . Two facts about the activities matter.

$$|z(X)| \le \varepsilon^{|X|}, \quad \varepsilon(\beta) = \sup_U |f(U)|; \qquad z(X) = 0 \ \text{(if some link lies in only one plaquette of}$$

The right-hand statement holds because that link appears in only one $f(AU_\ell B)$, and by the invariance of the Haar measure $\int f(AU_\ell B)\,dU_\ell = \int f = 0$. What survives are only **closed lumps**, in which every link is surrounded by two or more plaquettes. The smallest is the surface of a unit cube (6 plaquettes).

Example: for $\mathbb{Z}_2$, $f = \tanh\beta \cdot U$, and the activity of the cube is $\tanh^6\beta$. For $SU(2)$ we expand each face as $f = \sum_{j\,/=0} d_j\, u_j\, \chi_j$ by (3) of 5-02 and integrate by (5) of 5-02; only the terms in which every face carries the same representation j survive. For a closed surface the rule is known that its contribution is $d_j^{F-E+V}\, u_j^F$ (F, E, V are the numbers of faces, edges and vertices, $F - E + V$ is the Euler characteristic of the surface, and the surface of a cube is a sphere, so $6 - 12 + 8 = 2$) Known . For the cube this can be checked as follows. Gluing adjacent faces 5 times by the formula on the right of (5) ($1/d_j$ each time) turns the 6 faces into a single disc. In the product around its rim, the remaining 7 links cancel in order, starting with those adjacent as in $U_\ell\, U_{\bar\ell}^{-1}$, and the identity matrix remains. Hence $d_j^6 \cdot d_j^{-5} \cdot \chi_j(1) = d_j^2$.

$$z(\text{cube}) = \sum_{j=1/2,1,\dots} (2j+1)^2\, u_j^6 \approx 4u^6, \qquad u_j = \frac{I_{2j+1}(\beta)}{I_1(\beta)} \tag{8}$$

At $\beta = 1$, 7.682×10^{-4}, and direct Monte Carlo integration gives $7.63(8) \times 10^{-4}$ Computed .

05

Counting lumps — their number grows exponentially, and their weight falls exponentially

We fix four words (for details, 7-04 § 03 and § 04). We call the X of (6) **polymers** (lumps). Two lumps that share a link are called **incompatible**. A sequence of lumps joined into one by the incompatibility relation (the same lump may appear any number of times) is called a **cluster**. The logarithm of the left-hand side of (6) can be rewritten as a sum of terms, one for each cluster, and the question is whether that sum converges absolutely independently of the box Known .

Convergence is a contest between **how fast the number** of lumps grows and the weight ε^n. Let c_n be the number of sets of n plaquettes, connected through links, containing a root plaquette. A plaquette has 4 links and a link is contained in $2(d-1)$ plaquettes, so the plaquettes that share a link with it number $D = 4(2(d-1)-1) = 4(2d-3)$ (20 in four dimensions). There is the upper bound of Lemma 2.1(c) of Borgs–Chayes–Kahn–Lovász (2010) Known .

$$c_n \le \frac{1}{n}\binom{nD}{n-1}, \qquad \lim_{n\to\infty}\left(\frac{1}{n}\binom{nD}{n-1}\right)^{1/n} = \frac{D^D}{(D-1)^{D-1}} \approx 53.0 \quad (D=20)$$ (9)

Table 2 c_n in four dimensions (enumerated by Redelmeier's method) Computed

n	c_n	Ratio	Closed lumps	Upper bound (9)
2	20	20.0	0	20
3	458	22.9	0	590
4	11 132	24.3	0	20 540
5	280 175	25.2	0	784 245
6	7 218 768	25.8	4	31 763 004

The ratio is about 26 at $n = 6$, and the growth rate 53 of the upper bound is about twice that. There are no closed lumps for $n \le 5$, and the 4 at $n = 6$ are four cubes Computed . To guarantee convergence with this way of counting, ε must be smaller than the reciprocal of the growth rate of lumps (a value expected to be smaller than $1/26$ from the ratios in Table 2, and larger than $1/53$ from (9)).

As a metaphor, for each plaquette it is a contest between a cost of $\ln(1/\varepsilon)$ and a gain of about $\ln 26$ from the number of choices. Precisely, if the gain wins, the expansion merely becomes unusable; a phase transition need not occur (7-04 § 10).

06

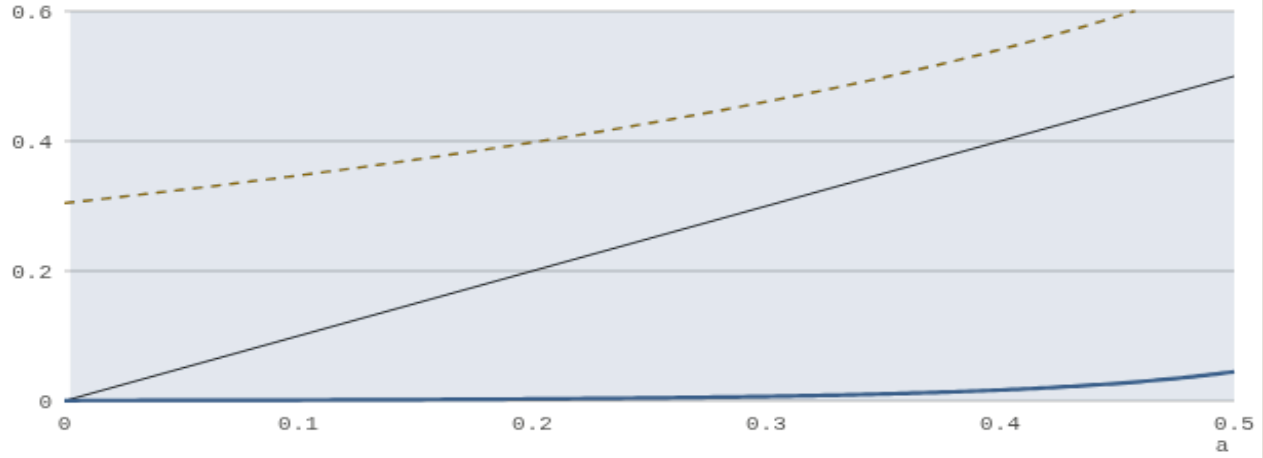
The condition for convergence — applying the Kotecký–Preiss condition

The Kotecký–Preiss condition turns the contest into a theorem (7-04 § 06). In the form of a theorem: if there is some $a > 0$ such that the left-hand inequality of (10) holds for every lump X, then the sum over clusters for the logarithm in (6) converges absolutely independently of the box Known . One reads it as giving each lump a budget $a|X|$, and requiring the total weight of its incompatible partners (multiplied by $e^{a|X'|}$) to fit within the budget.

$$\sum_{X' \text{ sharing a link with } X} |z(X')|\, e^{a|X'|} \le a|X| \quad \Longleftarrow \quad 8(d-1)\, G\big(\varepsilon e^{a}\big) \le a, \qquad G(x) = \sum_{1}$$

The right-hand side is the count that X has at most $4|X|$ links, and that a lump containing a given link contains one of its $2(d-1)$ plaquettes. c'_n is the number of closed lumps (Table 2 for $n \le 6$, (9) for $n \ge 7$). Maximising $\varepsilon = x\, e^{-8(d-1)G(x)}$ over $x = \varepsilon e^{a}$ gives the upper limit $\varepsilon_{\max}$ of the guarantee, and solving $\varepsilon(\beta) = \varepsilon_{\max}$ gives the range of β.

Figure 2 Condition (10) in four dimensions. The thin line is the right-hand side a, the thick line the left-hand side $24G(\varepsilon e^{a})$ (closed lumps), and the dashed line all connected lumps. If there is a band where the thick line lies below the thin line, convergence is guaranteed. The knob changes β and the buttons the group.



SU(2), β = 0.0100, ε = 0.0100, a satisfying the condition: 0.003–0.500

Computed The numbers in this figure were computed on this machine. If JavaScript is disabled, Table 3 holds the content.

Table 3 The upper limit of β guaranteed by (10) (0.0087 when all connected lumps are counted) Computed , and the condition of Shen–Zhu–Zhu converted to the Wilson β of this book Known (conversion of conventions in 5-01 Table 4)

Four dimensions	$\varepsilon_{\max}$	$\mathbb{Z}_2$	$U(1)$	$SU(2)$	$SU(3)$
Closed lumps	0.0161	0.0161	0.0160	0.0160	0.0160
Shen–Zhu–Zhu Known	–	–	–	0.0833	0.1875

What is guaranteed is about $\beta < 0.016$, and since $\varepsilon \approx \beta$ it hardly depends on the group. This is two orders of magnitude narrower than $\beta \leq 1$, where the strong-coupling formula agreed with the numbers in 5-02 Table 2. The sources are the supremum norm (for $SU(2)$ at $\beta = 0.1$, $\varepsilon = 0.104$, while $u = 0.025$), the crudeness of (9), and ignoring cancellations Computed . Table 3 is not "where the expansion stops".

07

What comes out of the expansion — the limit, analyticity, exponential clustering, the lattice gap

When (10) holds independently of the box, the following come out of 7-04 § 08 Known .

(a) Analyticity. The limit of $\ln Z_L$ per plaquette is an analytic function of β (it converges even when extended to complex β), and there is no phase transition in that range. **(b) Existence of the limit.** $\langle F \rangle_L$ converges without taking a subsequence (the influence of the boundary dies away exponentially, as in Table 1). What can be said is the existence of the limit along the sequence with that boundary condition; uniqueness independent of the boundary condition relies on the road of § 08. **(c) Exponential clustering.** In the connected correlation of two local observables only

clusters touching both supports survive, and joining a distance r needs about r plaquettes or more.

$$\big|\langle FG\rangle - \langle F\rangle\langle G\rangle\big| \le C_{F,G}\, e^{-m\,\mathrm{dist}(F,G)}, \qquad m>0 \quad (11)$$

Example: the smallest lump joining two plaquettes is a tube with $4t$ side plaquettes having those two as its ends (a closed surface together with the ends), and the leading term is $m_0 = -4\ln u$ (5-02 § 05). (11) justifies this independently of the box.

(d) The lattice gap. We normalise the transfer matrix T (5-03 § 05) so that its eigenvalue on the vacuum Ω is 1. For the state $F\Omega$ made by a gauge-invariant local observable F, with its vacuum component subtracted, (11) in the time direction says that $\langle F\Omega, T^t F\Omega\rangle$ decreases at the rate e^{-mt} for every F. Such states are dense in the space of all gauge-invariant states orthogonal to the vacuum, so by the spectral theorem the spectrum of T there lies in $[0, e^{-m}]$ (4-11 § 02). That is, $\|T\| \le e^{-m}$ on the gauge-invariant states orthogonal to the vacuum, and the gap of $-\ln T$ is at least m.

In the literature, Shen–Zhu–Zhu (2023) § 1 summarises the content of Osterwalder–Seiler (1978) as "the existence of a mass gap (exponential clustering) follows from convergence of the cluster expansion, so do existence of the infinite volume limit and analyticity of Schwinger functions in the inverse coupling.", and writes of uniqueness only weakly, that it "should also follow". Chatterjee (2018) § 6 writes that "quark confinement and the existence of a positive self-adjoint transfer matrix" were shown Known . The abstract of the original paper also says "We then prove existence and analyticity of the infinite volume limit of strongly coupled Yang-Mills theories on the lattice and we verify Wilson's confinement bound." (we have not read the text).

The forms machine-checked on this book's definitions are in 20-01 to 20-04.

08

Another road — Langevin dynamics and the Poincaré inequality

There is also a road that gets there without counting lumps. We prepare three words.

Three words — Langevin dynamics, Ricci curvature, the Hessian

Langevin dynamics is a dynamics that moves the link variables randomly in the form $dQ = -\nabla S_W(Q)\,dt + \sqrt{2}\,dB$. ∇ is the gradient on the group of the link variables, and B is Brownian motion on the group (an independent fluctuation on each link). It is built so that the balance between the force down the slope and the fluctuations leaves the measure (1) unchanged (has it as invariant measure) (7-09). **Ricci curvature** is how curved the group is, seen as a curved surface. Where it is positively curved like a sphere, geodesics leaving nearby points gather together. For $SU(N)$ it is the same

positive constant everywhere ($N/2$ in the normalisation of the inner product of Shen–Zhu–Zhu). The **Hessian** is the second derivative of the action S_W taken on the group.

The Poincaré inequality and the Bakry–Émery criterion

The key is the Poincaré inequality (1-10).

$$\mathrm{Var}_\mu(F) \le C_P\, \mathbb{E}_\mu\big[\,|\nabla F|^2\,\big] \quad (12)$$

That C_P does not depend on the box is what "a uniform estimate" means. In (9) of 1-10, in one dimension $V'' \ge \kappa > 0$ gave $C_P \le 1/\kappa$. On a group the curvature adds to this κ, and the Bakry–Émery criterion takes the following form Known .

$$\text{Ricci curvature} + \text{Hessian of the action} \;\ge\; K_S > 0 \quad\Longrightarrow\quad (12)\text{ holds with a box-independe}$$

The strong-coupling condition of Shen–Zhu–Zhu (Shen–Zhu–Zhu's Assumption 1.1) is $K_S = N/2 - 8N|\beta_{\mathrm{SZZ}}|(d-1) > 0$ with $SU(N)$. The term on the right comes from the estimate that the Hessian does not fall below $-8N|\beta_{\mathrm{SZZ}}|(d-1)$. The point is that the Hessian stays of order $|\beta|N$ (an example with one link is in 5-11). As a metaphor, it is the state in which the force pulling back to the bottom of a round bowl is stronger than the bumps of the action. Precisely, what is compared is the curvature $N/2$ with the size $8N|\beta_{\mathrm{SZZ}}|(d-1)$ by which the Hessian can become negative.

The theorem of Shen–Zhu–Zhu

Shen–Zhu–Zhu (2023), under $|\beta_{\mathrm{SZZ}}| < 1/(16(d-1))$ with $SU(N)$ ($\beta = N^2\beta_{\mathrm{SZZ}}$; conversion in 5-01 Table 4), write in the abstract “The finite volume Yang–Mills measures converge to this unique invariant measure in the infinite volume limit” and “the infinite volume measure has a strictly positive mass gap”. Their uniqueness theorem for the invariant measure (Shen–Zhu–Zhu's Theorem 1.2) shows that the invariant measure of the Langevin dynamics is unique (1), and that the limit of the measures on periodic boxes agrees with it independently of the subsequence (2). Since the group is compact, the sequence of measures on boxes is always tight, and the “every tight limit” of (2) may be read as a subsequential limit in the sense of § 02. Remark 1.3 states that replacing the periodic boundary by Dirichlet or other boundary conditions gives the same invariant measure in the limit. It is one answer, in the strong-coupling range, to the question of § 02 Known . For four-dimensional $SU(2)$ this is $\beta < 1/12$.

Note Two gaps

The $1/C_P$ of (12) is the gap in the artificial Langevin time, a different quantity from the gap of the transfer matrix (18-03). The mass gap of Shen–Zhu–Zhu is the

exponential decay of correlations in distance on the lattice (not in the artificial Langevin time) derived from (12) (Shen–Zhu–Zhu's Corollary 1.6), a statement in the same sense as (11).

09

Where it stops — the strong-coupling shore and the continuum-limit shore

Table 4 The scale of β for four-dimensional $SU(2)$ (action (1))

β	What is said	Level
< 0.016	The fit in this chapter (Table 3)	Computed
< 0.083	A unique limit and exponential decay of correlations (Shen–Zhu–Zhu)	Known
$\lesssim 1$	The strong-coupling formula agrees with the numbers (5-02 Table 2)	Computed
≈ 2	Roughening of the surface (5-02 § 07)	Physics
2.3 and above	Numbers extrapolated to the continuum limit (5-08)	Physics
$\to \infty$	The continuum limit (5-05). A limit of correlation functions was built by the construction published in 2026-11 (along a subsequence of volumes; 6-00). There is no theorem on uniqueness or on a gap in physical units (6-04)	The committee's assessment

The meaning of "infinite-volume limit" changes between the top and the bottom of the table. The limit in this chapter is at fixed lattice spacing. In the continuum limit, $a \to 0$ while the box in lattice units also goes to infinity, and the gap in lattice units should close as $m(\beta) \to 0$ (5-09) Physics . What is needed is a gap uniform in physical units, and there is no small ε there. The problem statement § 6 writes “no present ideas point the direction to establish the existence of a mass gap that is uniform in the volume. Nor do present methods suggest how to obtain the existence of the infinite volume limit $T^4 \to R^4$.” Known .

Also, analyticity says nothing outside the small disc of complex β guaranteed by (10). For four-dimensional $U(1)$ the area law fails at large β (Guth 1980; according to Chatterjee (2018) § 4, the fully rigorous proof is Fröhlich–Spencer 1982; 5-06), while that non-abelian groups have no critical point at finite β (the correlation length diverges only as $\beta \to \infty$) is at the level of Chatterjee (2018) § 5, “It is believed that ... many of the non-Abelian lattice models of interest have βc = ∞.” Physics (7-10).

What this chapter can say

Matter	Said	Not said
(2)–(8)	Existence of subsequential limits, Elitzur, the gas of lumps Known . Table 1 and the activity of the cube Computed	Uniqueness at large β
Tables 2 and 3	The numbers for $n \le 6$, and the crude sufficient condition (10) Computed	The best range in the literature, the β at which the expansion stops
(a)–(d), (11), (12)	An account of known results on the strong-coupling side (Osterwalder–Seiler from the abstract and secondary sources). Uniqueness independent of the boundary condition is, from the expansion in lumps, at the level of "should also follow"; Shen–Zhu–Zhu showed it along the road of Langevin dynamics Known	The details of the proofs. Uniqueness and the gap in the continuum limit. This chapter shows nothing about them

Sources and reproduction

Item	Kind	Source / tools
Quotations in § 01, § 02, § 07 and § 09	Text § 3, § 4, § 5, § 6	S. Chatterjee, arXiv:1803.01950 (2018)
Quotations in § 07 and § 08, Assumption 1.1, Theorem 1.2, Remark 1.3, Theorem 1.4, Corollary 1.6, formulas (1.2), (4.7), (4.8)	Abstract, text § 1 and § 4	H. Shen, R. Zhu, X. Zhu, Commun. Math. Phys. 400 (2023) 805–851, arXiv:2204.12737
Osterwalder–Seiler in § 07	Abstract (positivity of the transfer matrix, existence and analyticity of the infinite-volume limit at strong coupling, Wilson's confinement bound; the INSPIRE abstract). Bibliographic data only for the text	K. Osterwalder, E. Seiler, Ann. Phys. 110 (1978) 440–471
Original source (including (10))	Bibliographic data only (content from the two above and 7-04)	R. Kotecký, D. Preiss, CMP 103 (1986) 491. E. Seiler, *Gauge theories as a problem of constructive quantum field theory and statistical mechanics*, Springer (1982)
Quotation in § 03	Abstract	S. Elitzur, Phys. Rev. D 12 (1975) 3978–3982
(9)	Text, Lemma 2.1(c)	C. Borgs, J. Chayes, J. Kahn, L. Lovász, arXiv:1002.0115 (2010)
Enumeration method of Table 2	Bibliographic data only	D. H. Redelmeier, Counting polyominoes: yet another attack, Discrete Math. 36 (1981) 191–203

Item	Kind	Source / tools
The rule d^{F-E+V} in (8)	Text § 3.4 (the gluing formulas (3.17) and (3.19), and that $(\dim R)^{2-2p}$ appears for a closed surface; in the form of the two-dimensional theory)	S. Cordes, G. Moore, S. Ramgoolam, Lectures on 2D Yang–Mills theory, equivariant cohomology and topological field theories, arXiv:hep-th/9411210 (1994)
The problem statement in § 09	Text § 6 (checked word for word in the material)	A. Jaffe, E. Witten, Quantum Yang–Mills Theory, Clay Mathematics Institute (2000)
The m^* of (3) / the phases of $U(1)$	Bibliographic data only	C. N. Yang, Phys. Rev. 85 (1952) 808–816 / A. H. Guth, Phys. Rev. D 21 (1980) 2291, J. Fröhlich, T. Spencer, Commun. Math. Phys. 83 (1982) 411
Tables 1–3, the figures, (8)	Computed on this machine	Python 3, NumPy, SciPy: ising_box.py, animals.py, kp.py, cube.py, cube_check.py

Next chapter to read: 5-05 Weak coupling and the continuum limit (the opposite shore). For the general theory, 7-04.

Weak coupling and the continuum limit — raising β makes the lattice finer, and asymptotic freedom sets how fast

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — Raising β brings the lattice spacing $a(\beta)$ towards 0, and a is an output, fixed by calibrating against a physical quantity. The difference between "scaling" and "asymptotic scaling". Lattice errors vanish as a^2 and can be reduced by improved actions and continuum extrapolation.

Prerequisite chapters — 5-01 (Wilson's lattice) and 4-06 (the β function). For the general theory of the continuum limit, 2-08 § 08; for the mass gap, 4-11. The β of this chapter is the inverse lattice coupling $2N/g_0^2$ of 5-01, a different quantity from the β function $\beta(g)$ of 4-06. The "lattice β function" of § 03 refers to how this β depends on a, (3).

Said first — The asymptotic form of $a(\beta)$ and the a^2 errors are matters at the level of perturbation theory and numerics Physics . The continuum limit of the four-dimensional non-abelian lattice was built by the construction published in 2026-11, for a family whose bare coupling is chosen to match the running of asymptotic freedom (the committee's assessment is that it is proved in Lean; 6-00). A limit in the form with the scale set by a physical quantity, the limit of the ratios (4), and the mass gap have not been shown (6-04). This chapter too shows nothing about the continuum limit or the mass gap.

Order of this chapter

1. The continuum limit is approaching a critical point
2. The lattice spacing is an output — what to calibrate with
3. The lattice β function and asymptotic scaling — the $\Delta\beta$ that halves a (Table 1, Figure 1). Detour: the price of the continuum limit
4. Scaling — dimensionless ratios stand still

5. The form of lattice errors — a^2 seen in a free field (Table 2, Figure 2)
6. Symanzik's effective theory and improved actions — adding rectangles to remove a^2 (Table 3), logarithmic corrections in the quantum theory. Detour: the price of improvement
7. Continuum extrapolation — drawing a straight line against a^2
8. The continuum limit seen from mathematics — what has been shown and what has not

---

01

The continuum limit is approaching a critical point

A lattice theory has no unit of length; what comes out are numbers of the form "so many lattice spacings". For example, the correlation length $\xi_{\text{lat}}(\beta)$ when the correlation of two plaquettes decreases as $e^{-|x|/\xi_{\text{lat}}}$ (4-11). A mass in physical units is obtained by dividing by the lattice spacing.

$$\Delta = \frac{1}{a(\beta)\,\xi_{\text{lat}}(\beta)} \quad (1)$$

To hold Δ fixed and take $a \to 0$ on the lattice with $\beta = 2N/g_0^2$ (5-01), we need $\xi_{\text{lat}} \to \infty$, that is, to approach a critical point. Chatterjee (2018) § 5 summarises the physical usage as “the model has a continuum limit if there is a critical point βc ∈ [0,∞] such that as β → βc, the correlation length tends to infinity”, and for four-dimensional non-abelian models writes “It is believed … βc = ∞” Known . $\beta = \infty$ is $g_0 \to 0$, the same direction as asymptotic freedom (4-06).

There are two limits: first hold β fixed and make the box infinite (5-04), and afterwards take $\beta \to \infty$.

02

The lattice spacing is an output — what to calibrate with

How many fm a is, is decided afterwards by matching the lattice value $\widehat{Q} = a^k Q$ of a quantity Q of mass dimension k against its physical value.

$$aQ(\beta) = \left(\widehat{Q}(\beta)/Q_{\text{phys}}\right)^{1/k} \quad (2)$$

As standards one uses the string tension σ (Creutz 1980 fixed a by “hold fixed the string tension”), the length r_0 defined from the force, and t_0 of the gradient flow (Lüscher

2010) Known . r_0 is the length fixed by $r_0^2 F(r_0) = 1.65$ from the force $F(r)$ acting between a pair of static charges (Sommer 1994; 4-06 § 07). t_0 is a length fixed by smoothing the field with a diffusion equation (the gradient flow) (18-04 § 08 calls the same thing the Wilson flow).

Example: for pure SU(3), r_0/a is 2.922 at $\beta = 5.7$ and 9.74 at 6.4 (Table 1 of Necco–Sommer 2002; the values for $\beta \le 6.4$ are reproduced from earlier calculations). Setting $r_0 \approx 0.5$ fm (a reference value used in QCD; the introduction of Necco–Sommer) gives $a \approx 0.171$ fm and 0.051 fm (1 fm = 10^{-15} m) Computed . The physical value of r_0 is an input from outside the lattice.

Changing the standard changes $a(\beta)$ slightly at finite β, and the ratio of the two a is fixed by the ratio $\widehat{Q}_1/\widehat{Q}_2$ measured on the lattice. That this ratio settles down is the scaling of § 04. In the mathematical formulation (§ 08) too the choice of calibration is an issue, and the existence of $\varepsilon(\beta)$ in either convention is unproved (5-09).

03

The lattice β function and asymptotic scaling

To keep the same physics as the lattice is made finer, the bare coupling g_0 is weakened. How fast begins with the scheme-independent b_0, b_1 (4-06 § 05) Known . Reading $\mu = 1/a$ and rewriting in terms of $\beta = 2N/g_0^2$ gives

$$\frac{d\beta}{d\ln(1/a)} = 4Nb_0 + \frac{8N^2 b_1}{\beta} + O(\beta^{-2}),$$
$$a\Lambda_{\mathrm{L}} = (b_0 g_0^2)^{-b_1/(2b_0^2)} e^{-1/(2b_0 g_0^2)} \left(1 + O(g_0^2)\right) \quad (3)$$

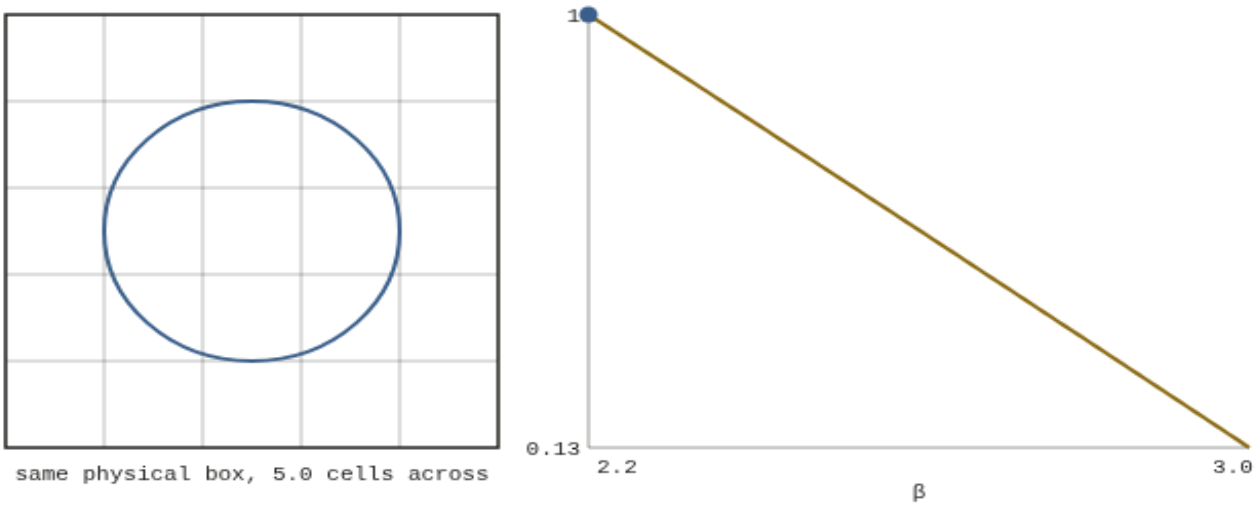
and the lower formula is formula (8) of 4-06, obtained by integrating the upper one. The constant of integration Λ_{L} (4-07) does not come out, but the ratio $a(\beta)/a(\beta')$ does. At one loop $a \propto e^{-\beta/(4Nb_0)}$, and the step that halves a is $\Delta\beta = 4Nb_0 \ln 2$: 0.2575 for SU(2) from $4Nb_0 = 11/(3\pi^2)$, and 0.5794 for SU(3) Computed .

Table 1 $a(\beta)$ from formula (3) (two loops) Computed

Group	β	a/a_{ref}	Lattice points in the same box ×	$\Delta\beta$ to halve
SU(2)	2.2	1	1	0.2760
SU(2)	2.6	0.3656	56.0	0.2732
SU(2)	3.0	0.1323	3263.5	0.2711
SU(3)	5.7	1	1	0.6155
SU(3)	6.3	0.5088	14.9	0.6121

Group	β	a/a_{ref}	Lattice points in the same box ×	$\Delta\beta$ to halve
SU(3)	6.9	0.2579	225.9	0.6092

Figure 1 Raising β makes the lattice in the same physical box (left) finer according to formula (3), and more lattice cells span the correlation length (circle). The right shows the logarithm of a/a_{ref}.



SU(2) β = 2.20, a/a(2.2) = 1.0000, ξ in lattice units ×1.00, lattice points ×1.0, Δβ to halve a = 0.2760

Computed The numbers in this figure were computed on this machine (formula (3) computed on the spot; agrees with Table 1 to 4 digits. If JavaScript is disabled, Table 1).

That lattice numbers follow (3) is called **asymptotic scaling**. Creutz (1980) saw for SU(2) "a logarithmic decrease of the bare coupling constant in a manner consistent with the perturbative renormalization-group prediction" Known . The step in pure SU(3) data is 0.354 at $\beta = 5.70$ and 0.524 at 6.20, short of the two-loop 0.61 (Table 4 of 4-06) Physics . The abstract of Lepage–Mackenzie (1993) says that expanding in a coupling defined by a physical quantity makes perturbation theory and numerics agree well. The abstract also says that if that coupling is used instead of the bare coupling, scaling already follows perturbation theory from about $\beta = 5.7$. Much of the deviation seen with the bare coupling comes from the choice of coupling.

Detour: the price of the continuum limit

A subsection you may skip.

Holding the box fixed and halving a multiplies the lattice points by 16 in four dimensions. For SU(2), raising β from 2.2 to 3.0 multiplies them by about 3,300 (Table 1). As a metaphor, it is "a picture where each doubling of resolution means 16 times the pixels, and each stroke gets slower too". Precisely, the cost is the number of lattice points times the autocorrelation time, and the latter is also expected to grow as a power of ξ_{lat} (5-07) Physics .

04

Scaling — dimensionless ratios stand still

What the continuum limit needs is that **ratios** of physical quantities stop depending on β.

$$R(\beta) = \widehat{Q}_1(\beta)/\widehat{Q}_2(\beta) = R_0\left(1 + s\,a^2 Q_2^2 + \cdots\right) \quad (4)$$

being constant within errors is called **scaling**, and that range of β the scaling window. Example: Necco–Sommer (2002) write, for the ratio r_c/r_0 of two lengths defined by the force, "No dependence on the resolution is seen within the errors of below 1%", and obtain 0.5133(24) by extrapolation Known . $m/\sqrt{\sigma}$ is looked at in 5-08.

Scaling is ratios standing still, and is necessary for the continuum limit. **Asymptotic scaling** is $a(\beta)$ itself following (3), an additional claim tying the lattice to the Λ of perturbation theory Physics . If two quantities deviate from (3) by the same amount, the deviation cancels in the ratio. Lattice results such as $m/\sqrt{\sigma}$ rely on scaling (18-04 § 05).

05

The form of lattice errors — a² seen in a free field

We see why the correction in (4) starts at a^2 in a free field, which can be solved exactly. In the symmetric difference $f(x+a) - 2f(x) + f(x-a)$, odd powers of a cancel. For a free scalar field of mass m, the transfer matrix (5-03) fixes the mass E seen on the lattice Known . Concretely, looking for the pole of the lattice propagator $1/(\hat{p}^2 + m^2)$ ($\hat{p}^2$ is formula (6) below) at zero spatial momentum and imaginary momentum $p = iE$ gives $\frac{4}{a^2}\sinh^2\frac{aE}{2} = m^2$, and rewriting this gives the upper formula below.

$$\cosh(aE) = 1 + \tfrac{1}{2}(am)^2,$$
$$\frac{E}{m} = 1 - \frac{(am)^2}{24} + \frac{3(am)^4}{640} - \cdots \quad (5)$$

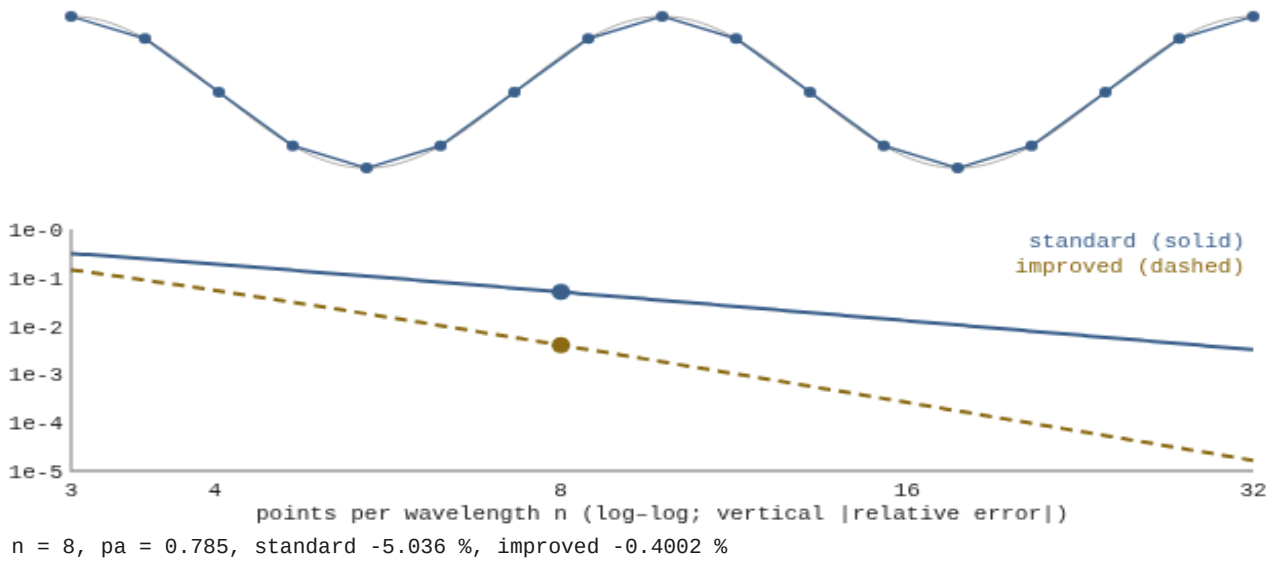
Table 2 The lattice mass (5) of the free field and the two-point extrapolation (8) Computed

am	$(E/m-1)/(am)^2$	Extrapolation error
1.0	−0.03758	-9.9×10^{-4}
0.4	−0.04093	-2.9×10^{-5}
0.2	−0.04148	-1.9×10^{-6}
0.1	−0.04162	-1.2×10^{-7}

The second column approaches $-1/24 = -0.04167$. The same holds in the spatial directions: applying the lattice Laplacian to a wave of wave number p gives $\hat{p}^2$ instead of p^2. Building a difference that uses points two steps away, with weights $\frac{4}{3}$, $-\frac{1}{3}$, removes the a^2 term. This is the simplest example of improvement.

$$\hat{p}^2 = \frac{4}{a^2}\sin^2\frac{pa}{2} = p^2\left(1 - \frac{(pa)^2}{12} + \cdots\right),$$
$$\hat{p}^2_{\text{imp}} = \frac{4}{3}\hat{p}^2 - \frac{1}{3a^2}\sin^2(pa) = p^2\left(1 - \frac{(pa)^4}{90} + \cdots\right) \quad (6)$$

Figure 2 Changing the number n of lattice points per wavelength moves the sample points above and the positions of the errors of the standard (solid) and improved (dashed) differences below (log–log).



Computed The numbers in this figure were computed on this machine ((6) computed on the spot). If JavaScript is disabled: $\hat{p}^2/p^2 - 1$ at $n = 4, 8, 16, 32$ is −0.189, −0.0504, −0.0128, −0.00321 for the standard difference and −0.0543, −0.00400, −0.000261, −0.0000165 for the improved one. Doubling n makes the standard error about 1/4 and the improved one about 1/16.

06

Symanzik's effective theory and improved actions

Symanzik (1983) read a lattice theory at small a as "the continuum theory plus terms with powers of a", and says in the abstract “Corrections to continuum theory results stemming from finite lattice spacing can be diminished systematically by use of lattice actions that also include suitable irrelevant terms” Known . For pure gauge theory it is thought to take the following form Physics .

$$\mathcal{L}_{\text{eff}} = \frac{1}{2}\operatorname{tr} F_{\mu\nu}F_{\mu\nu} + a^2\sum_{i=1}^{3} c_i(g_0)\,\mathcal{O}_i + O(a^4) \quad (7)$$

The first term has the same normalisation as 5-01 (g_0 put into the covariant derivative). $\mathcal{O}_i$ are gauge-invariant terms of mass dimension 6 ($\mathrm{tr} F^2$ has dimension 4, so a^2 multiplies them; 4-04). According to the abstract of Lüscher–Weisz (1985), if only spectral quantities (masses of stable particles, the static potential and so on) are to be improved, one of the three coefficients can be chosen freely Known . **Why there is no a^1**: a term of dimension 5 has two field strengths and one derivative, so 5 indices. The lattice is invariant under reversing each axis, so the indices of each direction must appear an even number of times, and an odd number of indices cannot be placed Physics . This is an argument about the action and spectral quantities (masses). For quantities that involve operators, such as the force between static charges, it has to be shown separately that the correction starts at a^2, and Necco–Sommer (2002) § 4 gives that argument.

Adding rectangles

We add 1×2 rectangles to the plaquette with weights c_0, c_1 ($c_0 + 8c_1 = 1$). A rectangle has two orientations in each plane, and has twice the area so acts 4 times as strongly on the leading term; hence 8. The lowest-order (classical) coefficients are attributed to Lüscher–Weisz (1985), and for the action that adds only rectangles they are $c_1 = -\frac{1}{12}$ ($c_0 = \frac{5}{3}$) Known . Table 3 reproduces this number. We check it with a smooth U(1) field on a two-dimensional torus (a sum of six wave numbers; the flux computed exactly with an antiderivative).

Table 3 Relative error of the action for a smooth U(1) field in two dimensions (classical) Computed

$L = 1/a$	**Wilson**	$/a^2$	**Improved** $c_1 = -\frac{1}{12}$	$/a^4$
4	−0.2718	−4.35	−0.1229	−31.5
16	−0.02121	−5.43	-1.06×10^{-3}	−69.2
32	−0.005370	−5.50	-6.94×10^{-5}	−72.8
64	−0.001347	−5.52	-4.39×10^{-6}	−73.7

The ratio of errors on halving a is 3.95→3.99 (→4) for Wilson and 15.2→15.8 (→16) for the improved action. The Wilson coefficient approaches the expected value $-\frac{1}{12}\int |\nabla F|^2 / \int F^2 = -5.522$. At $L = 16$, taking $c_1 = 0, -\frac{1}{24}, -\frac{1}{12}, -\frac{1}{8}$ makes the error divided by a^2 move linearly as $-5.43, -2.85, -0.27, +2.31$ Computed . For a four-dimensional non-abelian field, see Table 3 of 5-01.

Coefficients and logarithmic corrections in the quantum theory

In the quantum theory c_i depend on g_0 (one loop: Lüscher–Weisz 1985). Because of asymptotic freedom, the leading error comes with a logarithmic correction of the form $a^2[\bar{g}^2(a^{-1})]^{\hat{\gamma}_1}$, and Husung–Marquard–Sommer (2020) conclude for Yang–Mills that

“the log-corrections are a rather weak modification” Physics . Here $\bar{g}^2(a^{-1})$ is the running coupling at the scale $1/a$ (4-06), which approaches 0 like $1/\ln(1/(a\Lambda))$. $\hat{\gamma}_1$ is an exponent fixed for each correction term (a number that plays the role of an anomalous dimension). The error therefore takes the form a^2 times a power of $\ln(1/a)$.

Detour: the price of improvement

A subsection you may skip.

The Wilson action has reflection positivity and gives a positive transfer matrix, that is, a lattice Hamiltonian (5-03). The abstract of Lüscher–Weisz (1984) showed that adding operators of dimension 6 means “physical positivity is lost in general”: a transfer matrix can be defined but complex eigenvalues can appear, and the spectral decomposition of two-point functions always has contributions with negative weight Known . It is a choice between taking precision and taking a lattice Hamiltonian (4-10) (18-04).

07

Continuum extrapolation — drawing a straight line against a²

A computer cannot reach $a = 0$, so one measures a ratio R at several β, draws a straight line against a^2 on the horizontal axis, and takes the intercept.

$$R(a) = R_0 + c\,a^2 + O(a^4) \quad \Longrightarrow \quad R_0 \approx \tfrac{1}{3}\big(4R(a/2) - R(a)\big) \ (8)$$

The third column of Table 2 is the error of this two-point extrapolation, and the error drops to a^4. Even if the slope c differs from action to action, the intercept R_0 should be common (universality; § 08). For the Wilson case of Table 3 too, building it from $L = 16, 32$ takes the relative error from 0.54 % to 0.009 % Computed . In actual numerics, Necco–Sommer (2002) fit the force in the form $(1 + s(a/r_c)^2)$ and report that the slope s is not significant Known .

The pitfalls are points that are too coarse (at $L = 4$ in Table 3, $/a^2$ is off by 20 %), the logarithmic corrections of § 06, and boxes that are too small. All are procedures resting on the premise that $R(a)$ has the form (8), and the premise is not a theorem Physics .

08

The continuum limit seen from mathematics — what has been shown and what has not

Chatterjee (2018) § 3 writes “the general belief in the physics community is that β should scale like log(1/ε) in dimension four, although there are doubts about this belief and the question remains an open mathematical problem” (written in 2018) Known . Integrating the upper formula of (3) gives $\beta \approx 4Nb_0 \log(1/a)$, and this is that $\log(1/\varepsilon)$. The continuum string-tension problem in § 5 of the same paper (Chatterjee's Problem 5.2) asks whether, in the infinite-volume theory as $\beta \to \infty$, there exist sequences $\varepsilon(\beta) \to 0$, $c(\beta) \to \infty$ and a nonzero d such that, for a rectangular loop of height R and width T,

$$\log\left\langle W_{\gamma_{R/\varepsilon,\,T/\varepsilon}}\right\rangle = -c\,(R+T) - d\,RT + o(1) \quad (9)$$

holds. d is the continuum string tension, and about the perimeter term it says “it is not clear to me whether the renormalization term c(R+T) is indeed necessary” Known .

In two dimensions it has been shown that the Wilson, Villain and Manton lattice actions go to the same Yang–Mills measure (Chevyrev–Shen, the trivial bundle on the torus; 7-02) Known . Four-dimensional U(1) behaves differently (there is a massless photon phase on the weak-coupling side, and the critical point is at finite β; 5-06), and the renormalisation steps of the non-abelian case are treated in 7-07. In the four-dimensional non-abelian case, the construction published in 2026-11 showed, for a family of Wilson actions with a chosen sequence of spacings, that the correlations of renormalised local observables converge (the committee's assessment is that it is proved in Lean; the volume limit is along a subsequence; 8-02, 8-05). The bare coupling of that family is chosen so that the ratio of β to log(1/a) approaches a fixed value (10-06). Independence of the action and of the sequence of spacings, the limit of the ratios (4), and the continuum string-tension problem are not theorems (6-04). The tables and figures of this chapter are procedures that assume the limit, not evidence of its existence. Volume 2 reads this construction in the order: the objects (8-01 to 8-05), one step of renormalisation (9-01 to 9-07), and matching the coupling (10-01 onward).

What this chapter can say

Matter	Said	Not said
Continuum limit = critical point (1)	An account of the physical usage Known	That ξ_{lat} exists and diverges on a four-dimensional non-abelian lattice

Matter	Said	Not said
Two-loop $a(\beta)$ (3), Table 1, Figure 1	Computed from the universal b_0, b_1 Computed	That the actual $a(\beta)$ follows (3) (in numerics a deviation remains Physics)
Scaling (4)	An account of numerical support Physics	That the ratios have a limit (no theorem)
Errors start at a^2; improvement $c_1 = -\frac{1}{12}$	Free field and two-dimensional U(1) (classical) Computed . Account of the effective theory Physics	Proofs, coefficients and logarithmic corrections in the four-dimensional quantum theory
Continuum extrapolation (8)	For free and classical fields Computed	Validity for actual numerics
Existence of the continuum limit, the mass gap	—	This chapter shows nothing about them

Sources and reproduction

Item	Kind	Source / tools
Quotations in § 01 and § 08	Primary source (text checked)	S. Chatterjee, “Yang–Mills for probabilists”, arXiv:1803.01950 (2018) § 3, § 5
Calibration by the string tension in § 02 and § 03	Primary source (abstract checked)	M. Creutz, “Monte Carlo Study of Quantized SU(2) Gauge Theory”, Phys. Rev. D 21 (1980) 2308
r_0 and t_0 in § 02	Primary source (abstract checked)	R. Sommer, Nucl. Phys. B 411 (1994) 839, arXiv:hep-lat/9310022 / M. Lüscher, JHEP 1008 (2010) 071, arXiv:1006.4518
r_0/a in § 02 (the values for $\beta \le 6.4$ in Table 1 are reproduced from earlier calculations) and $r_0 \approx 0.5$ fm, r_c/r_0 in § 04, the a^2 of the force in § 06, the fit in § 07	Primary source (text checked; bibliographic data only for § 4)	S. Necco, R. Sommer, Nucl. Phys. B 622 (2002) 328, arXiv:hep-lat/0108008
Lepage–Mackenzie in § 03	Primary source (abstract checked)	G. P. Lepage, P. B. Mackenzie, Phys. Rev. D 48 (1993) 2250, arXiv:hep-lat/9209022
Symanzik in § 06	Primary source (abstract checked)	K. Symanzik, Nucl. Phys. B 226 (1983) 187, 205

Item	Kind	Source / tools
Coefficients in § 06, the transfer matrix in the detour	Primary source (abstract checked. The number $-\frac{1}{12}$ is not checked in the text; the abstract goes as far as saying the coefficients were fixed at lowest order)	M. Lüscher, P. Weisz, Commun. Math. Phys. 97 (1985) 59; Erratum 98 (1985) 433 / Phys. Lett. B 158 (1985) 250 / Nucl. Phys. B 240 (1984) 349
Logarithmic corrections in § 06	Primary source (abstract checked)	N. Husung, P. Marquard, R. Sommer, Eur. Phys. J. C 80 (2020) 200, arXiv:1912.08498
Two-dimensional universality in § 08	Primary source (abstract checked)	I. Chevyrev, H. Shen, arXiv:2302.12160; Comm. Pure Appl. Math. (2026) (volume and pages not checked)
Tables 1–3, Figures 1 and 2, numbers in the text	Computed on this machine	Python 3.12, NumPy: `a_of_beta.py` (including the conversion of a to fm in § 02), `freefield.py`, `improve2d.py`. The formulas of the figures are checked against `jscheck.js`. For b_0, b_1 and the SU(3) step, 4-06

Next chapter to read: 5-06 How U(1) differs from the non-abelian case.

How U(1) differs from the non-abelian case — on the same lattice, an abelian group changes phase on the way

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — If the group on a four-dimensional lattice is $U(1)$, at large β one enters a phase of massless photons, and that there is a phase transition on the way is a theorem (at least for actions of Villain type) (Guth, Fröhlich–Spencer). For $SU(2)$ no jump is seen in the numbers, and it is believed that the correlation length does not diverge at finite β ($\beta_c = \infty$). The precise form of "because it is non-abelian".

Prerequisite chapters — 5-02 The strong-coupling expansion. We also borrow the phase transitions of 2-08 and the one-loop coefficient of 4-06.

Said first — The theorems are about $U(1)$, finite groups and three dimensions. For four-dimensional $SU(N)$, neither $\beta_c = \infty$ nor the gap has been proved. The numbers are observations on a small 6^4 lattice.

Order of this chapter

1. Two groups on the same lattice — sums of angles and products of matrices
2. On the strong-coupling side they cannot be told apart — both have the area law
3. Weak-coupling U(1) is a gas of photons — the Coulomb force (Figure 1)
4. Between the two shores the line breaks — Guth, Fröhlich–Spencer (Figure 2)
5. Detour: monopole loops — four dimensions and three
6. No jump is seen for SU(2) — only a crossover
7. The source of the difference is the running of the coupling — C_A and $\beta_c = \infty$
8. The precise form of "because it is non-abelian" — a table of counterexamples
9. What remains as mathematics

01

Two groups on the same lattice — sums of angles and products of matrices

Wilson's lattice (5-01) has the same form for every compact group. Group elements are placed on the edges, and the product around the four edges of a plaquette is weighted. For $U(1)$ (2-03 § 08) each edge carries an angle θ_ℓ, and the product becomes a sum of angles θ_P. For $SU(2)$ each edge carries a 2×2 matrix U_ℓ, and the product is the matrix product U_P.

$$S_{U(1)} = \beta \sum_P (1 - \cos\theta_P), \quad \theta_P = \theta_{x,\mu} + \theta_{x+\hat\mu,\nu} - \theta_{x+\hat\nu,\mu} - \theta_{x,\nu}, \quad \beta = \frac{1}{g^2}; \qquad S_{SU(2)}$$

The convention for β is chosen so that the action returns to the Maxwell or Yang–Mills action in the classical continuum limit Known . The $\beta = 1/g^2$ of $U(1)$ is not the $2/g^2$ obtained by putting $N=1$ into the $\beta = 2N/g^2$ of $SU(N)$. This is because the generators are normalised differently (for $U(1)$ the edge carries e^{-iagA_μ} (the sign is the same as in (2) of 5-01), for $SU(N)$ $\operatorname{tr} T^a T^b = \delta^{ab}/2$). For $U(1)$, $\theta_P \approx -a^2 g F_{\mu\nu}$ and $1-\cos\theta_P \approx \theta_P^2/2$ make the action approximately $\frac{\beta g^2}{4}\sum_x a^4 \sum_{\mu,\nu} F_{\mu\nu}^2$, and matching it to Maxwell's $\frac14\int F^2$ gives $\beta = 1/g^2$. The difference shows in the form of the formulas. θ_P is a **linear expression** in the edge angles (the lattice curl), and the only non-linearity of the $U(1)$ action is the periodicity of $\cos$, that is, the compactness of the group. For $SU(2)$, writing $U_\ell = e^{i\vartheta_\ell\cdot\sigma}$, the exponent of U_P contains terms with the commutator $[\vartheta,\vartheta]$ (1-03). This is the self-interaction on the lattice (3-02). Example: if the angles on the four edges are $0.1, 0.2, 0.3, 0.4$, then $\theta_P = -0.4$, and the order of addition does not matter. For $SU(2)$ with $A = e^{0.3i\sigma_1}$ and $B = e^{0.3i\sigma_3}$, $\frac12\operatorname{tr}(ABA^{-1}B^{-1}) = 0.9847$, while the reordered $\frac12\operatorname{tr}(ABB^{-1}A^{-1}) = 1$ Computed . For $U(1)$ the sum of angles is 0 in both cases. Cyclic permutations (such as $BA^{-1}B^{-1}A$), however, do not change the value, because $\operatorname{tr} XY = \operatorname{tr} YX$.

02

On the strong-coupling side they cannot be told apart — both have the area law

On the side where β is small the two groups behave alike. The variable u of the character expansion of 5-02 is

$$u_{U(1)} = \frac{I_1(\beta)}{I_0(\beta)}, \qquad u_{SU(2)} = \frac{I_2(\beta)}{I_1(\beta)}, \qquad \langle W_{R\times T}\rangle \approx u^{RT}, \quad \sigma \approx -\log u \ (2)$$

(I_n are modified Bessel functions), and since $u < 1$ both have the area law. Example: for $U(1)$ at $\beta = 0.5$, $u = 0.2425$ and $\sigma \approx 1.417$; at $\beta = 1$, $u = 0.4464$ and $\sigma \approx 0.807$ Computed . The area law and exponential decay at small β are theorems for every compact group (Osterwalder–Seiler 1978) Known . They are a property of the lattice being coarse and do not use non-commutativity (5-02 § 08).

The difference shows only in numerical coefficients. As $\beta \to 0$, $u_{U(1)} \approx \beta/2$ and $u_{SU(2)} \approx \beta/4$, so $\sigma \approx -\log u$ grows in both cases in the form $\log(1/\beta)$, and the difference is the constant $\log 2$. At $\beta = 1$, $u_{SU(2)} = 0.2402$ and $\sigma \approx 1.426$, larger than for $U(1)$ at the same β ($\sigma \approx 0.807$) Computed . The convention (1) differs from group to group, so comparing sizes at the same β means little. What matters is that both have an area law of the same form.

03

Weak-coupling U(1) is a gas of photons — the Coulomb force and the perimeter law

When β is large the weight demands $\theta_P \approx 0$, and with $1 - \cos\theta_P \approx \theta_P^2/2$ the action is quadratic. Since θ_P is linear, this is a lattice Maxwell field without interaction (photons).

$$S_{U(1)} \approx \frac{\beta}{2} \sum_P \theta_P^2, \qquad \langle \theta_\mu(k)\theta_\nu(-k)\rangle = \frac{\delta_{\mu\nu}}{\beta\,\hat{k}^2}, \qquad \hat{k}^2 = \sum_\mu 4\sin^2\frac{k_\mu}{2} \quad (3)$$

The right-hand formula of (3) is the form for one fixed gauge (Feynman gauge, 4-08) Known . Gauge-invariant quantities (Wilson loops) do not depend on this choice.

That $1/\hat{k}^2$ diverges as $\hat{k} \to 0$ is what "massless" means, and correlations of the field strength decrease as the power $|x|^{-4}$ Known .

Per point there are 3 physical degrees of freedom (4 edges minus 1 for the gauge), and 6 plaquettes. From the Gaussian integral $\int x^2 e^{-x^2/2}dx / \int e^{-x^2/2}dx = 1$, the mean action of one mode with quadratic action $\frac{\lambda}{2}x^2$ is $1/2$ (equipartition). Hence $6 \cdot \frac{\beta}{2}\langle\theta_P^2\rangle = \frac{3}{2}$, that is,

$$\langle \cos\theta_P \rangle \approx 1 - \frac{1}{4\beta} \;\; (U(1)), \qquad \left\langle \tfrac{1}{2}\mathrm{tr}\, U_P \right\rangle \approx 1 - \frac{3}{4\beta} \;\; (SU(2)) \quad (4)$$

($SU(2)$ has 3 generators). Example: for $U(1)$ at $\beta = 2$, 0.875, and Monte Carlo gives 0.865 (Table 2) Computed .

The force between static charges

An $R \times T$ Wilson loop is the amplitude of a pair of static charges held for time T. For a Gaussian field it can be written in closed form, and as $T \to \infty$

$$V(R) = -\lim_{T\to\infty} \frac{1}{T} \log\langle W_{R\times T}\rangle = \frac{G_3(0) - G_3(R)}{\beta}, \qquad G_3(r) = \int_{[-\pi,\pi]^3} \frac{d^3k}{(2\pi)^3} \frac{e^{ik\cdot r}}{\sum_i 4\sin^2($$

remains Known . G_3 is the Green function of the three-dimensional lattice, $1/(4\pi R)$ at large distance. So V has the Coulomb form and stays at the finite value $G_3(0)/\beta$. It does not confine. Computing with $G_3(r,0,0) = \int_0^\infty e^{-6t} I_r(2t) I_0(2t)^2 dt$ gives $G_3(0) = 0.252731$ Computed . The $1/6$ at $R = 1$ is a check from $6G_3(0) - 6G_3(1) = 1$.

Computed Table 1 $\beta V(R)$ of weak-coupling $U(1)$ (formula (5)) and the Coulomb approximation

R	$\beta V(R)$	$G_3(0) - \frac{1}{4\pi R}$
1	0.16667	0.17315
4	0.23244	0.23284
16	0.24775	0.24776
∞	0.25273	0.25273

Figure 1 The static potential of $U(1)$. The knob β changes the strong-coupling σR (dashed) and the weak-coupling Coulomb form (solid)



Computed The numbers in this figure were computed on this machine (the solid line is the calculation of Table 1; the dashed line computes the series for I_1/I_0 on the spot). Near $\beta \approx 1$ neither approximation can be used (§ 04).

For finite squares $R \times R$ too, we obtained $-\beta \log\langle W\rangle$ from a Fourier sum on a periodic 64^4 lattice. At $R = 1, 4, 8, 16$ it is $0.250,\ 1.631,\ 3.584,\ 7.557$: divided by the area R^2 it goes towards 0 as $0.250 \to 0.030$, and divided by the perimeter $4R$ it settles to almost a constant, $0.063,\ 0.102,\ 0.112,\ 0.118$ (the perimeter law). The value per perimeter approaches $G_3(0)/2 \approx 0.126$ as $R \to \infty$ (the same as the limit of long rectangles). It approaches with corrections of $1/R$ and logarithms from the corners Computed .

04

Between the two shores the line breaks — Guth and Fröhlich–Spencer

That the same picture remains when the approximations are dropped is a theorem for $U(1)$. The theorem, however, is about actions of Villain type (Guth's abstract). The β $\cos$

Villain type is the action whose weight is the periodic Gaussian $\sum_n e^{-\frac{\nu}{2}(\theta+2\pi n)^2}$ in place of $e^{\beta\cos\theta}$. The abstract of Fröhlich–Spencer does not state the type of action (the text was not obtained). (2), (4), Table 2, Figure 2 and the β_T below are numbers for the Wilson action (1).

The logic of the theorem is short: if the string tension is positive at small β and 0 at large β, then $\sigma(\beta)$ cannot be an analytic function. An analytic function that is 0 on an interval is 0 everywhere. A point where a thermodynamic quantity is not analytic is a phase transition (2-08 § 04).

$$\sigma(\beta) > 0 \ (\beta < \beta_1), \quad \sigma(\beta) = 0 \ (\beta > \beta_2) \implies \sigma \text{ is not analytic somewhere in } [\beta_1, \beta_2] \quad (6)$$

The second half was shown by Guth (1980): for an action of Villain type with $g^2 < 0.168$, a lower bound on Wilson loops "has the form of a Coulomb interaction and guarantees that electric charges are not confined", and together with the strong-coupling result of Osterwalder–Seiler he concludes "this model has at least one phase transition" (abstract) Known .

The abstract of Fröhlich–Spencer (1982) is "We give a new, elementary proof for the existence of a deconfining transition to a massless (QED) phase in the four-dimensional U(1) lattice gauge theory" Known . Chatterjee (2018) § 4 introduces it as "A fully rigorous proof of Guth's theorem".

Convergence to the continuum limit — Driver

The abstract of Driver (1987) deals with the continuum limit. "compact U(1) lattice gauge theory with general energy function converges to a renormalized free electromagnetic field on the current sector as the lattice spacing approaches zero" comes with the condition "provided the coupling constant is sufficiently large". For the Wilson action it says "it is possible, by judicious choice of the Gibbs state, to get convergence for arbitrary coupling strengths" Known . The "current sector" is the range of observables that smear the field with smooth currents. Whether the "coupling constant" of the abstract is on the side of β or of g has not been checked, because we have not been able to read the text.

The jump seen in numbers

For the Wilson action the transition point is reported numerically as $\beta_T = 1.011128(11)$, with evidence of a first-order transition (Arnold et al. 2001, conference proceedings; the abstract of the follow-up proceedings hep-lat/0210010 also writes "strong evidence for a first-order phase transition") Physics . A first-order transition is one in which the mean plaquette jumps at β_T, and there the two phases can coexist.

Because they can coexist, a run started in one phase stays in that phase for a while, and runs that raise β and runs that lower it disagree. This is hysteresis.

On this machine we compared, with the Metropolis method on a 6^4 lattice, a run lowering β from above (cold start) with a run raising it from below (hot start) (at each β, 100 sweeps discarded and 100 sweeps measured).

Figure 2 The mean plaquette. The buttons choose the group, and choosing a point of β with the knob changes the vertical line and the reading (cold ●, hot ○, leading strong-coupling term dashed, leading weak-coupling term dotted)



Computed The numbers in this figure were computed on this machine (the points are Monte Carlo, of which Table 2 is an excerpt; the curves compute (2) and (4) on the spot).

Computed Table 2 The mean plaquette on a 6^4 lattice (in brackets, the error from bin averages;

underestimated near the transition)

Group	β	Cold	Hot	Strong coupling	Weak coupling
U(1)	0.98	0.5742(55)	0.5541(27)	0.4393	0.7449
U(1)	1.00	0.6425(39)	0.5684(35)	0.4464	0.7500
U(1)	1.02	0.6756(18)	0.6150(38)	0.4534	0.7549
U(1)	2.00	0.8652(4)	0.8641(5)	0.6978	0.8750
SU(2)	1.8	0.4440(10)	0.4371(18)	0.3994	0.5833
SU(2)	2.2	0.5685(20)	0.5638(18)	0.4645	0.6591
SU(2)	4.0	0.7997(8)	0.7994(6)	0.6580	0.8125

For $U(1)$ the two runs disagree for $\beta = 0.98$–1.02, and at $\beta = 1$ the difference is 0.074. It can be read as each run staying in its original phase (the hysteresis above). The $U(1)$ transition was also seen in the numerics of Creutz–Jacobs–Rebbi (1979) Physics . This table comes from short runs on a small lattice and has no power to decide the order (5-07).

05

Detour: monopole loops — four dimensions and three

A section you may skip.

θ_P does not distinguish integer multiples of 2π. Bringing the θ_P of the six faces of a small cube back into $(-\pi, \pi]$ and adding them can leave an integer multiple of 2π, and that is a lattice magnetic monopole. In four dimensions monopoles are closed loops (world lines); a loop of length ℓ pays about $\beta\varepsilon\ell$ in action and can be drawn in about μ^ℓ ways.

$$\sum_\ell \mu^\ell e^{-\beta\varepsilon\ell} = \sum_\ell e^{-(\beta\varepsilon - \log\mu)\ell} \tag{7}$$

If $\beta\varepsilon < \log\mu$, long loops fill space, disorder the lines of force and produce the area law. When β is large only short loops remain, and photons survive. Estimates of this kind are in Banks–Myerson–Kogut (1977) Physics . Precisely, (7) leaves out the interaction between loops, and the theorem is the one in § 04.

In three dimensions monopoles become points, and a gas of points causes Debye screening (the weakening of the far field by a gas of point charges) at every coupling. Polyakov (1977) showed confinement for three-dimensional compact $U(1)$ by this mechanism (a semiclassical argument) Physics . Göpfert–Mack (1982) proved for the Villain action that the “string tension α is nonzero for all values of the coupling constant g^2” Known . The same abstract writes that in the continuum limit the string tension in physical units goes to infinity. It is an example that confines even though abelian, and where the scale does not become a single one (17-02 § 03).

06

No jump is seen for SU(2) — only a crossover

For $SU(2)$ in Table 2, the difference between the two runs is at most about 0.007, about a tenth of that for $U(1)$. The difference 0.0069 at $\beta = 1.8$ is about 3 times the error in the table, but it is a difference between runs thermalised for only 100 sweeps at each β, and as β varies the cold and hot runs do not split into two systematic branches (how to tell insufficient thermalisation from a jump is in 5-07). The plaquette follows the leading strong-coupling term for $\beta \lesssim 1.8$, approaches the leading weak-coupling term for $\beta \gtrsim 3$, and joins the two smoothly around $\beta \approx 2.2$ Computed .

The abstract of Creutz (1980) fixes the string tension, makes the lattice finer, sees the bare coupling decrease logarithmically as the renormalisation group predicts, and writes “This supports the coexistence of confinement and asymptotic freedom” Physics .

The peak of the specific heat around the crossover is said to be the “shadow” of a nearby critical end point in a family of actions with an added term in the adjoint representation (Bhanot–Creutz 1981) Physics . Whether there is a phase transition can depend on the choice of action, so we consider it separately from the continuum limit. That for four-dimensional $SU(2)$ there is no point at finite β where the correlation length diverges is a believed statement, not a theorem. Chatterjee (2018) § 5 writes “It is believed ... βc = ∞”, that is, $\beta_c = \infty$ (5-05 § 01).

07

The source of the difference is the running of the coupling — C_A and $\beta_c = \infty$

What separates $U(1)$ and $SU(2)$ is whether the effective coupling changes with distance (4-06). At one loop

$$\mu \frac{dg}{d\mu} = -b_0 g^3, \qquad b_0 = \frac{11\, C_A}{48\pi^2}, \qquad \frac{1}{g^2(L)} \approx \frac{1}{g_0^2} - 2b_0 \log \frac{L}{a} \tag{8}$$

C_A is the Casimir number of the adjoint representation, N for $SU(N)$ and 0 for $U(1)$ Known . The right-hand formula is the value of the coupling g_0 at spacing a seen at length L.

With no matter fields, b_0 is 0 for $U(1)$, 0.04644 for $SU(2)$ and 0.06966 for $SU(3)$ Computed .

For $U(1)$, $b_0 = 0$, so if β is large the Gaussian approximation is self-consistent at all distances, and the Coulomb force of § 03 continues to large distances. For $SU(2)$,

however large β is, taking L large enough brings the right-hand side of (8) towards 0, and the coupling becomes strong. The rough guide is $L \sim a\, e^{\beta/(4Nb_0)}$, the same form as the $a(\beta)$ of 5-05 § 03. That there is no point at finite β where the correlation length becomes infinite is what "$\beta_c = \infty$" means Physics . Example: for $SU(2)$ at one loop, raising β by 1 multiplies the correlation length by about $e^{1/0.3715} \approx 15$ ($4Nb_0 = 11/(3\pi^2)$) Computed . This is a one-loop guide only; corrections from two loops and beyond, and the breaking of asymptotic scaling in the range of β actually used, are treated in 5-05.

08

The precise form of "because it is non-abelian" — four phrasings and counterexamples

Using known examples with groups and dimensions exchanged, we list which phrasings fail (see also 16-08 § 04).

Known Table 3 Phrasings, and examples that break them (on the side of large β)

Phrasing	Example that breaks it	Source
Abelian means no confinement	Three-dimensional $U(1)$ has the area law at every coupling (Villain action)	Göpfert–Mack
Non-abelian means confinement	Four-dimensional non-abelian finite groups are thought not to confine Physics	Chatterjee's lecture material, 5-10
Exponential decay is because it is non-abelian	For finite groups, abelian or not, exponential decay at weak coupling is a theorem	Adhikari–Cao
The number of phases is decided by whether the group is abelian	Four-dimensional abelian $\mathbb{Z}_N$ (with N large) has at least three phases: confining, massless (QED-like) and screened	Elitzur–Pearson–Shigemitsu, Guth (Villain type, $N > 37$), Fröhlich–Spencer; numerics by Creutz–Jacobs–Rebbi

The "screened" of Table 3 is the “screened” of Elitzur–Pearson–Shigemitsu, the phase in which Wilson loops follow the perimeter law and excitations are massive. It is a different meaning from the Debye screening of § 05 (which produces confinement). The same paper derives this sequence of phases from the existence of the phases of the O(2) model, using self-duality and correlation inequalities (abstract). Numerically, the abstract of Creutz–Jacobs–Rebbi (1979) reports two transitions for $N \geq 5$ Physics .

What matters is not only "non-abelian" but that the group is **continuous**. Continuity matters, however, for whether massless waves remain at the end of sufficiently large β. For a finite group the edge values are discrete, and at sufficiently large β the edges freeze to the identity, so no massless waves like photons can be made Physics . On the

other hand $\mathbb{Z}_N$ (with N large) has a massless phase at intermediate β (row 4 of Table 3). Exponential decay at large β for finite groups is a different thing from the gap of the continuum theory (5-10). $U(1)$ is continuous, so it has massless waves, and since $C_A = 0$ they survive.

The precise form

For a four-dimensional continuous non-abelian compact group, (i) there is no point at finite β where the correlation length diverges ($\beta_c = \infty$), and (ii) as $\beta \to \infty$ the gap in lattice units $1/\xi$ goes to 0, but dimensionless ratios (for example $m/\sqrt{\sigma}$) stay at positive values Physics .

(ii) is the form of the lattice correlation-length problem of 5-09 (Chatterjee's Problem 5.1). Neither (i) nor (ii) has been proved. It is more accurate to say that the role of non-commutativity is not "to make a gap" but "to break the massless weak-coupling picture at large distances".

09

What remains as mathematics

On the side of theorems are the strong-coupling results of § 02, Guth and Fröhlich–Spencer (actions of Villain type) and Driver (with the conditions of the abstract) in § 04, Göpfert–Mack in § 05, rows 1 and 3 of Table 3, and Guth and Fröhlich–Spencer in row 4 Known . (7) and Polyakov in § 05, and row 2 of Table 3, are at the level of physics. On the side of belief: that for four-dimensional $SU(N)$ there is no point at finite β where the correlation length diverges ($\beta_c = \infty$, Chatterjee § 5), the area law for all β, and that the gap remains in physical units in the continuum limit Physics . Depending on the action, there can be bulk (lattice-scale) phase transitions not accompanied by a divergence of the correlation length. Lattice calculations of $SU(N)$ with large N report such a transition on the strong-coupling side, limiting the range that can be computed (abstract of Lucini–Teper–Wenger 2005) Physics . This does not contradict $\beta_c = \infty$. The lattice area-law problem of Chatterjee (2018) (Chatterjee's Problem 4.1) asks for the area law at all $\beta > 0$ for four-dimensional non-abelian compact groups, and writes that at large β it “remains elusive” (5-02 § 08).

What the $U(1)$ theorem teaches is that an argument extending strong-coupling results to weak coupling without distinguishing groups is wrong. What must be shown for $SU(N)$ includes a negative statement, "the divergence of the correlation length and the breaking of the area law do not occur at finite β", and the tools to say that over the whole range of β exist at present only on the strong-coupling side (7-10, 16-07).

What this chapter can say

Matter	Said	Not said
Leading terms (2)(3)(4), estimate (7)	Standard approximations Known . Consistent with Table 2 at both ends Computed . (7) is a physical estimate Physics	Errors at intermediate β. The coefficients of (7)
Static potential, perimeter law	A bounded Coulomb force and the perimeter law within the Gaussian approximation Computed	Estimates without the approximation
The phase transition of four-dimensional $U(1)$ (actions of Villain type)	An account of the theorems Known . Observation of hysteresis for the Wilson action on 6^4 Computed	The order and position of the transition. The form of the theorem for the Wilson action
$\beta_c = \infty$ for $SU(2)$	That no jump is seen on a small lattice Computed	A proof that the correlation length does not diverge at finite β
"Because it is non-abelian"	Sorting the phrasings by known counterexamples Known	Proofs of (i)(ii). This chapter shows nothing about the mass gap or the continuum limit

Sources and reproduction

Item	Kind	Source / tools
§ 04, Table 3	Abstract (text not obtained)	A. H. Guth, Phys. Rev. D 21 (1980) 2291
§ 04, Table 3	Abstract (text not obtained; the abstract does not give the type of action)	J. Fröhlich, T. Spencer, Commun. Math. Phys. 83 (1982) 411–454
Continuum limit in § 04	Abstract (text not obtained)	B. K. Driver, Commun. Math. Phys. 110 (1987) 479–501
Strong-coupling theorem in § 02	Bibliographic data only	K. Osterwalder, E. Seiler, Ann. Phys. 110 (1978) 440–471
§ 04, § 06, § 09	Text § 4, § 5	S. Chatterjee, arXiv:1803.01950 (2018)
Bulk transition in § 09	Abstract	B. Lucini, M. Teper, U. Wenger, “Properties of the deconfining phase transition in SU(N) gauge theories”, JHEP 0502 (2005) 033, arXiv:hep-lat/0502003

Item	Kind	Source / tools
β_T in § 04	Abstract (conference proceedings)	G. Arnold, T. Lippert, T. Neuhaus, K. Schilling (author order as on arXiv), “Finite size scaling analysis of compact QED”, Nucl. Phys. B Proc. Suppl. 94 (2001) 651–656, arXiv:hep-lat/0011058
Follow-up proceedings in § 04	Abstract	G. Arnold, B. Bunk, T. Lippert, K. Schilling, Lattice 2002, arXiv:hep-lat/0210010
Numerics in § 04, row 4 of Table 3	Abstract	M. Creutz, L. Jacobs, C. Rebbi, Phys. Rev. D 20 (1979) 1915
§ 05	Abstract	T. Banks, R. Myerson, J. B. Kogut, Nucl. Phys. B 129 (1977) 493 / A. M. Polyakov, Nucl. Phys. B 120 (1977) 429
§ 05, Table 3	Abstract	M. Göpfert, G. Mack, Commun. Math. Phys. 82 (1982) 545–606 (the year is the Crossref year of publication; INSPIRE gives 1981)
§ 06	Abstract	M. Creutz, Phys. Rev. D 21 (1980) 2308 / G. Bhanot, M. Creutz, Phys. Rev. D 24 (1981) 3212
Table 3	Abstract	A. Adhikari, S. Cao, Ann. Probab. 53 (2025) 140, arXiv:2202.10375 / S. Elitzur, R. B. Pearson, J. Shigemitsu, Phys. Rev. D 19 (1979) 3698
Row 2 of Table 3	Text of the lecture slides (the subject is the area law: “It is not expected to hold for 4D Abelian theories at large β, nor even for 4D non-Abelian theories with finite gauge groups”). The same slides mention that Cao (2020) obtained the leading term of Wilson loops at large β for every finite group in four dimensions (Cao's paper: bibliographic data only, not checked)	S. Chatterjee, “Yang–Mills on the lattice: New results and open problems” (lecture slides; year not checked, the PDF was created in 2020)
Table 2, Figure 2	Computed on this machine	Python 3, NumPy. mc_u1.py, mc_su2.py (Metropolis method, arguments 6 100 100, 1 to 2 minutes per run), approx.py
The example of ordering in § 01, the comparison of u in § 02	Computed on this machine	Python 3, SciPy. order_su2.py

Item	Kind	Source / tools
Table 1, the perimeter law of § 03, Figure 1, b_0 in § 07	Computed on this machine	Python 3, SciPy. coulomb.py (the integral for G_3, the Fourier sum on 64^4), g3table.py

Next chapter to read: 5-07 The Monte Carlo method (the method that made Table 2, and how to read its errors).

The Monte Carlo method — make configurations that follow the weight, count the correlations, and give the error

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows

- How the integral for a lattice expectation value (tens of thousands of dimensions) is estimated by a sample mean over a Markov chain that follows the weight.
- How to count the extra error caused by correlation between samples (the autocorrelation time, binning, the Γ method).
- Critical slowing down: as the continuum limit is approached, the correlation time grows as ξ^z (for a free field, $z = 2$ by a formula).

Prerequisite chapters — 5-01 Wilson's lattice, 1-08 Probability and measure. For the mathematical side, see 18-03 § 06– § 08.

Said first — The Monte Carlo method estimates expectation values on a finite lattice with error bars; it does not prove the mass gap or the continuum limit. The calculations in this chapter are small examples: one point of $SU(2)$ on 6^4, and a two-dimensional free field.

Order of this chapter

1. Why integrate with random numbers
2. Markov chains and detailed balance
3. Moving one link at a time — the step size (Figure 1, Table 1)
4. Heat bath and overrelaxation
5. Thermalisation and autocorrelation — three update methods (Table 2)
6. Estimating the error — bins, the jackknife, the Γ method
7. Critical slowing down — $z = 2$ for a free field (Figure 2, Table 3)
8. HMC and topological freezing

9. Monte Carlo seen from mathematics — the line between theorems and observations

---

The β of this chapter is the coupling of the Wilson action (5-01), not an inverse temperature.

01

Why integrate with random numbers — an integral in fifteen thousand dimensions

A lattice expectation value is an integral over all links (5-01 § 06; dU_ℓ is the Haar measure).

$$\langle O \rangle = \frac{1}{Z} \int O(U)\, e^{-S_W(U)} \prod_\ell dU_\ell \quad (1)$$

A four-dimensional lattice of side 6 has 5,184 links, and $SU(2)$ is 3-dimensional per link, so it is an integral in 15,552 dimensions Computed . So we make configurations $U^{(n)}$ that follow the weight e^{-S_W}/Z and take the sample mean.

$$O_N = \frac{1}{N} \sum_{n=1}^{N} O\big(U^{(n)}\big), \qquad \text{if independent, } \operatorname{Var} O_N = \frac{\operatorname{Var} O}{N} \quad (2)$$

The error decreases as $1/\sqrt{N}$ whatever the dimension (1-08) Known . Drawing uniformly does not work. At $\beta = 2.3$ on 6^4, the mean of the 7,776 plaquettes is 0.6026 in equilibrium, while uniform configurations scatter around 0 with standard deviation 0.00567, so the equilibrium value is 106 standard deviations away Computed .

02

Markov chains and detailed balance — making a distribution that stands still

We make a random rule that changes the configuration one step at a time (a Markov chain, with transition probability $P(x, y)$), so that after a long time it follows the weight π. A sufficient condition for the distribution not to change in one step (stationarity) is **detailed balance**.

$$\sum_x \pi(x) P(x, y) = \pi(y), \qquad \pi(x) P(x, y) = \pi(y) P(y, x) \quad (3)$$

Summing the right-hand side over x gives the left. A chain satisfying the right-hand side is called **reversible**. If moreover one can get from anywhere to anywhere (irreducible), the time average of a finite-state chain converges to the expectation value (the ergodic theorem). If it is also free of periodicity (aperiodic), the distribution itself approaches π (the convergence theorem) Known . For continuous state spaces too, under conditions such as the transitions having a positive density, there are a law of large numbers and a central limit theorem of the same form (Tierney 1994) Known .

The **Metropolis method** (Metropolis et al. 1953; the first application was the equation of state of two-dimensional hard spheres) proposes a candidate y with a symmetric proposal $q(x,y) = q(y,x)$, accepts it with the following probability, and stays put if it is rejected.

$$A(x \to y) = \min\bigl(1,\ \pi(y)/\pi(x)\bigr) = \min\bigl(1,\ e^{-\Delta S}\bigr) \quad (4)$$

If $\pi(y) \le \pi(x)$, then $\pi(x) q \cdot \pi(y)/\pi(x) = \pi(y) q \cdot 1$, so (3) holds, and Z is not needed. Example: two states, $\pi = (0.2, 0.8)$, with the proposal "go to the other one", gives $P = \begin{pmatrix} 0 & 1 \\ 0.25 & 0.75 \end{pmatrix}$ with eigenvalues $1, -0.25$, and the distribution from state 0 is $(0.1992,\ 0.8008)$ after 5 steps Computed . The deviation decreases as powers of the eigenvalue of largest absolute value other than 1 (here -0.25), and this is the source of the autocorrelation time of § 05.

03

Moving one link at a time — Metropolis and the step size

In four dimensions 6 plaquettes contain the link U_ℓ, and the part of the action that depends on U_ℓ has the following form.

$$S_W = -\frac{\beta}{N} \operatorname{Re}\operatorname{tr}\bigl(U_\ell A_\ell\bigr) + (\,U_\ell\text{-independent terms}\,), \qquad A_\ell = \sum_{6\ \text{plaquettes}} \text{staples} \quad (5)$$

A staple is the product of the three edges of a plaquette other than U_ℓ (5-01 § 03). A unit that goes once through all links is called a **sweep**.

The Metropolis proposal is near the current value. If the step size δ is large, rejections increase; if it is small, the chain moves only a little. We look at a single $U(1)$ link (weight $e^{\kappa \cos\theta}$).

Figure 1 A single $U(1)$ link ($\kappa = 2$). The knob sets the step size δ, and "Play" advances the chain. Left: θ over the last 300 steps; right: the histogram and the target density

Computed The numbers in this figure were computed on this machine (the readings are from Table 1; the chain in the figure is an example run on the spot).

Computed Table 1 Metropolis on a single link (200,000 steps each). The acceptance rates are exact values from numerical integration (agreeing with the runs to within 0.002). Exact $\langle\cos\theta\rangle = I_1(2)/I_0(2) = 0.69777$

δ	Acceptance rate	$\tau_{\mathrm{int}}(\cos\theta)$	$\langle\cos\theta\rangle$
0.1	0.975	242(37)	0.747(17)
0.25	0.937	63.9(5.5)	0.695(10)
1	0.757	5.90(18)	0.6972(31)
2.5	0.494	1.94(4)	0.6990(18)
π	0.426	2.00(4)	0.6990(18)

Widening δ from 0.1 to 2.5 halves the acceptance rate and makes τ_{int} (§ 05) about 120 times smaller. A high acceptance rate does not indicate efficiency. At $\delta = 0.1$ the mean is off from the exact value by 2.9 standard deviations; for a slow chain, the error estimate is itself uncertain.

04

Heat bath and overrelaxation — redrawing from the conditional distribution

The **heat-bath method** redraws a single link directly from the conditional distribution $\propto e^{\frac{\beta}{2}\mathrm{tr}(U_\ell A_\ell)}$ with the rest held fixed. A sum of $SU(2)$ matrices is a real multiple of an $SU(2)$ matrix (the quaternions of 1-04), so we can write $A_\ell = kV$ ($V \in SU(2)$); $X = U_\ell V$ also moves with the Haar measure, and the weight is $e^{\beta k x_0}$ ($x_0 = \frac{1}{2}\mathrm{tr}X$). $SU(2)$ is the three-sphere $x_0^2 + |\vec{x}|^2 = 1$, and the Haar measure is the uniform measure on it. The slice at fixed x_0 is a two-sphere of radius $\sqrt{1-x_0^2}$, with area proportional to $1-x_0^2$. When x_0 moves by dx_0, the width (arc length) on the three-sphere is $dx_0/\sqrt{1-x_0^2}$. The product is $\sqrt{1-x_0^2}\,dx_0$, so

$$p(x_0) \;\propto\; \sqrt{1-x_0^2}\,e^{\beta k x_0} \qquad (-1 \le x_0 \le 1) \;\text{(6)}$$

we draw x_0 from this, choose the remaining direction uniformly, and set $U_\ell = XV^{-1}$ Known . This is the method said to have been used in the $SU(2)$ calculation of Creutz (1980). Kennedy–Pendleton (1985) is an improvement that draws (6) quickly, and Cabibbo–Marinari (1982) is a method that updates $SU(N)$ one $SU(2)$ subgroup at a time.

Overrelaxation is a move that goes far without changing the action; for $SU(2)$ it is $U_\ell \mapsto V^{-1}U_\ell^{-1}V^{-1}$. With $X' = X^{-1}$ the action does not change, two applications return to the start, and it preserves the Haar measure, so it satisfies detailed balance. On its own it is not ergodic, so it is mixed with the heat bath (Adler 1981, Creutz 1987, Brown–Woch 1987). Note that a sweep in a fixed order is in general not reversible (the chain seen backwards in time is the one updated in reverse order), and only stationarity remains Known .

05

Thermalisation and autocorrelation — forgetting the starting point and counting the correlations

A chain starts either from a **cold start**, with all links set to the identity matrix, or from a **hot start** drawn uniformly, and the first stretch is discarded (thermalisation). For the plaquette at 6^4, $\beta = 2.3$ (equilibrium 0.6026, the heat bath of Table 2), the heat bath enters the equilibrium band from both sides in 20 sweeps, while the hot side of Metropolis lags, at 0.482 after 20 sweeps and 0.593 after 150 Computed . Agreement of the two starting points is only a necessary condition (5-06 § 04; the mathematical counterpart is the mixing time of 18-03 § 06).

Even in equilibrium the configurations are not independent. We define the **autocorrelation function**, the correlation of the time series at a separation of t steps divided by the variance, and the **integrated autocorrelation time**

$$\rho(t) = \frac{\langle \delta O_n \, \delta O_{n+t} \rangle}{\langle \delta O^2 \rangle}, \qquad \tau_{\mathrm{int}} = \frac{1}{2} + \sum_{t \geq 1} \rho(t), \qquad \operatorname{Var} O_N \approx \frac{2\tau_{\mathrm{int}} \operatorname{Var} O}{N} \tag{7}$$

($\delta O = O - \langle O \rangle$, $N \gg \tau_{\mathrm{int}}$). The effectively independent samples number $N/(2\tau_{\mathrm{int}})$ (formulas (23) and (25) of Wolff 2004) Known . For the heat bath of Table 2 that is $9{,}500/(2 \cdot 5.08) \approx 935$. The time τ_{exp} of the slowest decay is set by $\lambda_\star$, the eigenvalue of the transition of largest absolute value other than 1, as $\tau_{\mathrm{exp}} = -1/\log \lambda_\star$ (in the example of § 02, $\lambda_\star = 0.25$).

On the mathematical side, slowness is measured by the **Poincaré constant** C_P. C_P is the smallest constant such that $\operatorname{Var} f \leq C_P \, \mathbb{E}|\nabla f|^2$ for every function f (1-10 § 07; for a discrete chain the right-hand side is replaced by half the mean square change in one step, $\frac{1}{2}\sum_{x,y} \pi(x) P(x,y) (f(y) - f(x))^2$). **Langevin dynamics** is a continuous-time dynamics that adds noise to motion down the gradient of the action, and there $\tau_{\mathrm{exp}} = C_P$ holds with equality. For a discrete-time chain, the following three are different quantities Known (Levin–Peres–Wilmer § 12.2, 18-03 § 06).

- $C_P = 1/\gamma$: $\gamma = 1 - \lambda_2$ is the gap between 1 and the second largest eigenvalue λ_2.
- The relaxation time $t_{\mathrm{rel}} = 1/\gamma_\star$: $\gamma_\star = 1 - \lambda_\star$ is the absolute gap.
- $\tau_{\mathrm{exp}} = -1/\log \lambda_\star$.

When $\gamma_\star$ is small, the ratio of the last two is $1 + O(\gamma_\star)$. The **Polyakov loop** in the last column of Table 2 is the quantity obtained by taking, at each spatial point, the trace of the product of links once around the time direction, and averaging over space (here we look at its absolute value). It is an example of a slowly moving quantity spread over the whole lattice.

Computed Table 2 Four-dimensional $SU(2)$, 6^4, $\beta = 2.3$ (first 500 sweeps discarded). τ_{int} in sweeps. The cost is the number of staple computations per link

Method	Cost	Plaquette	τ: plaquette	τ: |Polyakov|
Metropolis	1	0.60291(57)	42(10)	79(24)
Heat bath	1	0.60261(21)	5.1(6)	6.0(7)
Heat bath + overrelaxation 3	4	0.60297(19)	2.1(2)	0.87(6)

The sample sizes are 9,500, 9,500 and 4,500. At the same sample size the error of Metropolis is 2.7 times that of the heat bath, consistent with $\sqrt{42/5.1} \approx 2.9$. Comparing τ multiplied by the cost, overrelaxation loses for the plaquette, 8.3 against 5.1, and

gains for the Polyakov loop, 3.5 against 6.0. What gains are slow global quantities Computed . For $SU(3)$ too, overrelaxation is reported to have shrunk the τ of the magnitude of the Polyakov loop from 28 sweeps to 9 (Brown–Woch 1987), and the tendency agrees Physics .

06

Estimating the error — bins, the jackknife, the Γ method

τ_{int} is estimated from the data. **Binning** cuts the time series into blocks of length B and treats the block means as independent. One reads off where the error levels off for $B \gg \tau_{\mathrm{int}}$ (Flyvbjerg–Petersen 1989). The **jackknife** gives the error of a non-linear quantity f from the means $\bar{b}_{(i)}$ with one block left out at a time.

$$\sigma_f^2 = \frac{n-1}{n} \sum_{i=1}^{n} \left(f(\bar{b}_{(i)}) - f(\bar{b}_{(\cdot)}) \right)^2, \qquad n = N/B \tag{8}$$

It is used for Creutz ratios and masses (5-02 § 06, 5-08). For the heat-bath plaquette of Table 2, (bin error / naive error)^2 grows as 1.00, 2.85, 6.23, 10.19 for $B = 1, 5, 20, 100$, reaching the Γ-method $2\tau_{\mathrm{int}} = 10.2$. For $B = 200$–1000 (47 to 9 blocks) it fluctuates downwards, 7.4 to 7.7. With few blocks the error of the error matters (for the square of the ratio the relative error is about $\sqrt{2/(n-1)}$) Computed .

The **Γ method** estimates the sum (7) directly, cut off at a window W.

$$\tau_{\mathrm{int}}(W) = \frac{1}{2} + \sum_{t=1}^{W} \rho(t), \qquad \langle (\delta\tau_{\mathrm{int}})^2 \rangle \approx \frac{4}{N} \left(W + \frac{1}{2} - \tau_{\mathrm{int}} \right) \tau_{\mathrm{int}}^2 \tag{9}$$

The right-hand side is formula (42) of Wolff (2004). It is derived by error propagation from the approximate formula for the error of the windowed sum given by Madras–Sokal (1988) (Wolff's formula (40)) Known . If W is short there is bias; if long, noise matters. The τ_{int} of this chapter use Wolff's (2004) automatic window ($S = 1.5$). Wolff asks one to “verify by eye” that $\tau_{\mathrm{int}}(W)$ is flat. In the lower panel of Figure 2 (§ 07) one can check by eye that it is flat around the window W of the vertical line.

07

Critical slowing down — seeing $z = 2$ by a formula for a free field

As the continuum limit is approached the correlation length ξ (in lattice units) grows (5-05 § 01), and the τ of local updates grows too. Physics writes $\tau \propto \xi^z$ and calls z the dynamical exponent (the review of Hohenberg–Halperin 1977, 18-03 § 07) Physics . The lattice free field (1-08 § 07) is where this can be checked by a formula.

$$S = \frac{1}{2}\sum_x \Big[\sum_\mu (\phi_{x+\hat\mu} - \phi_x)^2 + m^2\phi_x^2\Big], \qquad \phi_x \,|\, \text{neighbours} \sim \mathcal{N}\Big(\frac{\sum_{y\sim x}\phi_y}{2d+m^2}, \frac{1}{2d+m^2}\Big) \quad (10$$

In a checkerboard heat bath, let M_e be the sum over even points and M_o over odd. An odd point neighbours $2d$ even points, so $M_e' = rM_o +$ (noise), with $r = 2d/(2d+m^2)$. For the zero momentum $M = M_e + M_o$

$$\rho_M(t) = \frac{1+r}{2}\, r^{2t-1} \ (t \ge 1), \qquad \tau_{\text{int}}(M) = \frac{1}{2(1-r)} = \frac{1}{2} + \frac{d}{m^2}, \qquad \tau_{\text{exp}} = \frac{-1}{\log r^2} \approx \frac{d}{m^2}$$

From $\cosh(1/\xi) = 1 + m^2/2$, $\xi \approx 1/m$, so $\tau \approx d\xi^2$, and for this dynamics $z = 2$ Computed . The intuition is a random walk: a local update takes ξ^2 steps to carry information a distance ξ. Precisely, the $r = 1 - m^2/(2d+m^2)$ of (11) approaches 1 as $m \to 0$, so the fraction that shrinks in one sweep becomes $1 - O(\xi^{-2})$.

Figure 2 Two-dimensional free field, checkerboard heat bath. The knob sets m (top: $\rho_M(t)$, points measured, curves (11); bottom: $\tau_{\text{int}}(W)$, dashed line (11), vertical line the automatic window W). "Overlay" makes the horizontal axis t/ξ^2, and the four coincide

```
m = 0.25 (ξ = 4.01, 64²): τ_int measured 33.14 ± 3.91, (11) 32.5, τ_int/ξ² = 2.02
```

Computed The numbers in this figure were computed on this machine (the curves compute (11) on the spot. If JavaScript is disabled, Table 3).

Computed Table 3 Two-dimensional free field, checkerboard heat bath (side 16 to 64, 50,000 to 60,000 sweeps). τ in sweeps

m	ξ	$\tau_{\mathrm{int}}(M)$ **measured**	**(11)**	**(11)**$/\xi^2$
1	1.039	2.57(10)	2.5	2.316
0.5	2.020	8.17(53)	8.5	2.082
0.25	4.010	33.1(3.9)	32.5	2.021
0.125	8.005	125(23)	128.5	2.005

Each doubling of ξ multiplies $\tau_{\mathrm{int}}(M)$ by about 4. The cost of a calculation of the same precision is estimated to grow roughly as ξ^{d+z} (5-05) Physics . For Langevin dynamics

(§ 05) the relaxation time is $1/m^2 = C_P$ (18-03; in the free-field normalisation where the field values ϕ_x themselves are the coordinates in which gradients are measured; changing the choice of coordinates changes C_P only by a constant factor). Both are proportional to m^{-2} ($\propto \xi^2$). The τ of the heat bath is measured in sweeps and the τ of Langevin in Langevin time, so we do not compare the coefficients.

08

HMC and topological freezing — a global step, and a wall hard to cross

HMC (hybrid Monte Carlo, Duane et al. 1987) moves all links at once. Momenta P_ℓ (traceless Hermitian matrices; real linear combinations of the T^a of 1-03) are given normal distributions, and

$$H(P,U) = \frac{1}{2}\sum_\ell \operatorname{tr} P_\ell^2 + S_W(U) \tag{12}$$

the classical mechanics is run with the leapfrog method and accepted with probability $\min(1, e^{-\Delta H})$. The leapfrog method can be run backwards in time and preserves phase-space volume, so even with step errors e^{-S_W} is exactly the stationary distribution Known . Estimating the autocorrelation time of HMC by the Langevin C_P is an assumption (the generators are different).

As the continuum limit is approached, the field splits into sectors of topological charge Q (the number counting how the field is twisted, integer-valued for continuum fields; for its definition on the lattice see 3-05 § 08), and small steps find it hard to move between them. Schaefer–Sommer–Virotta (2011) reported, for HMC in pure gauge theory, an "effective dynamical critical exponent of about 5" for Q^2, and held that Wilson loops can be decoupled from this slow mode Physics . Lüscher–Schaefer (2011) showed a method that lets topological charge flow in and out through open boundary conditions in the time direction Physics . As a metaphor it is crossing passes in a landscape with many valleys. Precisely, if the measure near the boundaries of sectors is small the Poincaré constant becomes large (18-03 § 08) Known , but whether that is the cause of freezing has not been established.

09

Monte Carlo seen from mathematics — the line between theorems and observations

On a finite lattice (1) is a probability measure on a compact space (5-01 § 06). What can be said as theorems is the convergence of time averages by the ergodic theorem (for continuous state spaces, under the conditions of § 02) and the central limit theorem.

For a reversible chain, if the asymptotic variance is finite, a central limit theorem of the form (7) holds (Kipnis–Varadhan 1986) Known . What remains observation is the error bar. If there is a slow mode invisible at finite N, the error bar comes out too small ($\delta = 0.1$ in Table 1, § 08). Extrapolation from finite volume and finite a (5-05 § 07, 5-09) needs further assumptions.

The glueball masses and string tension of 5-08 are physical results that rely on that extrapolation Physics , not a proof of the existence and the gap of the problem statement (4-11). On the other hand CP is a mathematical object, and at strong coupling a Poincaré inequality for lattice Yang–Mills has been shown (Shen–Zhu–Zhu 2023, 5-04) Known . For four-dimensional Yang–Mills there is no theorem fixing the ξ-dependence of CP on the side $\xi \to \infty$.

For how this book uses this, see 23-08 § 01 and § 03 (checking numerical inequalities by interval arithmetic, under the label computer-assisted).

What this chapter can say

Matter	Said	Not said
Update methods	That Metropolis, heat bath, overrelaxation and HMC preserve the stationary distribution; (6) Known	The speed of mixing on a four-dimensional lattice
Tables 1 and 2	τ_{int} and errors in small examples Computed	Generalisation to large lattices. That the error bars are upper bounds
Critical slowing down	(11) and $z = 2$ for the free field with a checkerboard heat bath Computed . The general $\tau \propto \xi^z$ is a hypothesis Physics	The z of four-dimensional Yang–Mills
Topological freezing	The number in the abstract (effective exponent about 5) Physics	A mathematical explanation of freezing
The mass gap	—	This chapter shows nothing about the mass gap or the continuum limit

Sources and reproduction

Item	Kind	Source / tools
§ 02	Abstract	N. Metropolis et al., J. Chem. Phys. 21 (1953) 1087

Item	Kind	Source / tools
The ergodic theorem and convergence theorem in § 02, the relaxation time in § 05	Text	D. A. Levin, Y. Peres, E. L. Wilmer, *Markov Chains and Mixing Times*, 2nd ed.: Theorem C.1 (ergodic theorem, Appendix C.1), Theorem 4.9 (convergence theorem, § 4.3), § 12.2, Theorems 12.4–12.5 (relaxation time and mixing time)
Continuous state spaces in § 02 and § 09	Abstract	L. Tierney, Ann. Statist. 22 (1994) 1701
§ 04, § 05 (Brown–Woch's 28→9 sweeps)	Abstract (the attribution of the heat bath to Creutz not checked)	M. Creutz, PRD 21 (1980) 2308 / N. Cabibbo, E. Marinari, PLB 119 (1982) 387 / A. D. Kennedy, B. J. Pendleton, PLB 156 (1985) 393 / S. L. Adler, PRD 23 (1981) 2901 / M. Creutz, PRD 36 (1987) 515 / F. R. Brown, T. J. Woch, PRL 58 (1987) 2394
(7)(9) and the automatic window in § 05 and § 06	Text (formulas (23), (25), (40), (42), (50)–(52); (42) is error propagation from (40))	U. Wolff, CPC 156 (2004) 143 (erratum: CPC 176 (2007) 383), arXiv:hep-lat/0306017
§ 06 (the origin of formula (40) for the error of the windowed sum, binning)	Abstract	N. Madras, A. D. Sokal, J. Stat. Phys. 50 (1988) 109 / H. Flyvbjerg, H. G. Petersen, J. Chem. Phys. 91 (1989) 461
The dynamical exponent z in § 07	Abstract	P. C. Hohenberg, B. I. Halperin, Rev. Mod. Phys. 49 (1977) 435
§ 08	Abstract	S. Duane, A. D. Kennedy, B. J. Pendleton, D. Roweth, PLB 195 (1987) 216 / S. Schaefer, R. Sommer, F. Virotta, NPB 845 (2011) 93, arXiv:1009.5228 / M. Lüscher, S. Schaefer, JHEP 07 (2011) 036, arXiv:1105.4749
§ 09	Bibliographic data only / abstract	C. Kipnis, S. R. S. Varadhan, CMP 104 (1986) 1 / H. Shen, R. Zhu, X. Zhu, CMP 400 (2023) 805, arXiv:2204.12737
Tables 1–3, Figures 1 and 2, the numbers in § 01, § 02 and § 05–§ 07	Computed on this machine	Python 3, NumPy, SciPy. one_link_metropolis.py, su2_algos.py, analyze_su2.py, free_field.py, exact_tint.py, free_field_ext.py (extension for Figure 2), gamma_method.py (checked with AR(1)), small_numbers.py

Next chapter to read: 5-08 Measuring glueballs and the string tension (reading masses and the string tension from configurations made by the methods of this chapter).

Measuring glueballs and the string tension — read masses from how correlations decrease, and form the ratio $m/\sqrt{\sigma}$

This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — A lattice mass is read from how fast the correlation of time-slice operators decreases in imaginary time, and the value is an upper bound approached from above. Smoothing and the variational method speed up the approach. The string tension is extracted from periodic strings by assuming a string correction, and the ratio $m(0^{++})/\sqrt{\sigma}$ is 3.78–3.84 for $SU(2)$ and 3.41–3.61 for $SU(3)$ in primary sources (3.35–3.61 including Meyer's value from secondary information).

Prerequisite chapters — 5-07 The Monte Carlo method. We use the transfer matrix (5-03), the definition of the mass gap (4-11) and the variational principle (1-10). β is the inverse lattice coupling of 5-01, a different quantity from the β function $\beta(g)$ of 4-06.

Said first — Every number in this chapter is at the level of numerics, not a theorem. They are measured at finite box, β and time window, and pass through extrapolation and the string picture. The demonstration on this machine is three-dimensional $\mathbb{Z}_2$ gauge theory, not four-dimensional Yang–Mills. This chapter shows nothing about the existence of the mass gap.

Order of this chapter

1. A mass is how a correlation decreases — the effective mass is an upper bound
2. Name tags for glueballs — the rotation group of the cube
3. The noise does not decrease — the relative error grows exponentially
4. Fattening the operators — smoothing in space only
5. The variational method — optimising the upper bound
6. Measuring the string tension — Creutz ratios and periodic strings
7. A demonstration on this machine — three-dimensional $\mathbb{Z}_2$

8. Numbers from primary sources — a table of $m/\sqrt{\sigma}$
9. What the numbers say, and what they do not — upper bounds, windows, extrapolation, the string assumption

---

01

A mass is how a correlation decreases — the effective mass is an upper bound

Advancing time by one step on the lattice is applying the transfer matrix $T = e^{-aH}$ once (5-03). The eigenvalues of H are read from how the correlation of an operator $O(t)$, built from quantities on one time slice t, decreases. O is summed over space to give zero momentum, and its vacuum expectation value is subtracted (lattice units $a = 1$).

$$C(t) = \langle O(t)\, O(0)\rangle - \langle O\rangle^2 = \sum_{n\geq 1} |\langle n|O|\Omega\rangle|^2\, e^{-E_n t}, \qquad m_{\text{eff}}(t) = \ln \frac{C(t)}{C(t+1)} \quad (1)$$

$|\Omega\rangle$ is the vacuum, and E_n are the eigenvalues measured from the vacuum. The weights $|\langle n|O|\Omega\rangle|^2$ are squared absolute values, so they are 0 or more. The eigenvalues e^{-E_n} are positive reals because link-reflection positivity ($\beta \geq 0$) gives $T \geq 0$, and for $\beta > 0$, T does not have 0 as an eigenvalue (5-03 § chain, § transfer matrix). A sum of positive exponentials with weights of 0 or more is log-convex (by Cauchy–Schwarz, $C(t)^2 \leq C(t-1)C(t+1)$), so the effective mass $m_{\text{eff}}(t)$ decreases monotonically in t and approaches from above the energy of the lightest state that O touches Known . A value read at finite t is an **upper bound**, and no lower bound comes out (for an example with a finite window, 4-11 § window). On a lattice periodic in the time direction (length N_t), each term gets a backward $e^{-E_n(N_t - t)}$. For $N_t = 32$, $t \leq 7$ of § 07 it is negligible.

Example. For $C(t) = 0.8\, e^{-0.77t} + 0.2\, e^{-1.8t}$, m_{eff} goes down as 0.908, 0.824, 0.790, 0.777, 0.773, 0.771 for $t = 0, \ldots, 5$ Computed .

02

Name tags for glueballs — J^{PC} and the rotation group of the cube

Gauge-invariant states are called glueballs and are classified by spin J, parity P and charge conjugation C. P is the operation reversing the spatial coordinates $\vec{x} \to -\vec{x}$ (with the links mapped accordingly), and C is the operation mapping links to their

complex conjugates $U \to U^*$. For the trace of a loop, $\operatorname{tr} U^* = \operatorname{tr} U$, which is the trace of the loop with the opposite orientation, so C reverses the orientation of loops. The lightest is thought to be 0^{++}, and the next 2^{++} Physics .

The rotation group of the cube and its characters

The only rotations left on the lattice are the 24 that carry a cube onto itself, and the irreducible representations of this group O are A_1, A_2, E, T_1, T_2 (dimensions 1, 1, 2, 3, 3). "Name tag" is a metaphor; precisely, it is the pair of the irreducible representation of O to which a state belongs and the eigenvalues (± 1) of P and C.

The representation of spin J splits on the lattice into a sum of these. Characters decide how it splits. The character $\chi(g)$ of a representation is the trace of the matrix representing the element g. The 24 elements of O fall into five sets on which characters take the same value (conjugacy classes): the identity (1), 90° rotations about axes through face centres (6), 180° rotations about the same axes (3), 120° rotations about diagonal axes (8), and 180° rotations about axes through edge midpoints (6). The number of times a representation Γ appears is the following sum, which comes from a theorem, the orthogonality of characters of finite groups (here we only use it) Known .

$$n_\Gamma = \frac{1}{24} \sum_{g \in O} \chi_J(g)\, \chi_\Gamma(g), \qquad \chi_J(\theta) = \frac{\sin\big((2J+1)\theta/2\big)}{\sin(\theta/2)}$$

θ is the rotation angle of g. For example, for $J = 2$, $\chi_2 = (5, -1, 1, -1, 1)$, and $n_E = \frac{1}{24}(5 \cdot 2 + 3 \cdot 1 \cdot 2 + 8 \cdot (-1) \cdot (-1)) = 1$.

Known **Characters of O** The columns are the five conjugacy classes above (number of elements).

	Identity (1)	90° (6)	180° (3)	120° (8)	Edge 180° (6)
A_1	1	1	1	1	1
A_2	1	−1	1	1	−1
E	2	0	2	−1	0
T_1	3	1	−1	0	−1
T_2	3	−1	−1	0	1

Computed Known **Table 1** How the representation of spin J splits under the rotation group O of the cube (computed from the characters of the five conjugacy classes; we checked that the dimensions add up to $2J+1$).

J	0	1	2	3	4
Splitting	A_1	T_1	$E + T_2$	$A_2 + T_1 + T_2$	$A_1 + E + T_1 + T_2$

0^{++} is measured in A_1^{++}, but A_1 also contains $J = 4$. Lucini–Teper 2001 themselves raise the possibility that the excitation 0^{++*} is the lightest 4^{++}. 2^{++} appears in E^{++} and T_2^{++}, and the two agreeing is a sign that rotational symmetry is restored (4.582(14) and 4.578(11) for $SU(\infty)$ in Athenodorou–Teper 2021) Physics .

Example. The sum of spatial plaquettes in the three orientations $P_{12} + P_{23} + P_{31}$ is A_1^{++}, and the difference $P_{12} - P_{23}$ is E^{++}. For $SU(2)$, $U^* = gUg^{-1}$ ($g = i\sigma_2 \in SU(2)$, where σ_2 is the Pauli matrix of 1-04), so C is the same as making the same gauge transformation at every point. Hence every gauge-invariant state has $C = +$.

03

The noise does not decrease — the signal falls exponentially, and the fluctuation stays

$C(t)$ is estimated by an average over N configurations (how configurations are made and their errors: 5-07). The variance of the product $O(t)O(0)$, which sets the error, approaches a constant of about $\langle O^2 \rangle^2$ as t grows, and does not decrease. The signal falls as $e^{-E_{\min} t}$, so

$$\frac{\delta C(t)}{C(t)} \approx \frac{\text{const.}}{\sqrt{N}} e^{+E_{\min} t} \quad (2)$$

This argument is standard in lattice spectroscopy and is said to go back to Parisi 1984 Known . Glueballs have a large $E_{\min}$, and 0^{++} also needs the vacuum subtracted, so it bites especially hard.

Example. In the demonstration of § 07 ($E_{\min} \approx 0.77$, about 35,000 configurations), the relative errors of the plaquette $C(t)$ were 0.004, 0.011, 0.025, 0.050, 0.121, 0.214, 0.379, 0.657 for $t = 0, \ldots, 7$ Computed . The factor 1.8–2.9 per step is of the same order as $e^{0.77} \approx 2.2$.

04

Fattening the operators — smoothing and blocking in space only

A plaquette has size a, but the physical length $1/\sqrt{\sigma}$ is about $7.5a$ for $SU(2)$ at $\beta = 2.60$ (Table 2). So we "fatten" the operators. Precisely, we replace links by averages over nearby paths, so that the operator overlaps strongly with extended states.

$$U_i(x) \longmapsto \text{Proj}\Big[U_i(x) + \alpha \sum_{j \neq i} \big(U_j(x)\, U_i(x+\hat{j})\, U_j(x+\hat{i})^\dagger + (\text{reversed}) \big) \Big] \quad (3)$$

i, j run over spatial directions only, and $\mathbf{Proj}$ is the operation that brings the sum back to the group. There is also blocking, which builds links of length $2a$. Lucini–Teper 2001 cite Teper 1987 and Albanese et al. 1987 (APE smoothing) for these as standard methods Known . Averaging only in space is the key: the operator stays on one time slice, so the positive weights of (1) and "approaching from above" are kept.

Quality is measured by the fraction c_0^2 of the weight in the ground state. For the best operators of Lucini–Teper 2001, for 0^{++} it was from about 0.95 ($SU(2)$) to about 0.90 ($SU(5)$) on coarse lattices, and from about 0.92 to about 0.85 on fine lattices. For periodic strings it is from about 0.90 to about 0.75 on both. The larger N, the worse the overlap Physics .

For the three-dimensional $\mathbb{Z}_2$ of § 07, the rough fraction $C(2)e^{2m}/C(0)$ ($m = 1/1.296$; it comes out on the large side) was 0.57 for the plaquette and 0.85 for the 3×3 loop. With a smoothing in which $\mathbf{Proj}$ is a majority vote of the link and two staples, 0.59 rose to 0.66 in one step Computed .

05

The variational method — the generalised eigenvalue problem, and optimising the upper bound

We line up candidates $O_1, \dots, O_N$ and let the correlations themselves choose the best linear combination. For the correlation matrix $C_{ij}(t) = \langle O_i(t) O_j(0) \rangle - \langle O_i \rangle \langle O_j \rangle$ we solve

$$C(t)\, v_n = \lambda_n(t, t_0)\, C(t_0)\, v_n, \qquad E_n^{\mathrm{eff}}(t) = \ln \frac{\lambda_n(t, t_0)}{\lambda_n(t+1, t_0)} \quad (4)$$

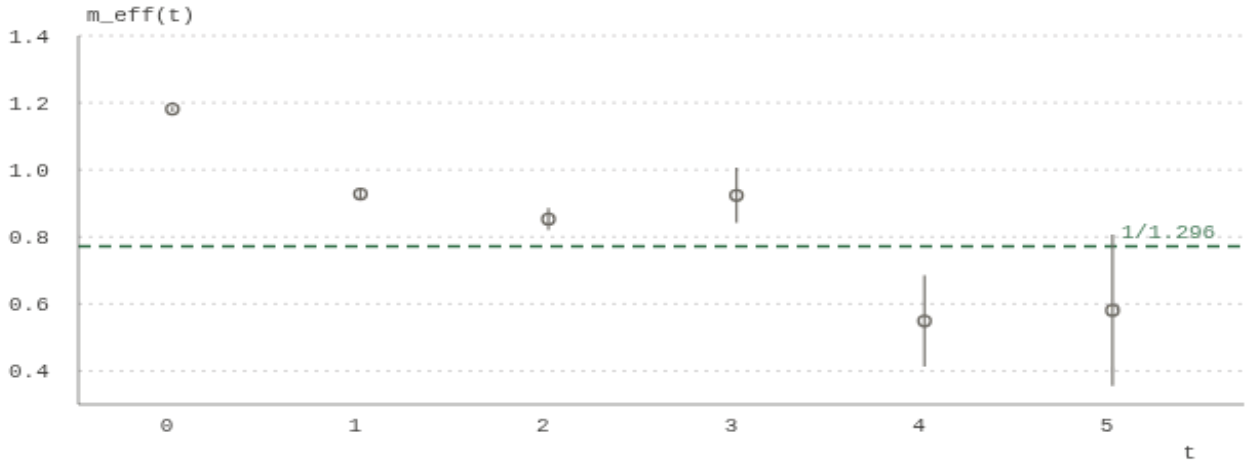
The variational idea began with Wilson's 1981 talk, and is said to have been used for glueballs by Berg 1982 and Michael–Teasdale 1983. The form of the generalised eigenvalue problem with t_0 and the estimate of its corrections are attributed to Lüscher–Wolff 1990 (according to the introduction of Blossier et al. 2009) Known . Blossier et al. 2009 show that taking $t_0 \geq t/2$ makes the corrections vanish as $\exp\!\big(-(E_{N+1} - E_n)t\big)$. The $t_0 = 0$ of Figure 1 is outside this condition, and is used only in the sense of the upper bound of formula (5).

For $t_0 = 0$, $t = 1$ it is the Rayleigh–Ritz variational principle itself (1-10, 2-05). For $\psi = \sum_i v_i O_i |\Omega\rangle$ (with the vacuum component subtracted)

$$\lambda_{\max}(1, 0) = \max_{v} \frac{v^\dagger C(1)\, v}{v^\dagger C(0)\, v} = \max_{\psi} \frac{\langle \psi | T | \psi \rangle}{\langle \psi | \psi \rangle} \leq e^{-E_1} \quad (5)$$

so $-\ln \lambda_{\max}$ is the best upper bound within the space spanned by the candidates Known . The variational method improves the upper bound, but does not change the fact that it is an upper bound.

Figure 1 The effective mass of 0^{+} in three-dimensional $\mathbb{Z}_2$. Choosing an operator with the buttons changes the points. The dashed line is $1/1.296$ from the primary source




```
1×1: m_eff(0) = 1.181, m_eff(3) = 0.924±0.082 (reference 0.772)
```

Computed The numbers in this figure were computed on this machine.

Computed The numbers of Figure 1 (jackknife errors, 39 bins. The variational basis is 1×1, $1 \times 2 + 2 \times 1$, 2×2, 3×3, with $t_0 = 0$).

t	0	1	2	3	Fraction
1×1	1.181(7)	0.928(17)	0.853(33)	0.924(82)	0.57
2×2	0.940(6)	0.858(14)	0.841(29)	0.842(70)	0.78
3×3	0.883(6)	0.818(13)	0.831(29)	0.795(61)	0.85
Variational (4)	0.860(6)	0.824(13)	0.831(29)	0.803(63)	0.87

How to read it. At $t = 0$ the variational value is lowest (formula (5)); every column goes down to 0.80–0.85 at $t = 2, 3$ and approaches the reference value 0.772 from above. For $t \geq 4$ the errors exceed 0.1 (Figure 1). The rightmost column is the rough fraction (§ 04).

06

Measuring the string tension — Creutz ratios and periodic strings

The string tension σ is the slope of the energy $V(R)$ of two separated static charges. An $R \times T$ Wilson loop (5-01) decreases like $e^{-V(R)T}$, and at long distances the following form is used (17-03) Physics .

$$V(R) = \sigma R + \mu - \frac{\pi(d-2)}{24R}, \qquad \chi(R,R) = -\ln \frac{W(R,R)\,W(R-1,R-1)}{W(R,R-1)\,W(R-1,R)} \quad (6)$$

d is the spacetime dimension. In the Creutz ratio χ (attributed to Creutz 1980) the perimeter and corner contributions cancel. Assuming the form (6) and ignoring finite-T corrections, $\chi(R,R) \approx V'(R-\tfrac{1}{2}) = \sigma + \pi(d-2)/\big(24(R-\tfrac{1}{2})^2\big)$ approaches σ from above Physics .

Periodic strings

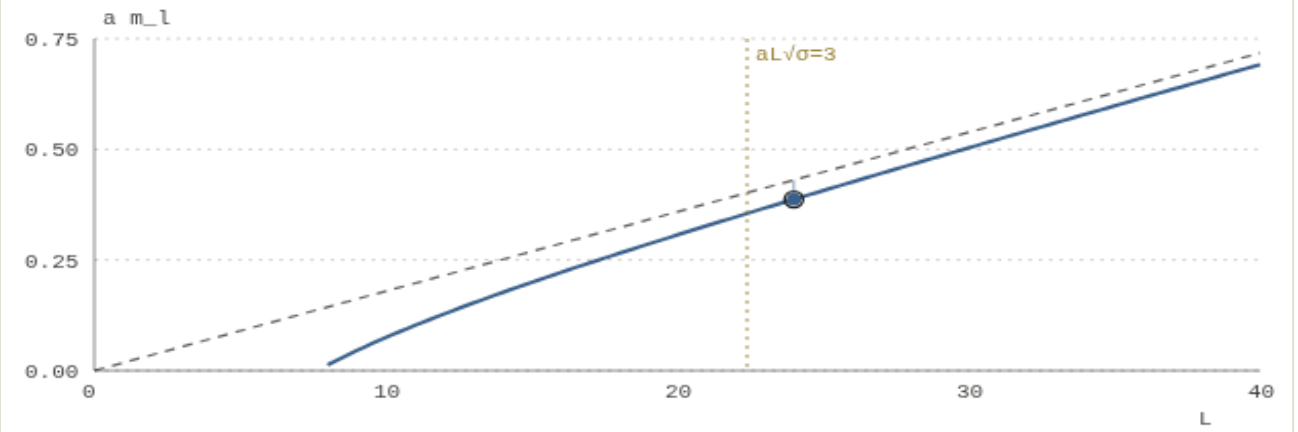
The other way is to read, by the same method as § 01, the mass m_l of a closed flux tube going once around the spatial period L (a periodic string). Lucini–Teper 2001 extracted σ in four dimensions from the following formula (the correction is in general $\pi(d-2)/(6L)$; in three dimensions $\pi/(6L)$; formula (7) of 17-03 § closed strings).

$$a\,m_l = a^2\sigma L - \frac{\pi}{3L} \quad (7)$$

The authors write 'in using eqn(4) we are assuming that we have linear confinement … Nambu-Gotto universality class' (Nambu–Goto is the approximation that treats the string as a membrane with action proportional to its area; 17-03), and caution that with two values of L 'we cannot claim to have evidence for both the 1/L functional form and for its coefficient'. Only lattices with $aL\sqrt{\sigma} \geq 3$ are used. The authors write that to study the string correction itself, periodic strings are much better than Wilson loops: the coefficient of the correction is larger ($\pi/3$ against $\pi/12$), and the Coulomb term of static charges (which has the same $1/r$ form and dominates at short distances) does not mix in.

Periodic strings do not mix with glueballs because of centre symmetry. Multiply all links in direction i crossing the plane $x_i = c$ of spatial direction i (at all times) by an element of the centre of the group (an element commuting with every element; -1 for $SU(2)$; 1-04). Every plaquette either does not contain these links or contains them as a there-and-back pair, so the action does not change. A loop going once around in direction i crosses the plane once and picks up one centre element (for $SU(2)$ its sign changes), while contractible closed loops do not change. In the confining phase this symmetry is not broken, so the correlation between the two is 0 Known .

Figure 2 Moving L changes the difference between the string mass (solid) and σL (dashed), and the deviation of the naive $\sqrt{m_l/L}$. $SU(2)$, $\beta = 2.60$. The circle is the measured value at $L = 24$ in the same paper



L = 24: a m_l = 0.3870, correction fraction 10.1%, naive a√σ = 0.1270 (94.8% of the true value)

Computed The numbers in this figure were computed on this machine (formula (7)).

Computed **Table 2** $a\sqrt{\sigma}$ recomputed by formula (7) from the am_l of Table 1 of Lucini–Teper 2001 ($SU(2)$). The master values of the six β and masses are in 16-05 Table 3.

β	L	am_l	$a\sqrt{\sigma}$	Correction fraction	$L\,a\sqrt{\sigma}$
2.25	8	1.301(17)	0.4231(25)	9.1%	3.38
2.30	10	0.861(11)	0.3108(18)	10.8%	3.11
2.40	12	0.745(9)	0.2634(14)	10.5%	3.16
2.475	16	0.585(8)	0.2016(12)	10.1%	3.23
2.55	20	0.453(4)	0.1590(6)	10.4%	3.18
2.60	24	0.387(4)	0.1340(6)	10.1%	3.21

How to read it. The $a\sqrt{\sigma}$ of the paper are reproduced by formula (7) within rounding (within 1 in the error of the last digit). $L\,a\sqrt{\sigma}$ is 3.1–3.4 at every β, so the box is almost constant in physical length. The correction is about 10% of σL, and leaving it out underestimates $\sqrt{\sigma}$ by about 5% Computed .

07

A demonstration on this machine — three-dimensional $\mathbb{Z}_2$ gauge theory

We check the same way of measuring in the simplest gauge theory: three-dimensional $\mathbb{Z}_2$ gauge theory, with ± 1 on the links and the Wilson action $S = -\beta \sum_p U_p$. It is known to map by duality to the three-dimensional Ising model (attributed to Wegner 1971; for how the correspondence is used, Agostini et al. 1997). The mass of 0^+ corresponds to the inverse correlation length of the Ising model, and the string tension to the interface tension (the free energy per unit area carried by the boundary between up and down regions; 2-09 § low temperature) Known .

Agostini et al. 1997 give, at $\beta = 0.72484$, $\xi_{0^+} = 1.296(3)$ (stable for $N_s \geq 16$, Table 2 of that paper), $\sigma = 0.05558(10)$ and $\sigma\xi^2 = 0.0934(6)$ (Table 6, from the literature on interface tension), so $m/\sqrt{\sigma} = 3.27$ Physics .

On this machine we made about 35,000 heat-bath updates on a $16^2 \times 32$ lattice (three independent runs), measuring the operators of § 04 and § 05 and the periodic string every time, and Wilson loops once every 5 (majority-vote smoothing in two other runs).

Computed **Table 3** Demonstration and reference values for three-dimensional $\mathbb{Z}_2$ at $\beta = 0.72484$. σ is the $m_{\mathrm{eff}}(2)$ of the periodic string put into $m_l = \sigma L - \pi/(6L)$. The rough guide for Creutz ratios is the crude approximation $\chi \approx V'(R - \frac{1}{2})$, without corrections for finite T (the loops being square).

Quantity	This machine	Reference value
m_{0^+} (variational, $t = 3$)	0.803(63)	0.7716(18)
m_l ($L = 16$)	0.872(39)	0.857 (from the formula)
σ	0.0566(24)	0.05558(10)
$m/\sqrt{\sigma}$	3.38(27)	3.27
$\chi(R,R)$, $R = 4, 5, 6$	0.0698(6), 0.0645(11), 0.0664(26)	Guide $\sigma + \frac{\pi}{24(R-1/2)^2}$: 0.066, 0.062, 0.060

How to read it. Both the mass and the string tension agree with the reference values within errors. The central value of the mass lies above the reference value, but the difference is about 0.5 times the error and cannot be distinguished. Being an upper bound is a property of the exact expectation value $C(t)$; an estimate from finitely many configurations can also come out below because of noise ($t \geq 4$ in Figure 1). The value at $t = 3$ with $t_0 = 0$ is a ratio of two largest eigenvalues, and is not a strict upper bound like formula (5) either. The Creutz ratios are 16–19% larger than σ even at $R = 5, 6$, slightly above the guide. The periodic string gave σ to 4% accuracy from the single point $L = 16$.

The ratio differs from theory to theory. The continuum value of Agostini et al. 1997 is $m_{0^+}/\sqrt{\sigma} = 3.08(3)$, and the value for three-dimensional $SU(2)$ that Table 9 of that paper cites from other literature is 4.763(31) Physics .

08

Numbers from primary sources — a table of $m/\sqrt{\sigma}$

These are continuum-limit values from primary sources for four-dimensional pure Yang–Mills theory (for the extrapolation procedure, 5-09 § ratios).

Physics Computed **Table 4** Continuum limits of $m/\sqrt{\sigma}$, and $2^{++}/0^{++}$ divided on this machine (errors combined as independent. Only the Meyer row is the ratio from the source's table; divided on this machine it is 1.47(4)). LT = Lucini–Teper 2001, AT = Athenodorou–Teper 2020 ($SU(3)$) and 2021 ($SU(2)$, $SU(\infty)$; 2^{++} is E^{++}). $SU(\infty)$ is the $1/N^2$ extrapolation.

Source, group	0^{++}	2^{++}	$2^{++}/0^{++}$
LT $SU(2)$	3.844(61)	5.59(15)	1.45(5)
LT $SU(3)$	3.607(87)	5.13(22)	1.42(7)
LT $SU(4)$	3.49(14)	5.21(21)	1.49(9)
LT $SU(5)$	3.38(16)	4.88(38)	1.44(13)
LT $SU(\infty)$	3.37(15)	4.93(30)	1.46(11)
AT $SU(2)$	3.781(23)	5.349(20)	1.415(10)
AT $SU(3)$	3.405(21)	4.894(22)	1.437(11)
AT $SU(\infty)$	3.072(14)	4.582(14)	1.492(8)
Meyer $SU(3)$ (via GKN II)	3.347(68)	4.916(91)	1.46(5)

Morningstar–Peardon 1999 and Chen et al. 2006 report for $SU(3)$, in units of the Sommer scale r_0 (the distance at which the force F between static charges satisfies $r_0^2 F(r_0) = 1.65$), $r_0 m(0^{++}) = 4.21(11)(4)$ and $4.16(11)(4)$; dividing by the $r_0\sqrt{\sigma} = 1.160(6)$ of AT 2020 gives 3.63(10) and 3.59(10) Computed . It is a conversion combining numbers from different sources and is only a rough guide.

How to read it. The 0^{++} ratio goes down with N, and $2^{++}/0^{++}$ is 1.4–1.5 for every N. Even for the same group the sources differ: the difference 0.063 between LT and AT for $SU(2)$ is about 1 times the combined error, and the difference 0.20 for $SU(3)$ about 2.3 times Computed . This chapter does not identify the cause of the differences. Conversion to GeV is treated in 5-09 § detour, and large N in 17-04.

09

What the numbers say, and what they do not — upper bounds, windows, extrapolation, the string assumption

Table 4 is numerical support for the physical view that the positive gap is about 3 to 4 times $\sqrt{\sigma}$. To make it a mathematical statement, four things remain.

1. **Upper bounds.** The $a\,m$ at each β is an upper bound at finite t or a fit to a plateau (§ 01, § 05). No lower bound comes out of a finite window (4-11 § window).

2. **Boxes and quantum numbers.** In a finite box, two glueballs or a pair of periodic strings also appear with the same quantum numbers (Morningstar–Peardon 1999 say in their abstract that they distinguished them). Spin is identified through the representations of the cube (§ 02).
3. **Extrapolation.** Continuum-limit values are extrapolations that assume the form of the a^2 corrections (5-09, 5-05). That the limits of the ratios of masses and string tensions exist has not been shown (the construction published in 2026-11 built limits of correlations of local fields, but does not treat the mass gap or Wilson loops; 6-04).
4. **The string assumption.** σ is extracted by assuming linear confinement and corrections of Nambu–Goto type (§ 06). Confinement at large β for four-dimensional non-abelian groups has not been proved, so using this σ as evidence of confinement is circular (17-03).

What the mass gap problem (4-11) asks for is the continuum theory and a **lower bound** on the gap, while every method in this chapter approaches from above. What to hold fixed when taking the limit is treated in 5-09.

What this chapter can say

	Content
Said	That the effective mass decreases monotonically and gives an upper bound on the lowest energy. That the generalised eigenvalue problem at $t_0 = 0, t = 1$ is the Rayleigh–Ritz upper bound. That smoothing in space only keeps the operators on one time slice. How J splits under the cube group (Table 1) Known .
At the level of numerics	The $m/\sqrt{\sigma}$ of primary sources (Table 4) and the reproduction of $a\sqrt{\sigma}$ by formula (7) (Table 2) Physics Computed . That the demonstration with three-dimensional $\mathbb{Z}_2$ agreed with the reference values within errors (Figure 1, Table 3) Computed .
Not said	That the continuum limit of four-dimensional Yang–Mills and the limit of $m/\sqrt{\sigma}$ exist; a lower bound on the gap. An independent check of the string correction. The cause of the spread in the $SU(3)$ values. This chapter shows nothing about the existence of the mass gap.

Sources and reproduction

Item	Kind	Source / tools
am_l, $a\sqrt{\sigma}$ and am of $SU(N)$ (Table 1 of the paper), formula (4), continuum values (Table 5 of the paper), the operators and the variational procedure, c_0^2, the question about 4^{++}, verbatim quotations	Primary (text)	B. Lucini, M. Teper, JHEP 0106 (2001) 050, arXiv:hep-lat/0103027
$m/\sqrt{\sigma}$, $r_0\sqrt{\sigma}$, E^{++} and T_2^{++}	Primary (text checked against the checked material)	A. Athenodorou, M. Teper, JHEP 11 (2020) 172, arXiv:2007.06422; JHEP 12 (2021) 082, arXiv:2106.00364
$r_0 m$, abstract	Primary (text checked against the material)	C. Morningstar, M. Peardon, Phys. Rev. D 60 (1999) 034509, arXiv:hep-lat/9901004; Y. Chen et al., Phys. Rev. D 73 (2006) 014516, arXiv:hep-lat/0510074
Meyer's $SU(3)$ values and the ratio 1.46(5)	Secondary (original not obtained)	The values of H. B. Meyer, arXiv:hep-lat/0508002, from the table of U. Gürsoy, E. Kiritsis, F. Nitti, JHEP 0802 (2008) 019, arXiv:0707.1349
ξ_0+ of three-dimensional $\mathbb{Z}_2$ (Table 2 of the paper), σ and $\sigma\xi^2$ (Table 6), 3.08(3), how the dual correspondence is used	Primary (text)	V. Agostini, G. Carlino, M. Caselle, M. Hasenbusch, Nucl. Phys. B 484 (1997) 331, arXiv:hep-lat/9607029
4.763(31) for three-dimensional $SU(2)$	Secondary (via Table 9 of Agostini et al. 1997; original not checked)	Table 9 of the same
The corrections of the generalised eigenvalue problem and the condition $t_0 \geq t/2$, the history of the variational method (introduction of the paper)	Primary (text)	B. Blossier et al., JHEP 0904 (2009) 094, arXiv:0902.1265
Starting points of the variational method (Wilson's 1981 talk, Berg 1982)	Bibliographic data only (from the introduction of Blossier et al. 2009; original not checked)	Quotation in the same introduction

Item	Kind	Source / tools
The variational method, smoothing, blocking, Creutz ratios, the noise argument, the string correction, duality	Bibliographic data only (attributions are "said to be")	Michael–Teasdale, Nucl. Phys. B 215 (1983) 433 (bibliographic data from the citation in Blossier et al. 2009); M. Lüscher, U. Wolff, Nucl. Phys. B 339 (1990) 222; M. Albanese et al., Phys. Lett. B 192 (1987) 163; M. Teper, Phys. Lett. B 183 (1987) 345; M. Creutz, Phys. Rev. D 21 (1980) 2308; G. Parisi, Phys. Rept. 103 (1984) 203; M. Lüscher, Nucl. Phys. B 180 (1981) 317; F. J. Wegner, J. Math. Phys. 12 (1971) 2259
Definition of the Sommer scale $r0$	Bibliographic data only (the defining formula checked against the checked material)	R. Sommer, Nucl. Phys. B 411 (1994) 839, arXiv:hep-lat/9310022
Table 1, Table 3 and Figure 1 (heat bath, jackknife, generalised eigenvalue problem), Table 2 and Figure 2, the ratios and conversions in Table 4, the example in § 01	Computed	python3 (numpy, scipy). Scripts and output are kept in the record of this book's investigation. The majority-vote smoothing was run as three runs; one was cut off and its record was damaged, so it is not used

Next chapter to read: 5-09 What to hold fixed when taking the limit (turning numbers in lattice units into physical quantities, and what to hold fixed in going to $\beta \to \infty$).

What to hold fixed when taking the limit — what gives lattice numbers a unit is a scale we choose

This chapter organises established knowledge. § 05 and § 06 are a map of open questions.

What this chapter shows — Every number a lattice calculation returns is a dimensionless number "in units of the lattice spacing", and physical units are attached only once one quantity is held fixed. What has meaning independently of how one fixes it are dimensionless ratios (for example $m/\sqrt{\sigma}$). The lattice correlation-length problem (Chatterjee's Problem 5.1) asks for the correlation length in lattice units to diverge, and "the gap in lattice units is uniformly positive" is the opposite of that.

Prerequisite chapters — 5-05 Weak coupling and the continuum limit ($a(\beta)$ and the lattice version of asymptotic freedom), 4-11 The definition of the mass gap (the gap, the correlation length, uniformity in the volume). We use the notation of 5-01 (β, links, Wilson loops) and "a critical point is where the lattice disappears" from 2-08 § continuum limit.

Said first — This chapter is about how to pose the question. Both the lattice correlation-length problem and the continuum string-tension problem (Chatterjee's Problem 5.2) are unsolved, and this chapter shows nothing about the existence of the continuum limit or about the mass gap. All numbers are lattice calculations from primary sources (at the level of numerics), and recalculations of them on this machine.

Order of this chapter

1. Lattice numbers have no unit — all one can measure are numbers in units of the lattice spacing
2. Choosing one scale — hold fixed one of σ, r_0, the correlation length or Λ (Detour: MeV is borrowed)
3. Dimensionless ratios — quantities from which the choice of scale drops out, and extrapolation in a^2
4. Misreading "a uniform gap" — hold the gap in lattice units fixed, and the physical mass goes to infinity

5. Reading the lattice correlation-length problem — the correlation length finite at each β, and divergent as $\beta \to \infty$
6. The continuum string-tension problem and where the ratio comes from — another formulation, fixing the scale with Wilson loops
7. Uniform in what — three kinds of "uniform": volume, lattice units, physical units
8. Fixing alone is not enough — three types of trivial limit
9. The order of limits — first the volume, then β. The shortcut on the numerical side

---

01

Lattice numbers have no unit — all one can measure are numbers in units of the lattice spacing

Pure Yang–Mills lattice theory (5-01 § action) has only one parameter: the β multiplying the action. The lattice sits on the integer points $\mathbb{Z}^4$, and nowhere is a "metre" or an "MeV" written.

So every number a lattice calculation returns is a number in units of one lattice step. A mass comes out only as am (the product of the lattice spacing a and the mass m), a string tension as $a^2\sigma$, and a correlation length as $\xi = 1/(am)$ (how many lattice spacings). In this book we write m_{lat} for the mass in lattice units and ξ for the correlation length in lattice units.

$$m_{\text{lat}}(\beta) = a\,m, \qquad \xi(\beta) = \frac{1}{m_{\text{lat}}(\beta)}, \qquad \sigma_{\text{lat}}(\beta) = a^2\sigma \tag{1}$$

Each left-hand side is a function of β alone, and has no dimension. The a and m on the right are not yet fixed separately at this stage. What is fixed on the lattice is only the product.

As a metaphor, it is like measuring things with a ruler on which the spacing of the marks is not written. One can say "4 marks", but not "so many centimetres". Precisely, given β, lattice theory determines a set of dimensionless functions such as m_{lat} and σ_{lat}, and a is defined only once a physical value is assigned to one of them.

Example. We copy the numbers of a lattice calculation from a primary source (Lucini–Teper 2001, $SU(2)$) into Table 1. At $\beta = 2.60$, $a\sqrt{\sigma} = 0.13395$ means "the square root of the string tension is 0.134 times the inverse lattice spacing", and $m_{\text{lat}} = 0.514$ means

"the mass of the lightest scalar glueball is 0.514 times" (we write this glueball 0^{++}: the state with spin 0, positive parity and positive charge conjugation; notation in 5-08); both are unitless numbers Physics .

Computed **Table 1** The numbers used in this chapter ($a\sqrt{\sigma}$, $m_{\text{lat}} = a\,m(0^{++})$ and lattice sizes are from Table 1 of Lucini–Teper. The six m_{lat} are the same as in 16-05 Table 3 (the master copy). $\xi = 1/m_{\text{lat}}$, and $L\,a\sqrt{\sigma}$, the side of the box measured in units of $1/\sqrt{\sigma}$, are computed on this machine).

β	Lattice	$a\sqrt{\sigma}$	m_{lat}	ξ	$L\,a\sqrt{\sigma}$
2.25	8^4	0.4231(25)	1.390(30)	0.72	3.38
2.30	10^4	0.3108(17)	1.090(33)	0.92	3.11
2.40	12^4	0.2634(14)	0.953(19)	1.05	3.16
2.475	16^4	0.2016(13)	0.754(10)	1.33	3.23
2.55	20^4	0.15896(63)	0.586(10)	1.71	3.18
2.60	24^4	0.13395(62)	0.514(8)	1.95	3.21

02

Choosing one scale — hold fixed one of σ, r_0, the correlation length or Λ

To attach units, one chooses a quantity with a dimension and fixes its value by convention. This is called **setting the scale**. If the chosen quantity is Q (of mass dimension) and the agreed value Q_*, the lattice spacing is fixed as follows.

$$aQ(\beta) := \frac{(aQ)(\beta)}{Q_*} \quad (2)$$

The numerator is a lattice number, the denominator a conventional number. Other quantities can be converted to physical units as $m = m_{\text{lat}}(\beta)/aQ(\beta)$. There are many ways to choose.

- **The string tension** σ: Creutz 1980, with 'hold fixed the string tension', saw numerically that holding the string tension fixed makes the bare coupling decrease logarithmically (abstract) Physics .
- **The Sommer scale** r_0: the length fixed by $F(r_0)\,r_0^2 = 1.65$ for the force F between static charges (Sommer 1994) Physics .
- **The correlation length itself**: Chatterjee 2018 § 5 explains sending the lattice spacing ϵ to 0 so that the correlation length approaches a finite positive value Known .

- **The Λ of asymptotic freedom**: the two-loop formula gives $a\Lambda_{\mathrm{L}}$ as a function of β, and Λ_{L} is held fixed (5-05, 4-06 § two loops).

The ratio of the lattice spacings made with two choices Q and Q' can be written with lattice numbers alone.

$$\frac{aQ(\beta)}{aQ'(\beta)} = \frac{(aQ)(\beta)}{(aQ')(\beta)} \cdot \frac{Q'_*}{Q_*} \quad (3)$$

The first factor on the right is the lattice value of the dimensionless ratio Q/Q'. If it settles to a constant as $\beta \to \infty$, the two scales differ only by a constant factor in the limit. That it settles is not a theorem but a matter supported by numerics (§ 03).

Example. We compare the scale fixed by the string tension with the scale fixed by the two-loop Λ_{L}. For $SU(2)$, with $g^2 = 4/\beta$, $b_0 = 11N/(48\pi^2)$ and $b_1 = \frac{34}{3}\left(\frac{N}{16\pi^2}\right)^2$,

$$a\Lambda_{\mathrm{L}} = (b_0 g^2)^{-b_1/(2b_0^2)}\, e^{-1/(2b_0 g^2)} \quad (4)$$

Dividing the $a\sqrt{\sigma}$ of Table 1 by (4) gives Table 2 (higher-order corrections to (4) are not included).

Computed **Table 2** The ratio $\sqrt{\sigma}/\Lambda_{\mathrm{L}}$ of the two scales for $SU(2)$.

β	2.25	2.30	2.40	2.475	2.55	2.60
$a\Lambda_{\mathrm{L}} \times 10^3$	6.704	5.914	4.600	3.809	3.152	2.777
$\sqrt{\sigma}/\Lambda_{\mathrm{L}}$	63.1	52.6	57.3	52.9	50.4	48.2

At $\beta = 2.25$ and 2.30 the lattice is coarse ($a\sqrt{\sigma} > 0.3$) and the ratio goes up and down, so comparisons are made at 2.40 and above. Between $\beta = 2.40$ and 2.60 the ratio moves from 57 to 48, by about 16%. In this range of β, "holding σ fixed" and "holding the Λ_{L} of the bare coupling fixed" do not give the same scale. This distinction (scaling and asymptotic scaling) is treated in 18-04 § the continuum limit and a(β) and in 5-05 Physics .

Detour: MeV is borrowed

A subsection you may skip.

There is no experiment corresponding to pure $SU(2)$ Yang–Mills theory, and the values of $\sqrt{\sigma}$ or r_0 written in "MeV" are conventions brought in from outside. The same holds for $SU(3)$. Athenodorou–Teper 2020 apply $r_0 = 0.472(5)$ fm, which comes from lattice QCD with quarks, to the continuum-limit $r_0\sqrt{\sigma} = 1.160(6)$, obtain $\sqrt{\sigma} = 485(6)$ MeV, and write 0^{++} as 1.653(26) GeV Physics . In the same place the authors themselves caution 'Strictly speaking this is of course not possible: the pure gauge theory and QCD … are two different theories'.

As a metaphor, it is like measuring with a ruler borrowed from the country next door. Precisely, what pure Yang–Mills theory predicts are only the dimensionless ratios of the next section, and the values in GeV are those multiplied by one convention from outside.

03

Dimensionless ratios — quantities from which the choice of scale drops out, and extrapolation in a^2

Dividing two lattice numbers cancels the lattice spacing.

$$\frac{m\mathrm{lat}(\beta)}{\sqrt{\sigma\mathrm{lat}(\beta)}} = \frac{a\,m}{a\sqrt{\sigma}} = \frac{m}{\sqrt{\sigma}} \quad (5)$$

This ratio has the same value whichever scale is chosen, since the Q_* of formula (2) appears nowhere. Pure Yang–Mills theory has no dimensionless parameter left to choose freely after renormalisation (dimensional transmutation, 4-07 § one parameter disappears). So what has meaning as a prediction of the theory are only the limiting values of ratios such as (5) Physics .

At finite β lattice corrections remain in the ratios too. For the Wilson action their leading term is thought to be proportional to a^2, and Lucini–Teper extrapolate in the following form (formula (5) of the paper; the coefficient c is really a power series in g^2, but they state that they treat it as a constant) Physics .

$$\frac{m(a)}{\sqrt{\sigma(a)}} = \frac{m(0)}{\sqrt{\sigma(0)}} + c\,a^2\sigma \quad (6)$$

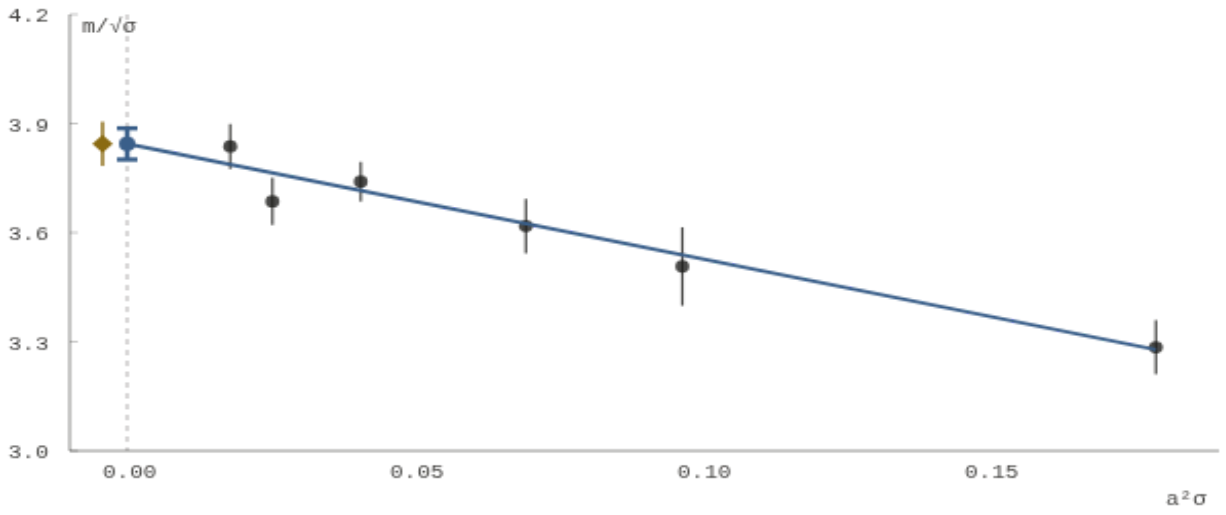
The procedure of the paper is to fit all points and, if the fit is poor, to drop points one at a time starting from the coarsest. Tables 3 and 4 redo the fits on this machine from the numbers of Table 1.

The fit is least squares with weights $1/\delta_i^2$ from the error δ_i of each point. The goodness of fit is measured by the following quantity.

$$\chi^2 = \sum_i \left(\frac{y_i - (\text{value on the line})}{\delta_i} \right)^2$$

Here y_i is the $m/\sqrt{\sigma}$ of each point. If χ^2 divided by the degrees of freedom (number of points − 2) is around 1, the fit matches the size of the errors. The confidence level is the probability that a χ^2 larger than this would occur by chance if the line were correct; the smaller it is, the worse the fit.

Figure 1 Vertical axis $m(0^{++})/\sqrt{\sigma}$, horizontal axis $a^2\sigma$ (coarser lattices to the right). Dropping points on the coarse side with the buttons changes the line and the intercept (the value extrapolated to $a = 0$). The diamond is the value 3.844(61) of Table 5 of the paper



6 points fitted: intercept 3.844(43), slope c = −3.16(53), χ²/dof = 0.60

Computed The numbers in this figure were computed on this machine (points from Table 3, lines from the fits of Table 4).

Computed **Table 3** The ratio at each β (errors propagated treating the errors of m_{lat} and $a\sqrt{\sigma}$ as independent). The rightmost column is the "misreading" value used in § 04 (with m_{lat} held at 1.390).

β	2.25	2.30	2.40	2.475	2.55	2.60
$a^2\sigma$	0.1790	0.0966	0.0694	0.0406	0.0253	0.0179
$m/\sqrt{\sigma}$	3.285(74)	3.507(108)	3.618(75)	3.740(55)	3.686(65)	3.837(62)
Misreading	3.29	4.47	5.28	6.89	8.74	10.38

Computed **Table 4** Weighted fits of formula (6).

Points used	Intercept $m(0)/\sqrt{\sigma(0)}$	Slope c	χ^2/dof	Confidence level
6 points (2.25–2.60)	3.844(43)	−3.16(53)	0.60	0.66
5 points (2.30–)	3.855(61)	−3.5(1.3)	0.77	0.51
4 points (2.40–)	3.849(72)	−3.3(1.8)	1.15	0.32
3 points (2.475–)	3.84(11)	−2.9(3.6)	2.28	0.13

How to read it. The lattice ratio moves from 3.29 to 3.84, but put on a straight line in $a^2\sigma$ the intercepts line up at 3.84–3.86. The central value of the 6-point intercept agrees with the 3.844 of Table 5 of the paper (the paper's error assignment is larger). The ratio mostly increases, approaching from below, but drops once, within errors, at $\beta = 2.475 \to 2.55$ Physics . A more recent calculation (Athenodorou–Teper 2021, 5-08 § numbers) gives 3.781(23) Physics .

What can be said here goes only as far as "assuming a^2 corrections, the numbers in this range line up at one value". It is not a proof that the limit exists. How the quantities are measured (the variational method, extracting the string tension) is treated in 5-08.

04

Misreading "a uniform gap" — hold the gap in lattice units fixed, and the physical mass goes to infinity

Here we come to an easily misread point. One is tempted to read the "mass gap" on the lattice as "the gap in lattice units $m_{\mathrm{lat}}(\beta)$ is at least some fixed positive number for all β". That the gap on a finite lattice is positive is nearly obvious (4-11 § finite boxes and uniformity), and it seems enough to make it uniform in β.

But a mass in physical units is the mass in lattice units divided by the lattice spacing.

$$m_{\mathrm{phys}}(\beta) = \frac{m_{\mathrm{lat}}(\beta)}{a(\beta)}, \qquad \inf_{\beta} m_{\mathrm{lat}}(\beta) \geq c > 0 \ \text{ and } \ a(\beta) \to 0 \implies m_{\mathrm{phys}}(\beta) \geq \frac{c}{a(\beta)} \to \infty$$

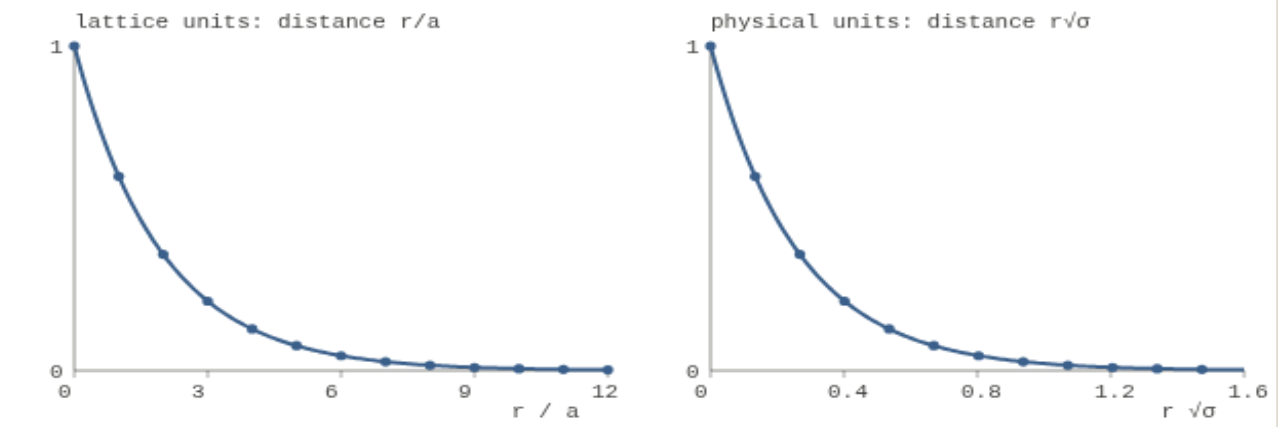
In the continuum limit $a \to 0$, so if the gap in lattice units is uniformly positive, the physical mass goes to infinity. Correlations shrink to zero width in physical distance, and the correlation between two separated points becomes 0. This is not a limit in which "the mass is positive" but one in which "nothing is transmitted".

For the mass to remain at a finite positive value in the continuum limit, m_{lat} must on the contrary go to 0 as fast as a.

$$m_{\mathrm{lat}}(\beta) \sim m_{\infty}\, a(\beta) \ \ (\beta \to \infty), \qquad \text{that is,} \quad \xi(\beta) \to \infty \ \text{(8)}$$

As a metaphor, when the marks on a ruler are made finer and finer, an object of fixed length should span more and more marks. If it always spans 2 marks, the object is shrinking. Precisely, (8) says that the correlation length in lattice units ξ diverges while the correlation length in physical units $a\xi$ is kept constant (the form seen exactly for the Gaussian chain of 2-08 § continuum limit).

Figure 2 Choosing β with the knob stretches the decay of the correlation $e^{-m_{\mathrm{lat}} r/a}$ on the left (distance in lattice units), while on the right (physical distance in units of $1/\sqrt{\sigma}$) the shape of the decay hardly changes. Pressing "Misreading" holds m_{lat} at its value for $\beta = 2.25$, and the decay on the right becomes steep



β = 2.60, a√σ = 0.134, m_lat = 0.514, ξ = 1.95, m/√σ = 3.84

Computed The numbers in this figure were computed on this machine ($a\sqrt{\sigma}$ and m_{lat} from Table 1; the curves are exponentials drawn on the spot).

From $\beta = 2.25 \to 2.60$ the lattice spacing shrinks by a factor of about 3.2 and m_{lat} by about 2.7. The two shrink at almost the same rate, so the ratio is almost constant. The misreading column of Table 3 keeps growing in proportion to $1/a$. Note that ξ in this range does not reach 2 lattice spacings, and the numbers are not looking at the limit itself Physics .

05

Reading the lattice correlation-length problem — the correlation length finite at each β, and divergent as $\beta \to \infty$

There is a statement of the direction of the previous section, written clearly as a mathematical question: the lattice correlation-length problem (Mass gap) in § 5 of Chatterjee's survey "Yang–Mills for probabilists" (arXiv:1803.01950, 2018) Known . We translate the gist.

> Consider any compact non-abelian Lie group $G \subseteq U(N)$ $(N \geq 2)$ and any infinite-volume limit of four-dimensional lattice gauge theory. For each $x \in \mathbb{R}^4$ take the nearest plaquette p_x, and let $f_\beta(x)$ be the correlation of W_{p_0} and W_{p_x}. Show that for every $\beta > 0$ there is some $\xi(\beta) \in (0, \infty)$ such that

$$\lim_{|x| \to \infty} \frac{\log f_\beta(x)}{|x|} = -\frac{1}{\xi(\beta)} \quad (9)$$

> Show further that $\lim_{\beta \to \infty} \xi(\beta) = \infty$.

The β of the lattice correlation-length problem corresponds to $\beta' = \beta/N$ in 5-01 § action, but as far as $\beta \to \infty$ is concerned there is no difference.

Here the correlation is the connected correlation $\langle W_{\gamma 1} W_{\gamma 2} \rangle - \langle W_{\gamma 1} \rangle \langle W_{\gamma 2} \rangle$, and the same section writes that the logarithm is taken because there may be a power correction in front of the exponential. It continues 'The number ξ represents the reciprocal of the mass of the lightest glueball in the theory.'

This question has two parts.

1. **Finiteness at each** β: $\xi(\beta) < \infty$, that is, correlations decrease exponentially in lattice units. This corresponds to the gap being positive in lattice units (4-11 § correlation length).
2. **Divergence as** $\beta \to \infty$: $\xi(\beta) \to \infty$. This is the direction of (8) in § 04. The gap in lattice units must go to 0.

The first part alone does not contradict "uniformly positive in lattice units". The second part does. The lattice correlation-length problem asks for the first part at each β separately, and the second part explicitly denies uniformity.

What this question does not ask

We also note what this question **does not ask**. It does not ask for the rate of divergence. What asymptotic freedom predicts is a form like $\xi \propto e^{\beta/(4Nb0)}$ (β in the Wilson normalisation of 5-01; the reciprocal of the one-loop formula $a \propto e^{-\beta/(4Nb0)}$ of 5-05, which in Chatterjee's β' reads $e^{N\beta'/(4Nb0)} = e^{\beta'/(4b0)}$). The ratio $m/\sqrt{\sigma}$ does not appear in the wording either. That the limiting theory exists, and that it satisfies the Osterwalder–Schrader axioms (4-10), are separate questions. Even if the lattice correlation-length problem were shown, the mass gap of the Clay problem statement (0-03 § mass gap) would not follow directly.

Example. The ξ of Table 1 grows from 0.72 to 1.95 with β. These numbers agree with the direction of the second part, but ξ growing over finitely many β is not a proof of divergence in the limit Physics .

> 06
>
> **The continuum string-tension problem and where the ratio comes from — another formulation, fixing the scale with Wilson loops**

The continuum string-tension problem (Continuum limit) in the same § 5 gives another way to fix the scale Known . For the rectangular Wilson loop $W_{R,T}$ (5-01

§ plaquettes and Wilson loops), it asks to show that there are sequences with $\epsilon(\beta) \to 0$ and $c(\beta) \to \infty$ as $\beta \to \infty$, and a nonzero constant d, such that for every R, T

$$\log\langle W_{R/\epsilon,\, T/\epsilon}\rangle = -c\,(R+T) - d\,RT + o(1) \quad (10)$$

$o(1)$ is a term that goes to 0 as $\beta \to \infty$. RT is the area of the rescaled loop, and the same section writes that d corresponds to the string tension of the continuum theory. $R+T$ is a quantity proportional to the perimeter of the rescaled loop (half the perimeter; the factor 2 is absorbed into c). The d here is the coefficient of the area, not the spacetime dimension.

In (10), all that is asked is that there exist an $\epsilon(\beta)$ "in whose scale the coefficient of the area of Wilson loops is constant". That is, it is a formulation that fixes the scale with the string tension. The explanatory paragraph of § 5 fixes the scale with the correlation length instead. The continuum string-tension problem does not presuppose the lattice correlation-length problem. For neither $\epsilon(\beta)$ has existence been shown.

One more caution. About the perimeter term $c(R+T)$, Chatterjee writes in the same section 'it is not clear to me whether the renormalization term c(R + T) is indeed necessary. It seems entirely possible that the limit in (5.2) holds without the renormalization term.' Subtracting the perimeter term is one proposal for how to pose this question.

Putting the two questions together with the same scale

The ratio $m/\sqrt{\sigma}$ appears when the two questions are put together with the **same** $\epsilon(\beta)$. Suppose the continuum string-tension problem holds and moreover $\epsilon(\beta)\,\xi(\beta) \to \ell \in (0, \infty)$. The physical correlation length is $\epsilon\xi \to \ell$, the physical mass $1/\ell$ and the string tension d, so

$$\frac{m}{\sqrt{\sigma}} = \frac{1}{\ell\sqrt{d}} = \lim_{\beta\to\infty} \frac{1/\xi(\beta)}{\sqrt{\sigma_{\mathrm{lat}}(\beta)}}, \qquad \sigma_{\mathrm{lat}}(\beta) \approx d\,\epsilon(\beta)^2 \quad (11)$$

and ϵ cancels and disappears. This is (5) of § 03 written in the limit. This construction is not in Chatterjee's text; it reads the two questions together. Note that the two limits existing simultaneously with the same ϵ is itself an assumption.

Example. The lattice values of the right-hand side of (11) are the $m/\sqrt{\sigma}$ row of Table 3. At $\beta = 2.60$ it is $(1/1.95)/0.134 \approx 3.84$.

> 07

Uniform in what — three kinds of "uniform": volume, lattice units, physical units

The phrase "a uniform gap" has no definite meaning until one says uniform in what. At least three kinds of "uniform" appear in discussions of the mass gap.

Uniform in what	Form of the formula	Meaning in the continuum limit	Where it appears
In the volume L	$\inf_L m(L) \geq c > 0$	Needed. The gap does not close as the box grows	Problem statement § 6.5 (quoted below), 4-11 § finite
In β, in lattice units	$\inf_\beta m_{\text{lat}}(\beta) \geq c > 0$	By (7), the physical mass is infinite. The opposite of what is wanted	The form given by strong-coupling estimates (small β)
In β, in physical units	$\inf_\beta m_{\text{lat}}(\beta)/a(\beta) \geq c > 0$	What is wanted. Meaningful together with the choice of $a(\beta)$	The form of the lattice correlation-length problem and the continuum string-tension problem put together (§ 06)

The Clay problem statement § 6.5 writes 'New ideas are needed to prove the existence of a mass gap that is uniform in the volume of space-time.' Known . The next sentence, 'Such a result presumably would enable the study of the limit as T4 → R4', has in mind the limit from compact spacetime T^4 to $\mathbb{R}^4$. The uniformity here is that of the first row, uniformity in the volume. It is a requirement separate from the third row, and both are needed.

Theorems of the form of the second row exist. Shen–Zhu–Zhu show, for $SU(N)$ with β small enough (in four dimensions $\beta < 1/48$ in the normalisation of the paper, where the inverse coupling is $N\beta$; in the Wilson β of this book $\beta < N^2/48$: 5-01 Table 4), uniqueness of the infinite-volume measure and exponential decay of correlations (abstract) Known . Earlier, Osterwalder–Seiler 1978 proved the area law rigorously at small β (according to the text of Chatterjee 2018 § 4) Known . The abstract of that paper says it also showed that at strong coupling (small β) the infinite-volume limit exists and is analytic in β (text not read). These are results with β held in a small range, theorems about a world with the lattice spacing held fixed. They sit on the side opposite the continuum limit ($\beta \to \infty$) (5-02 § tubes, 5-04).

The constants of these estimates do not extend to $\beta \to \infty$, and do not reach the "behaviour as the scale is made finer" asked about in the continuum limit.

For gauge theories with finite groups, exponential decay of correlations of gauge-invariant quantities at large β has been proved (Adhikari–Cao, abstract) Known . This is an example in which the form of the first part of the lattice correlation-length problem

holds at large β, but since the group is not continuous it is a different setting from the question asking for $\xi \to \infty$ in the second part (5-10).

08

Fixing alone is not enough — three types of trivial limit

Fixing one scale and taking $\beta \to \infty$ looks as if it would give some limit. But there is no guarantee that the limit obtained is the theory sought. We list three types that fail.

Type 1: the correlation length does not grow

If $\xi(\beta)$ is bounded, then whatever $\epsilon(\beta) \to 0$ is chosen, $\epsilon\xi \to 0$. The physical correlation length is 0 and the mass infinite: exactly the misread world of § 04. What remains in the limit is a field in which no two points are correlated (like white noise).

Type 2: the correlation length grows, but a ratio goes to 0 or infinity

If two quantities diverge at different rates, then once the scale is fixed with one, the other escapes to 0 or infinity. For the three-dimensional compact $U(1)$ lattice theory, Göpfert–Mack proved the area law at every β (according to the text of Chatterjee 2018 § 4) Known . Even so, the abstract of that paper says, the continuum limit is a free scalar field, and the string tension in units of the mass goes to infinity (Villain action; text not read) Known . It is an example that confines at every lattice β, yet the ratio $m/\sqrt{\sigma}$ goes to 0.

Type 3: the ratio is finite, but it collapses to a Gaussian

Even if a limit can be taken with the correlation length fixed, if it is a free field (a Gaussian measure) it is not a non-trivial theory. For four-dimensional Ising-type models and φ^4 (a scalar field with a quartic self-interaction added; 7-03), the scaling limits near the critical point have been shown to be Gaussian (Aizenman–Duminil-Copin 2021) Known . This is why the problem statement asks for a "non-trivial" theory (0-03 § non-trivial, 7-08).

Putting the three together, the conditions on the limit sought are the following set.

$$\xi(\beta) \to \infty, \qquad \frac{m_{\mathrm{lat}}}{\sqrt{\sigma_{\mathrm{lat}}}} \to R_\infty \in (0, \infty), \qquad \text{the limit is not Gaussian} \ (12)$$

The second uses confinement, while the problem statement itself does not require confinement (4-11 § definition). Even if another scale such as r_0 is used instead of the string tension, the second of (12) takes the form "the ratio of two dimensionful

quantities is finite and positive". On the numerical side, Tables 3 and 4 support the second for $SU(2)$ Physics .

09

The order of limits — first the volume, then β. The shortcut on the numerical side

The theory of the problem statement needs two limits: volume $L \to \infty$, and lattice spacing $a \to 0$ (in four dimensions, tied to $\beta \to \infty$) (5-01 § the map of Part 5). The lattice correlation-length problem is written in the order of taking the infinite-volume limit **first** at each β, and considering $\beta \to \infty$ afterwards.

$$\xi(\beta) = \xi\big(\lim_{L\to\infty} \mu L, \beta\big), \qquad \text{then } \beta \to \infty \quad (13)$$

Even for the first limit, Chatterjee 2018 § 3 writes 'the uniqueness (or non-uniqueness) is in general unknown for lattice gauge theories in dimensions higher than two when β is large.' Known . This is why the lattice correlation-length problem says "any infinite-volume limit".

Numerical calculations take another road. The rightmost column of Table 1, $L\,a\sqrt{\sigma}$, is the side of the box, the lattice side L (in lattice points) measured in physical units.

Each time the lattice is made finer, the number of lattice points is increased, keeping the physical box at about $3.2/\sqrt{\sigma}$. That is, numerical extrapolation takes the road "hold the physical volume fixed and take $a \to 0$". The paper uses only lattices with $aL\sqrt{\sigma} \geq 3$, as the lengths at which the formula that obtains the string tension by subtracting the string correction $\pi/(3L)$ from the mass of a closed flux tube (formula (4) of the paper) can be used. The paper writes that finite-volume effects on glueballs also become invisible at this size Physics .

The construction published in 2026-11 also takes the road of holding the volume fixed and taking a $\to 0$, and then takes the limit along a subsequence of volumes. It does not say that the order of the two limits may be exchanged (8-05).

That this road and the road of (13) reach the same value, that is, exchanging the two limits, is something checked numerically, not a theorem. The exchange needs an estimate uniform in the size of the box, which is the requirement of the first row of the table in § 07.

Part 23 of this book uses the direction set here (the gap in lattice units should go to 0) as its starting point. The correspondence with the definitions declared there is in 23-01 The line of construction and the diffusive-relaxation-time assumption.

What this chapter can say

	Content
Said	That lattice numbers are dimensionless numbers in units of the lattice spacing, and physical units are attached by fixing one quantity (formulas (1)–(3), sorting out the definitions). That if the gap in lattice units is uniformly positive in β, the physical mass goes to infinity as $a \to 0$ (formula (7), an elementary consequence). The wording of the two problems (Chatterjee's Problems 5.1 and 5.2), Chatterjee's reservation about the perimeter term, and that uniqueness of the infinite-volume limit at large β is unknown (the text of Chatterjee 2018) Known . That putting the two questions together with the same ϵ makes the ratio $m/\sqrt{\sigma}$ appear in the form (11) (algebra under assumptions).
At the level of numerics	That $m/\sqrt{\sigma}$ for $SU(2)$ lines up at 3.84–3.86 by extrapolation in a^2, that $\sqrt{\sigma}/\Lambda_L$ moves by 16% over $\beta = 2.40$–2.60, and that the numerical box is kept at almost constant physical size (recalculations on this machine of the tables of Lucini–Teper) Physics Computed . The $O(a^2)$ form is an assumption, and extracting the string tension involves an assumption about the string correction.
Not said	Both the lattice correlation-length problem and the continuum string-tension problem are unsolved; neither $\xi(\beta) \to \infty$, nor the existence of the limit of $m/\sqrt{\sigma}$, nor the exchange of the order of limits has been shown. The statements about the continuum limit of three-dimensional compact $U(1)$ are at the level of the abstract; the text has not been checked. This chapter shows nothing about the construction of the continuum limit or about the mass gap.

Sources and reproduction

Item	Kind	Source / tools
The lattice correlation-length problem and the continuum string-tension problem (Problems 5.1 and 5.2), the sentence on uniqueness in § 3, the account of Göpfert–Mack in § 4, the reservation about the perimeter term	Primary (text)	S. Chatterjee, "Yang–Mills for probabilists", arXiv:1803.01950 (2018; PDF text § 3– § 5)
$a\sqrt{\sigma}$, $am(0^{++})$ and lattice sizes for $SU(2)$, the extrapolation formula (6), the continuum value 3.844(61), $aL\sqrt{\sigma} \geq 3$	Primary (text)	B. Lucini, M. Teper, JHEP 0106 (2001) 050, arXiv:hep-lat/0103027 (Table 1, formulas (4)(5), Table 5)
ξ and $L\,a\sqrt{\sigma}$ of Table 1, Tables 2–4, the numbers of Figures 1 and 2	Computed	python3 (weighted least squares, the confidence level of χ^2, the two-loop formula (4)). Scripts and output are kept in the record of this book's investigation

Item	Kind	Source / tools
'uniform in the volume of space-time'	Primary (text § 6.5)	A. Jaffe, E. Witten, "Quantum Yang–Mills theory" (the Clay problem statement) § 6.5
The prescription of holding the string tension fixed	Abstract	M. Creutz, Phys. Rev. D 21 (1980) 2308
The definition $F(r_0)r_0^2 = 1.65$ of r_0	Abstract	R. Sommer, Nucl. Phys. B 411 (1994) 839, arXiv:hep-lat/9310022
$r_0\sqrt{\sigma}$ for $SU(3)$, $\sqrt{\sigma} = 485(6)$ MeV, the authors' caution	Primary (text checked against the checked material)	A. Athenodorou, M. Teper, JHEP 11 (2020) 172, arXiv:2007.06422 (formulas (8)(9), Table 17, § 5.2)
$m(0^{++})/\sqrt{\sigma} = 3.781(23)$ for $SU(2)$	Primary (text checked against the checked material)	A. Athenodorou, M. Teper, JHEP 12 (2021) 082, arXiv:2106.00364 (Table 23)
The area law and continuum limit of three-dimensional compact $U(1)$	Abstract (continuum limit, Villain action; the INSPIRE abstract). Text not read	M. Göpfert, G. Mack, Commun. Math. Phys. 82 (1982) 545–606 (the year is the Crossref year of publication; INSPIRE gives 1981. The area law also checked in the text of Chatterjee 2018 § 4)
Scaling limits of four-dimensional Ising-type models and φ^4 are Gaussian	Abstract	M. Aizenman, H. Duminil-Copin, Ann. Math. 194 (2021) 163, arXiv:1912.07973
The area law and the infinite-volume limit at small β	Abstract (existence and analyticity of the infinite-volume limit, Wilson's confinement bound; the INSPIRE abstract). The area law also checked in the account of Chatterjee § 4	K. Osterwalder, E. Seiler, Ann. Phys. 110 (1978) 440
Uniqueness and exponential decay at strong coupling for $SU(N)$ ($\beta_{\mathrm{SZZ}} < 1/48$ in four dimensions)	Abstract	H. Shen, R. Zhu, X. Zhu, arXiv:2204.12737
Exponential decay of correlations at weak coupling for finite groups	Abstract	A. Adhikari, S. Cao, arXiv:2202.10375 (Ann. Probab. 53 (2025))

Next chapter to read: 5-10 Lattice gauge theory with finite groups (an example where correlations decrease at large β, yet there is no progress in the direction of the continuum limit).

Lattice gauge theory with finite groups — exponential decay at weak coupling is a theorem, but it lies on the side opposite the continuum limit

> This chapter organises established knowledge (sources carry the “Known” label in the text). This book's own investigation is mainly in Part 15 and Volume 3 (Parts 16–28 and the appendices).

What this chapter shows — If elements of a finite group are put on the edges, at weak coupling the edges freeze, and only defects that pay the gap Δ_G of the action remain. From this come the first-order formula for Wilson loops (Chatterjee, Cao) and the theorem of exponential decay of correlations (Adhikari–Cao). That decay is about the side where the correlation length goes to 0, that is, the side opposite the continuum limit.

Prerequisite chapters — 5-02. Tools: 5-01, 1-08, 4-11. § 07 reads quickly after 5-09.

Said first — All the theorems are about theories with finite groups at fixed lattice spacing. This chapter shows nothing about the mass gap or the continuum limit for continuous groups. "No confinement at weak coupling" is a physical view, not a claim of the theorems cited (§ 08).

Order of this chapter

1. Making the group finite — only the edge values become discrete
2. Frozen edges — the smallest defect and Δ_G
3. The shape of defects is set by the dimension — area law and perimeter law
4. The line breaks on the way — duality and a first-order transition (Table 1)
5. Fixed by the perimeter alone — Wilson loops (Figure 1, Table 2)
6. Exponential decay of correlations — Adhikari–Cao (Figure 2, Table 3) / Detour: non-abelian defects become links
7. The side opposite the continuum limit — the correlation length goes to 0
8. Three "gaps" — a gap, but no confinement (Table 4)
9. Even approaching through subgroups of SU(2) — $\Delta_G \to 0$

01

Making the group finite — the same lattice, with only the edge values discrete

Replacing the group on the edges by a finite group G leaves the definitions of the lattice, plaquettes, gauge transformations and Wilson loops unchanged (5-01 § 02). The integral over each edge becomes an average over $|G|$ elements. The action uses the character $\chi = \operatorname{tr}\rho$ of a faithful unitary representation ρ (of dimension d; $\rho(g) \neq I$ if $g \neq 1$), and we write σ_p for the product once around a plaquette p.

$$S_\Lambda(\sigma) = \sum_p \operatorname{Re}\big(\chi(1) - \chi(\sigma_p)\big), \qquad \mu_{\Lambda,\beta}(\sigma) = \frac{1}{Z_{\Lambda,\beta}}\, e^{-\beta S_\Lambda(\sigma)} \quad (1)$$

(Adhikari–Cao (1.2)(1.3)) Known . $\chi(g)$ is a sum of d eigenvalues of absolute value 1, so each term is 0 or more, and it is 0 only when $\sigma_p = 1$.

Example. For $\mathbb{Z}_2 = \{\pm 1\}$ the weight is proportional to $e^{\beta \sum_p \sigma_p}$: Ising lattice gauge theory, which Chatterjee (2020) calls the first lattice gauge theory, defined by Wegner (1971) Known (2-09). $\mathbb{Z}_N$ is $U(1)$ with angles restricted to multiples of $2\pi/N$. For the finite subgroups of $SU(2)$, $Q_8 = \{\pm 1, \pm i, \pm j, \pm k\}$ (the unit quaternions) and $\bar{T}$, $\bar{O}$, $\bar{I}$ (the rotation groups of the regular tetrahedron, octahedron and icosahedron, subgroups of $SO(3)$, pulled back to $SU(2)$; of orders 24, 48, 120; 1-12 § 08), $\chi = \operatorname{tr}$, and relative to Wilson's convention $\beta_W = 2\beta$.

A finite group has no Lie algebra, and no elements close to the identity. There is no classical continuum limit starting from the $U = e^{iaA}$ of 5-01 § 05; we read it as a lattice model with the same gauge structure.

02

Frozen edges — the smallest defect, and the gap Δ_G of the action

When β is large, the weight strongly favours configurations in which every plaquette is the identity. The only configurations on a box with this property are pure gauge ($\sigma_{(x,y)} = h_x h_y^{-1}$), and up to gauge there is just one, "every edge the identity". The weak-coupling theory is small fluctuations around it, and we say that "the edges freeze".

From here on we write D for the dimension of the lattice, and keep d for the dimension of the representation. The cheapest fluctuation is to change a single edge to $g \neq 1$. On a D-dimensional lattice that edge lies in $2(D-1)$ plaquettes, and each of their products is conjugate to $g^{\pm 1}$. Since $\chi(g^{-1}) = \chi(g)$,

$$\Delta S = 2(D-1)\,\mathrm{Re}\big(\chi(1)-\chi(g)\big) \;\geq\; 2(D-1)\,\Delta_G, \qquad \Delta_G := \min_{g\neq 1} \mathrm{Re}\big(\chi(1)-\chi(g)\big) \tag{2}$$

$\Delta_G > 0$ is equivalent to ρ being faithful (Cao § 1.2) Known . Example: for four-dimensional $\mathbb{Z}_2$, $\Delta_G = 2$, and the weight of this defect is $e^{-12\beta}$. This 12 appears in the theorem of § 05.

For $U(1)$, changing an edge to $e^{i\theta}$ costs $\approx (D-1)\theta^2$, which can be as small as one likes, so long waves (photons) are cheap to make (5-06 § 03). Δ_G is a different thing from the mass gap of 4-11.

03

The shape of defects is set by the dimension — points, loops, closed surfaces, and the area and perimeter laws

For $\mathbb{Z}_2$, taking a surface (a set of plaquettes) P with boundary γ, the inner edges appear twice each and cancel (a discrete Stokes; the existence of P is Lemma 3.1 of Chatterjee 2020) Known .

$$W_\gamma = \prod_{e\in\gamma} \sigma_e = \prod_{p\in P} \sigma_p = (-1)^{\#\{p\in P:\, \sigma_p = -1\}} \tag{3}$$

Here we use the dual lattice. It is the lattice whose vertices are the centres of the D-dimensional cubes, and the k-dimensional cells of the original lattice (edges, faces, cubes, ...) correspond one to one with the $D-k$-dimensional cells of the dual lattice. In Lemma 3.3 of the same paper, the negative plaquettes split into "vortices", and the dual of each vortex is closed. The dual of a plaquette ($k=2$) is a $(D-2)$-dimensional cell, so vortices are points in two dimensions, closed loops in three, and closed surfaces in four Known . Only vortices linked with γ change the sign (Lemma 3.2), and if their number N is Poisson-distributed with mean μ then $\mathbb{E}[(-1)^N] = e^{-2\mu}$ (1-08, 17-02 § 06).

Example. In two dimensions with free boundary the plaquettes are independent and $\mathbb{E}\sigma_p = \tanh\beta$, so for a loop of area A

$$\langle W_\gamma \rangle = (\tanh\beta)^A, \qquad \sigma_{\text{lat}} = -\ln\tanh\beta > 0 \quad (\text{for all } \beta) \tag{4}$$

Known . In two dimensions even finite groups have the area law at every β (for general groups, 7-02 § 03). In three and four dimensions at large β the vortices are small, only nearby ones link with γ, and their number is proportional to the perimeter (the perimeter law) Physics . The area law at small β is a theorem of the strong-coupling expansion (5-02 § 03) Known .

04

The line breaks on the way — duality and the first-order transition of four-dimensional $\mathbb{Z}_2$

Theorem 4.1 of Chatterjee 2020 writes the partition function of four-dimensional $\mathbb{Z}_2$ as the same theory on the dual lattice, moving the coupling to a λ that exchanges strong and weak Known .

$$\lambda(\beta) = -\tfrac{1}{2}\log\tanh\beta, \qquad \lambda(\beta_{\mathrm{sd}}) = \beta_{\mathrm{sd}} \iff \beta_{\mathrm{sd}} = \tfrac{1}{2}\ln\bigl(1+\sqrt{2}\bigr) = 0.44069 \quad (5)$$

The number in (5) is Computed . Duality maps transition points to transition points, so if there is exactly one transition it falls at the self-dual point Known . According to the abstract of Wegner (1971), the dual of the Ising model on the simple cubic lattice is a model of four-spin products (plaquettes), that is, three-dimensional $\mathbb{Z}_2$ gauge theory Known .

A first-order transition is one in which a quantity such as the mean plaquette jumps at one point of β. Disagreement (hysteresis) between runs raising β and runs lowering it is a sign of it. In four dimensions Creutz–Jacobs–Rebbi (1979) reported a first-order transition numerically, and reported that for $\mathbb{Z}_N$ with $N \geq 5$ there are two transitions that are not first order, one of which moves to $\beta \to \infty$ with N Physics .

On this machine too we ran the heat-bath method on a periodic 8^4 lattice (the run raising β from a random start, the run lowering it from all edges $+1$, averaging the second half of 200 sweeps at each β). The leading term $1 - 8e^{-12\beta}$ comes from the probability $\approx 4e^{-12\beta}$ that a defect of § 02 sits on one of the four edges.

Computed Table 1 The mean plaquette of four-dimensional $\mathbb{Z}_2$ on 8^4 (excerpt)

β	Run raising β	Run lowering β	$\tanh\beta$	$1 - 8e^{-12\beta}$
0.40	0.4214	0.4205	0.3799	0.9342
0.44	0.5025	0.9117	0.4136	0.9593
0.46	0.9206	0.9422	0.4301	0.9680
0.60	0.9928	0.9930	0.5370	0.9940

For $\beta = 0.43$–0.45 the two runs disagree (by 0.41 at 0.44), and that interval straddles the self-dual point Computed . These are short runs on a small lattice, with no power to decide the order of the transition (first order, or a transition without a jump) (5-07; the same comparison for $U(1)$ and $SU(2)$ is in 5-06 § 04).

05

Fixed by the perimeter alone — the first-order formula for Wilson loops at weak coupling

Theorem 1.1 of Chatterjee (2020): let ℓ be the number of edges of γ, and ℓ_0 the number of corner edges (edges lying in the same plaquette as another edge of γ; 8 for a rectangle). Then for β_0 and above, with universal constants C_1, C_2, Known

$$\left|\langle W_\gamma\rangle_\beta - e^{-2\ell e^{-12\beta}}\right| \le C_1\left(e^{-2\beta} + \sqrt{\ell_0/\ell}\right)^{C_2} \quad (6)$$

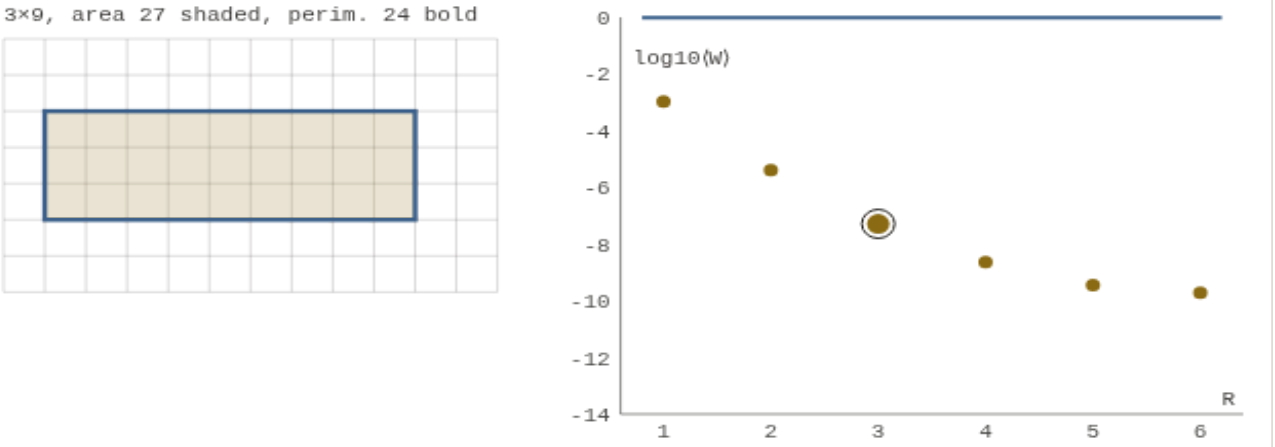
The paper reads it as "Nontrivial behavior happens if and only if ℓ is like a constant multiple of $e^{12\beta}$". Each edge becomes the centre of a defect of § 02 with probability $\approx e^{-12\beta}$ and changes the sign once; if their number is Poisson, this is the leading term of (6) (§ 03) Physics .

Cao (2020) extended this to any finite group. Its Theorem 1.2.1 is about a single self-avoiding loop γ, with conditions $\beta \ge (1000 + 14\log|G|)/\Delta_G$ and $\|A_\beta\|_{\mathrm{op}} < 1$. Under them the leading term is

$$\langle W_\gamma\rangle_\beta \approx \sum_{i=1}^{d} e^{-\ell\, r_\beta(1-\lambda_i(\beta))}, \qquad r_\beta = \sum_{g\neq 1} e^{-6\beta\,\mathrm{Re}(\chi(1)-\chi(g))} \quad (7)$$

where λ_i are the eigenvalues of the matrix A_β, the average of the $\rho(g)$ carried by defects weighted by the defect weights, and $\|A_\beta\|_{\mathrm{op}}$ is its operator norm (the largest factor by which it stretches a unit vector) Known . Because of the latter condition, $\mathbb{Z}_2$ ($A_\beta = -1$, norm 1) does not fall directly under this theorem (Cao's Remark 1.2.3) Known . Even so, putting $r_\beta = e^{-12\beta}$ and $\lambda = -1$ for $\mathbb{Z}_2$ makes the form of the leading term agree with (6) Computed . The leading term is a function of the perimeter alone: for the same perimeter, long and thin or square, it has the same value. In two dimensions, (4) was a function of the area alone.

Figure 1 Rectangles $R \times (12 - R)$ of perimeter 24. β, R and "Play" change $\log_{10}\langle W\rangle$ for the two-dimensional (4) (points) and the four-dimensional leading term (6) (horizontal line)



Computed The numbers in this figure were computed on this machine ((4)(6) computed on the spot). Example: at $\beta = 0.7$, for 1×11, 3×9, 6×6 the two-dimensional values are 3.9×10^{-3}, 1.2×10^{-6}, 1.3×10^{-8}, and the four-dimensional value is 0.989 for all. Theorem 1.1 guarantees only sufficiently large β (β_0 and above; the value is not stated explicitly in the paper). 0.4407 is a guide to the transition, and below it (dashed) the leading term does not fit even in form.

Computed Table 2 Wilson loops of four-dimensional $\mathbb{Z}_2$ on 8^4 (3000 sweeps, jackknife)

β	$R \times T$	$\langle W \rangle$	$-\ln W$ / area	$-\ln W$ / perimeter
0.40	1×1	0.42137(39)	0.864	0.216
0.40	2×3	0.00660(18)	0.837	0.502
0.50	1×1	0.97190(20)	0.0285	0.00713
0.50	3×3	0.91281(59)	0.0101	0.00760

At $\beta = 0.40$ the values divided by the area line up, and at $\beta = 0.50$ those divided by the perimeter Computed . The latter are about 1.5 times the $2e^{-6} = 0.00496$ of (6); just above the transition, corrections from corners and large vortices matter.

06

Exponential decay of correlations — the theorem of Adhikari–Cao and Δ_G

That separated observables become nearly independent (decay of correlations) had earlier results for abelian finite groups at weak coupling ($\mathbb{Z}_2$ in Chatterjee 2020 § 5, and the results of Seiler and Forsström listed in Adhikari–Cao § 1.1), and Adhikari–Cao (Ann. Probab. 2025) showed it in four dimensions including the non-abelian case. Let B_1, B_2 be rectangles separated by at least L in the ℓ^∞ distance (the largest difference of coordinates), and $|B|$ the number of plaquettes contained in a rectangle B. f_i is a function of the holonomies of loops in B_i (the product σ_γ of the edge elements along the loop; 5-01 § 02) that is conjugation-invariant in each argument ($f(h$

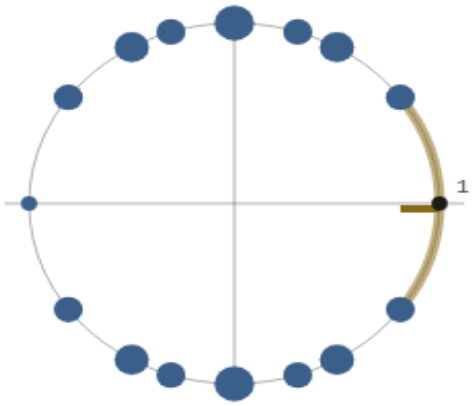
$^{-1}gh) = f(g)$). This includes any function of Wilson loops. $\|f\|_\infty = \max|f|$, and Cov is the covariance (1-08). If $\beta \geq (114 + 4\log|G|)/\Delta G$, then in every finite box Known

$$|\mathrm{Cov}(f_1, f_2)| \;\leq\; 4\big(4\cdot 10^{24}|G|^2\big)^{|B_1|+|B_2|}\,\|f_1\|_\infty\|f_2\|_\infty\, e^{-(\beta/2)\Delta G\,(L-1)} \quad (8)$$

It is an estimate independent of the box. The constant grows exponentially in the size of the supports of the observables (B_1, B_2), but does not depend on L. The authors write that it "can be interpreted as showing the existence of a mass gap in the setting of finite (non-Abelian) gauge groups at large β" (for small β, 5-04 § 07). The proof uses the fact that a map exchanging two independent configurations on a region preserves the measure, and bounds the covariance by "the probability that defects connect B_1 to B_2".

The condition and the rate are set by ΔG $(1 - \cos(2\pi/N)$ for $\mathbb{Z}_N$, $2 - 2\cos\theta_{\min}$ for subgroups of $SU(2)$).

Figure 2 The eigenvalues of the group elements placed on the unit circle. Choosing a group with the buttons changes the angle $\theta_{\min}$ closest to 1 (arc), $1 - \cos\theta_{\min}$ (thick line), and the condition of the theorem



Ī Δ_G = 0.381966

Ī (order 120) : θ_min = 36°、Δ_G = 0.3820, Adhikari–Cao condition β ≥ 348.59 (β_W ≥ 697.18)

Computed The numbers in this figure were computed on this machine (the groups were built by closing under products, and the eigenvalue angles and ΔG counted; the same as Table 3).

Computed Table 3 ΔG and the condition on β in the theorem ($\chi = \mathrm{tr}$ for subgroups of $SU(2)$)

Group	Order	$\theta_{\min}$	ΔG	Adhikari–Cao	Same in β_W
$\mathbb{Z}_2$	2	180°	2	58.39	—
$\bar{T}$	24	60°	1	126.71	253.42
$\bar{O}$	48	45°	0.5858	221.04	442.09
$\bar{I}$	120	36°	0.3820	348.59	697.18

The finer the group, the smaller Δ_G, and the condition grows as $1/\Delta_G$ ($\approx 2\pi^2/N^2$ for $\mathbb{Z}_N$). Cao 2020 calls its own condition (7) "very loose" (the threshold of Adhikari–Cao is about an order of magnitude smaller).

Detour: non-abelian defects become links

A subsection you may skip.

According to Adhikari–Cao § 1.2, in the abelian case separated defects appear independently, but in the non-abelian case two linked defect loops do not decompose however far apart they are Known . This is because the element a defect carries changes as $g \mapsto hgh^{-1}$ when it goes around another defect (in Q_8, $jij^{-1} = -i$ Computed). The paper writes of the non-abelian case that it does "not (as of yet) admit a cluster expansion".

07

The side opposite the continuum limit — the correlation length goes to 0

The constant in (8) depends on the size of the supports of the observables but not on L, so (8) gives an upper bound on the correlation length (4-11 § 07).

$$\xi(\beta) \;\le\; \frac{2}{\beta\,\Delta_G} \xrightarrow{\beta\to\infty} 0 \quad (9)$$

Known (a consequence of (8), with the definition of 4-11 § 07)

At β exactly at the condition, the upper bound is 0.0171 lattice spacings for $\mathbb{Z}_2$ and 0.0150 for $\bar{I}$. The perimeter $1/\big(r\beta(1-\lambda)\big)$ at which the leading term of (7) moves away from 1 is, at β exactly at Cao's condition, $10^{2721.1}$ for $\bar{T}$ and $10^{2780.1}$ for $\bar{I}$ Computed . The conditions are loose, so what matters is the direction: a finite group at weak coupling is a frozen lattice whose correlations do not reach even one lattice spacing.

The lattice correlation-length problem of 5-09 § 05 (Problem 5.1 of Chatterjee 2018, arXiv:1803.01950) asks, for four-dimensional non-abelian compact groups, for $\xi(\beta) < \infty$ at each β, and $\xi \to \infty$ as $\beta \to \infty$. Finite groups satisfy the first part at large β, and go to $\xi \to 0$, the exact opposite of the second Known . This is "Type 1" of 5-09 § 08, and the physical mass is infinite.

08

The precise form of "a gap, but no confinement" — three "gaps"

Theories with finite groups are cited as examples of "a gap, but no confinement". This "gap" has three meanings.

Known Table 4 Three meanings of "gap", and the state of finite groups at weak coupling

Meaning	Content	Finite group at weak coupling
(a) The gap of the problem statement	The spectral gap of a continuum, infinite-volume theory (4-11)	There is no continuum theory, so it does not apply
(b) The strong form	Exponential decay under any boundary condition (the definition of strong decay [Definition 2.3 of Chatterjee 2021])	Not shown (the Theorem from Uniform Decay to the Area Law [Theorem 2.4 of Chatterjee 2021] assumes a connected group, and finite groups do not fall under it literally)
(c) The weak form	Decay of covariances in a given measure (the form of (8))	Theorem (Adhikari–Cao)

The Theorem from Uniform Decay to the Area Law of Chatterjee (2021) takes the group G to be a closed connected subgroup of $U(n)$ and derives the area law from (b) for representations acting non-trivially on the centre Known . Non-trivial finite groups are not connected, so they do not literally fall under this assumption. Even so, p.8 of the same paper cites theories with finite groups as examples “believed to be ‘gapped’ … not confining at large β”, and explains why this does not contradict the Theorem from Uniform Decay to the Area Law: the decay there is in a weaker sense than the definition of strong decay (truncated two-point correlations under certain boundary conditions) Known . The paper goes on to say that for some theories it has been rigorously shown that there are several Gibbs measures. Read with the proviso "if the proof goes through without connectedness", (b) does not hold for non-confining finite groups.

The errors of (6)(7) do not go to 0 as the loop grows, so that the string tension is 0 at fixed β is not a claim of the theorems cited here (we have not checked, in this chapter, a result that states the perimeter law directly for abelian finite groups). It is the standard physical view (consistent with Table 2) Physics . Confinement and the gap are different statements (17-02 § 02).

09

Even approaching through subgroups of SU(2) — $\Delta_G \to 0$ and uniformity

According to the abstract of Petcher–Weingarten (1980), $\bar{T}$, $\bar{O}$ and $\bar{I}$ all have first-order transitions, and as the order is raised the critical coupling goes to 0. $\bar{I}$ agreed with the numbers for $SU(2)$ over a wide range of couplings Physics .

We make a rough guide. For $SU(2)$ at $\beta_W = 2.4$ in 5-02, $\langle 2 - \operatorname{tr} U_p \rangle = 0.739 > \Delta_{\bar{I}} = 0.382$. From the leading weak-coupling term $\langle \frac{1}{2} \operatorname{tr} U_p \rangle \approx 1 - 3/(4\beta_W)$ (5-06 § 04), $\langle 2 - \operatorname{tr} U_p \rangle \approx 3/(2\beta_W)$ equals Δ_G at

$$\beta_W \approx \frac{3}{2\Delta_G} : \qquad \bar{T}\ 1.50, \qquad \bar{O}\ 2.56, \qquad \bar{I}\ 3.93 \ (10)$$

Computed . It is the crude estimate that for β_W larger than this the fluctuations of $SU(2)$ are finer than the steps of the finite group, not the value of a transition point Physics .

The point is the direction of the constants. Bringing the subgroups closer to $SU(2)$ gives $\Delta_G \to 0$, and both the condition $\propto 1/\Delta_G$ and the rate $(\beta/2)\Delta_G$ degenerate. What is needed for $SU(2)$ is that the correlation length be finite in a regime where the action has no gap, and moreover diverge as $\beta \to \infty$. That mass gap is thought to come from the coupling growing stronger with distance (5-06 § 07, 4-07) Physics .

What this chapter can say

Matter	Said	Not said
(1)–(5)	Definitions, the smallest defect, vortices, the two-dimensional area law, duality Known . Tables 1 and 3 Computed	The vortex-gas reading (an estimate). The order of the four-dimensional transition
(6)(7)	Theorems on the leading term Known . Agreement of the form of the leading term for $\mathbb{Z}_2$, Table 2 Computed	That the string tension is 0 at fixed β
(8)(9)	Exponential decay and $\xi \to 0$ for four-dimensional finite groups at weak coupling Known	The strong form (b). Other dimensions
Relation to $SU(2)$	Reports of numerics for the subgroups Physics . The arithmetic of (10) Computed	The mass gap, the continuum limit and confinement for continuous groups. This chapter shows nothing about them

Sources and reproduction

Item	Kind	Source / tools
(1)(2)(8), quotations, the detour	Text § 1	A. Adhikari, S. Cao, arXiv:2202.10375v3; Ann. Probab. 53(1) (2025), doi:10.1214/24-AOP1702. The earlier results of Seiler and Forsström listed in § 1.1: bibliographic data only

Item	Kind	Source / tools
(3)(5)(6), quotations, $\mathbb{Z}_2$ in § 06	Text § 1– § 5	S. Chatterjee, arXiv:1811.09770v3; Commun. Math. Phys. 377 (2020) 307–340, doi:10.1007/s00220-020-03738-9
(7), the note on the conditions	Text § 1.2 (Theorem 1.2.1, Remarks 1.2.2–1.2.3)	S. Cao, arXiv:2001.05627v3; Commun. Math. Phys. 380 (2020) 1439–1505, doi:10.1007/s00220-020-03912-z
The Theorem from Uniform Decay to the Area Law and the quotation in § 08	Text § 2 (Definitions 2.1 and 2.3, Theorems 2.2 and 2.4, p.8)	S. Chatterjee, Commun. Math. Phys. 385 (2021) 1007–1039, arXiv:2006.16229
Duality and the transition in § 04	Abstract	F. J. Wegner, J. Math. Phys. 12 (1971) 2259; M. Creutz, L. Jacobs, C. Rebbi, Phys. Rev. Lett. 42 (1979) 1390, Phys. Rev. D 20 (1979) 1915
The subgroups in § 09	Abstract	D. Petcher, D. H. Weingarten, Phys. Rev. D 22 (1980) 2465
Tables 1–3, Figures 1 and 2, (5)(7)(10) and the numbers in § 07	Computed on this machine	python3 (calculations on this machine: group tables, heat bath for four-dimensional $\mathbb{Z}_2$ (about 100 seconds), comparison of area and perimeter)

Next chapter to read: 5-11 Detour: Bakry–Émery and one-link integration (seeing, on a single link, how theorems come out on the strong-coupling side).

Detour: Bakry–Émery and one-link integration — a criterion that gets a gap from curvature, and an example of integrating first to move the constant

> This chapter is this book's own investigation. § 02 and § 03 are established knowledge.

What this chapter shows — The Bakry–Émery criterion, "positive curvature gives a Poincaré inequality", and how conservative it is on a single link. On lattice $SU(2)$, integrating first a family of links containing exactly one link of every plaquette lowers the bending of the action to $O(\beta^2)$ and moves the threshold from $1/12$ to $\sqrt{2/27} \approx 0.272$ (a paper on this site), and the breakdown of the constant 27.

Prerequisite chapters — 5-02, 1-10. We also use 5-04 § 08 and 1-08 § 03 (the Haar measure of $SU(2)$).

Said first — This is about the side of small β at fixed lattice spacing, the shore opposite the continuum limit, and shows nothing about the continuum limit or the mass gap of the problem statement. The Lean theorems go as far as the lower bound on the bending and the combinatorics of the family; the mass gap is not a Lean theorem. Nor do we claim to have widened the strong-coupling range (§ 07).

Order of this chapter

1. Where this chapter stands — an example on the strong-coupling side
2. Curvature makes a gap — the criterion and its proof
3. Seen on a single link — the true gap grows with κ (Figure 1, Table 1)
4. Applying it to the whole lattice — 1/12 and 1/8
5. Integrating first — complete families (Table 2)
6. The breakdown of 27 — $36 \times 6 \div 8$
7. Comparing thresholds — a line and a parabola (Figure 2, Table 3)
8. Detour: complete families stop at four dimensions — equality in the Hamming bound
9. Limits and position of the method — what is Lean

The β of this chapter is in the $SU(2)$ convention $\beta = 4/g^2$, inversely proportional to the square of the coupling g (5-01). It plays the role of the inverse temperature of statistical mechanics (5-02).

$$S(U) = \beta \sum_P \Big(1 - \tfrac{1}{2}\operatorname{tr} U_P\Big), \qquad \mu_\beta(dU) = \frac{1}{Z}\, e^{-S(U)} \prod_\ell dU_\ell \tag{1}$$

Elements of $SU(2)$ are quaternions of length 1, and the group is the unit sphere S^3 (1-02). The inner product is $\langle u, v\rangle = \frac{1}{2}\operatorname{tr}(uv^\dagger)$.

01

Where this chapter stands — an example of a uniform estimate on the strong-coupling side

The key to the Langevin road of 5-04 § 08 is the **Bakry–Émery criterion**, which says "if the roundness (curvature) of the group beats the bumps of the action, a Poincaré inequality holds with a box-independent constant". This chapter looks at it concretely on a single link (§ 02, § 03) and on the whole lattice (§ 04– § 07).

What one gets is decay of correlations in lattice units at fixed lattice spacing (4-11), the shore opposite what is wanted in the continuum limit, "the gap in lattice units closing at the right rate" (5-09).

In 5-04 § 08 we used only the conclusion of the criterion. This chapter adds three things: the line of proof of the criterion (§ 02), a comparison with the true gap on a single link (§ 03), and that integrating part of the links first lowers the bending from first to second order in β and moves the constant of the threshold (§ 05– § 07).

The shortest way to read is to see first, in Figure 1, "where the criterion becomes conservative", then the threshold of the line $2 - c\beta$ in § 04, and finally the threshold of the parabola $2 - 27\beta^2$ in § 05– § 07. The breakdown in § 06 and the detour in § 08 can be skipped without losing the thread.

02

Curvature makes a gap — the criterion and a one-line proof

In 1-10 § 08 we saw, for $e^{-V}\,dx/Z$ on the line, "if $V'' \geq \kappa > 0$ then $C_P \leq 1/\kappa$". On a group the roundness of the space itself is added to this.

As a metaphor, two people who set off due north from the equator meet at the North Pole. Precisely, the Ricci curvature $\mathrm{Ric}(v, v)$ is the sum, over each direction

perpendicular to v, of how much a bundle of geodesics leaving in the direction of the unit vector v draws together; on the n-dimensional sphere of radius 1 it is $(n-1)|v|^2$ Known . Geodesics on the group are $\tau \mapsto U e^{\tau X}$ (X is a pure imaginary quaternion, an element of the Lie algebra; 1-03), and we write $\operatorname{Hess} V(X,X)$ for the second derivative along them. It corresponds to V'' on the line.

$$\operatorname{Ric}(X,X) = 2|X|^2 \ (S^3), \qquad \operatorname{Ric} + \operatorname{Hess} V \geq K > 0 \implies \operatorname{Var}_\mu(f) \leq \frac{1}{K}\int |\nabla f|^2 d\mu \quad (2)$$

The right-hand side is the Poincaré version of the criterion of Bakry–Émery (1985) ($\mu = e^{-V}\operatorname{vol}/Z$, with the condition at every point) Known . A logarithmic Sobolev inequality also follows. **No dimension enters the constant**, so if the lower bound on curvature does not depend on the box, neither does the Poincaré constant.

The line of proof is the same as in 1-10 § 08. With $Lf = -\Delta f + \nabla V \cdot \nabla f$ (Δ is the Laplacian on the group, different from the Δ of the mass gap) and $Lf = \lambda f$, integrating by parts twice (the integrated Bochner formula) gives Known

$$\lambda^2 \int f^2 d\mu = \int (Lf)^2 d\mu = \int \|\nabla^2 f\|^2 d\mu + \int (\operatorname{Ric} + \operatorname{Hess} V)(\nabla f, \nabla f)\, d\mu \ \geq \ K\lambda \int f^2 d\mu \quad (3)$$

($\int |\nabla f|^2 d\mu = \lambda \int f^2 d\mu$), so $\lambda \geq K$. (3) is the multidimensional form of $\int |g'|^2 + \int V'' g^2 = \lambda \int g^2$ in 1-10: V'' is replaced by $\operatorname{Hess} V$, and the roundness of the space is added as Ric. Example: on uniform S^3 the criterion says "gap ≥ 2", while the truth is 3 (first-order spherical harmonics; the first row of Table 1). It is conservative to the extent that it does not use the dimension.

For how this book uses this, see 24-06 § 02 (the convexity ceiling on a closed group is in 24-06 § 03).

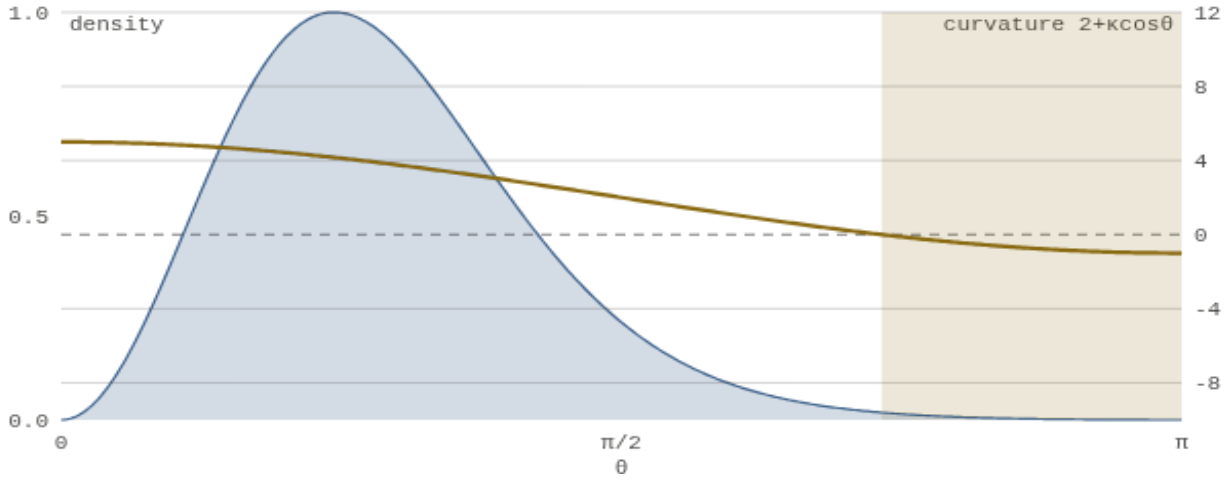
03 Seen on a single link — the true gap and the criterion

Holding fixed everything but a single link u, the weight comes from the six plaquettes containing u. The product of the remaining three links of a plaquette without u (a U shape) is called a **staple**. Using the sum $M = \sum_i V_i$ of the six staples $V_1, \dots, V_6$ with orientations aligned, the weight is $e^{\beta \langle u, M\rangle}$ (the same set-up as 23-02 Roads tried on the four-dimensional lattice, and walls; the distribution of the heat-bath method; 5-07). Taking the direction of M as the North Pole and θ as the angle from the North Pole, the Hessian of the height $\cos\theta$ is $-\cos\theta$ times the metric, so

$$\mu_\kappa(du) \propto e^{\kappa \cos\theta} du \ \ (\kappa = \beta|M| \leq 6\beta), \qquad \operatorname{Ric} + \operatorname{Hess} V = (2 + \kappa\cos\theta)|X|^2 \geq (2-\kappa)|X|^2 \ ($$

Known . All the criterion can say is a gap $\geq 2-\kappa$ for $\kappa<2$, but the true gap **grows** with κ (the measure concentrated near the North Pole is close to a Gaussian of variance $1/\kappa$). The criterion is tripped up by the negative curvature near the South Pole, where the measure carries almost no weight.

Figure 1 The knob and "Play" change κ. The fill is the density $e^{\kappa\cos\theta}\sin^2\theta$, the solid line the curvature $2+\kappa\cos\theta$ (right-hand scale), and the band the range where the curvature is negative.



κ=3.00, criterion 0.00, gap 4.1160, weight of negative curvature 0.0039

Computed The numbers in this figure were computed on this machine (the gaps are 41 points of the calculation of Table 1, the rest from formulas on the spot). If JavaScript is disabled, Table 1 holds the content.

Table 1 The gap λ_1 of μ_κ and the criterion Computed (eigenvalue problems for each angular momentum $l=0,1,2$ on S^2; steps 800 and 1600 agree to 5 digits. The rightmost column is the numerical integral $\int_{\arccos(-2/\kappa)}^{\pi} e^{\kappa\cos\theta}\sin^2\theta\,d\theta \,/\, \int_0^\pi$)

κ	$l=1$ = gap	Criterion	Weight of the region of negative curvature
0	3.0000	2.0	0
1	3.1323	1.0	0
2	3.5164	0	0
4	4.8782	–	0.0027
6	6.6746	–	0.00059
10	10.580	–	1.5×10^{-5}

The gap is set by $l=1$ (fluctuations tilting perpendicular to M). The criterion is a correct but crude tool.

Note. $\lambda_1 = 1/C_P$ is the gap in Langevin time, different from the gap of the transfer matrix (5-03) (18-03).

```
04
```

Applying it to the whole lattice — thresholds 1/12 and 1/8

The configuration space of the whole lattice is a product of S^3, and the Ricci curvature stays 2 on each link. The criterion asks for the following.

$$K(\beta) = 2 + \inf_U \lambda_{\min}\big(\operatorname{Hess} S(U)\big) > 0 \quad (5)$$

The lower bound has the form $-c\beta$. Shen–Zhu–Zhu (2023), with a plaquette-by-plaquette estimate, showed for $SU(N)$ under $|\beta_{\mathrm{SZZ}}| < 1/(16(d-1))$ uniqueness of the infinite-volume measure and exponential decay of correlations (abstract: "the infinite volume measure has a strictly positive mass gap") Known . The convention for $SU(2)$ is $\beta = 4\beta_{\mathrm{SZZ}}$, so in four dimensions this is $\beta < 1/12 = 0.0833$, and in the notation of this chapter $K = 2 - 24\beta$ ($\beta < 1/12$ is the same as in 5-04 Table 4).

On the other hand, the $\mathbb{Z}_2$ configuration with every plaquette -1 gives $\lambda_{\min}(\operatorname{Hess} S) = -16\beta$ (elementary) Paper . The numerical search of the paper found nothing below it, but this is not a proof (paper § 1, § 8). The ceiling $1/8$ comes from the elementary side alone: for this configuration $K \le 2 - 16\beta$, so **as long as the criterion is used without integrating, it tops out at $\beta < 1/8$.**

05

Integrating first — complete families and one-link integration

The paper on this site integrates part of the links exactly, as equalities (rather than bounding them by inequalities), before applying the criterion. The set I^* to be integrated is required to satisfy the following.

$$I^* \text{ is a complete family} \iff \text{for every plaquette } P,\ |P \cap I^*| = 1 \quad (6)$$

No two links of the family lie on the same plaquette, so each link can be integrated independently; and since every plaquette contains a link of the family, **not a single bare plaquette remains in the action after integration**. The density of the family is $1/4$, and four dimensions has a family of period 2 (counting and figures in 23-02 Roads tried on the four-dimensional lattice, and walls; why only up to four dimensions in § 08).

The integral over one link comes out in Catalan numbers. Under the Haar measure $\operatorname{tr} u = 2\cos\theta$ has the semicircle distribution (1-08 (4)), with even moments $\mathbb{E}[(2\cos\theta)^{2k}] = \mathrm{Cat}_k$ (the detour in 1-08). Expanding the exponential and using $\mathrm{Cat}_k/(2k)! = 1/(k!(k+1)!)$ gives

$$\int_{SU(2)} e^{\beta\langle u,M\rangle}\,du = \sum_{k\ge 0}\frac{\kappa^{2k}\,\mathrm{Cat}_k}{(2k)!\,4^k} = F_0\!\left(\frac{\kappa^2}{4}\right) = \frac{2I_1(\kappa)}{\kappa}, \qquad F_0(x)=\sum_{k\ge 0}\frac{x^k}{k!\,(k+1)!} \tag{7}$$

(I_1 is a modified Bessel function) Known . In the paper this is proved in Lean from the left invariance of the Haar measure alone Lean (`IsHaarS3.one_link`; the only assumptions are the Haar properties). The marginal density of the remaining links is $e^{-\Phi}$, a sum over the links s of the family ("stars").

$$\Phi = -\sum_{s\in I^*} G\big(|M_s|^2\big), \qquad G(t)=\ln F_0\!\left(\frac{\beta^2 t}{4}\right) = \frac{\beta^2 t}{8} - \frac{\beta^4 t^2}{384} + O(\beta^6 t^3) \tag{8}$$

For derivatives we use $F_1(x)=\sum_{k\ge 0} x^k/(k!\,(k+2)!) = F_0'(x)$ and $F_2 = F_1'$. By the chain rule

$$G'(t) = \frac{\beta^2}{4}\,\frac{F_1}{F_0}, \qquad G''(t) = \left(\frac{\beta^2}{4}\right)^2 \frac{F_0F_2 - F_1^2}{F_0^2} \qquad \left(F_i \text{ evaluated at } \frac{\beta^2 t}{4}\right)$$

so $G'\le \beta^2/8$ is the same as $2F_1\le F_0$, and $G''\le 0$ the same as the Turán-type inequality $F_1^2 \ge F_0F_2$.

The $O(\beta)$ term disappears, and the bending starts at $O(\beta^2)$ (loops of length 6 appear with coupling $\beta^2/4$; 23-02 § 02).

Table 2 The series (7) and the Bessel closed form, and the ratio $r = 2F_1/F_0 = G'(t)/(\beta^2/8)$ Computed (also agrees with numerical integration over the angle to 12 digits)

κ	$F_0(\kappa^2/4)$	$2I_1(\kappa)/\kappa$	r
1	1.130318207985	1.130318207985	0.9608
4	4.879732576852	4.879732576852	0.6580
8	99.96828419564	99.96828419564	0.4096

06

The breakdown of 27 — 36 × 6 ÷ 8

The central theorem of the paper is a lower bound on the second derivative of Φ along geodesics. It holds for every β and every configuration, with no smallness assumed Lean (`StaplePerturbation.hess_gauge`. The assumptions are that the period L is even and $L\ge 3$ (that is, an even number 4 or more; at $L=2$ opposite staples share links and the counting changes), $\|U_e\|=1$, and X_e pure imaginary. That the marginal density is $e^{-\Phi}$ is `WilsonOneLink.integral_star_links`).

$$\frac{d^2}{d\tau^2}\Phi\left(Ue^{\tau X}\right)\Big|_0 \geq -27\beta^2\sum_e \|Xe\|^2 \quad (9)$$

÷ 8 — one-link integration	× 36 — one star	× 6 — counting
$0 \leq G' \leq \beta^2/8,\ G'' \leq 0$ $r \leq 1$ in Table 2 ($\iff 2F_1 \leq F_0$). Concavity ($\iff F_1^2 \geq F_0F_2$) from a Catalan-number identity Lean (`turan_series`)	$(\lvert M_s\rvert^2)'' \leq 36\sum \|X\|^2$ Weighted Cauchy–Schwarz and a sum of squares Lean (`StarInequality.star_S` and `StapleCurve.second_deriv_le`; equality is `StapleCurve.second_deriv_sharp`)	Each remaining link lies in exactly 6 stars Lean (`star_count`; $L \geq 3$)

Joining them takes one line. With $t = \lvert M_s\rvert^2$, the second derivative of a star term is

$$-\big(G(t)\big)'' = -G'(t)\,t'' - G''(t)(t')^2 \ \geq\ -G'(t)\max(t'',0) \ \geq\ -\frac{\beta^2}{8}\cdot 36\sum_{\text{in the star}}\|X\|^2 \quad (10$$

and thanks to $G'' \leq 0$ the $(t')^2$ term does not get in the way. Summing over stars counts each link 6 times, giving $-\frac{\beta^2}{8}\cdot 36\cdot 6 = -27\beta^2$. $G' \leq \beta^2/8$ is the claim that "the fluctuations of a tilted Haar measure do not exceed those of the Haar measure", and this is specific to $SU(2)$ (§ 09).

Example. The star ratio $t''/(36\sum\|X\|^2)$ has maximum 0.371 over 200,000 random samples, and equality occurs for the configuration with $s = (+,+,+,-,-,-)$ and the three directions aligned to $s_i\, i$ (1.000000) Computed. The slack lies on the side of joining the stars (§ 09).

07

Comparing thresholds — a line and a parabola

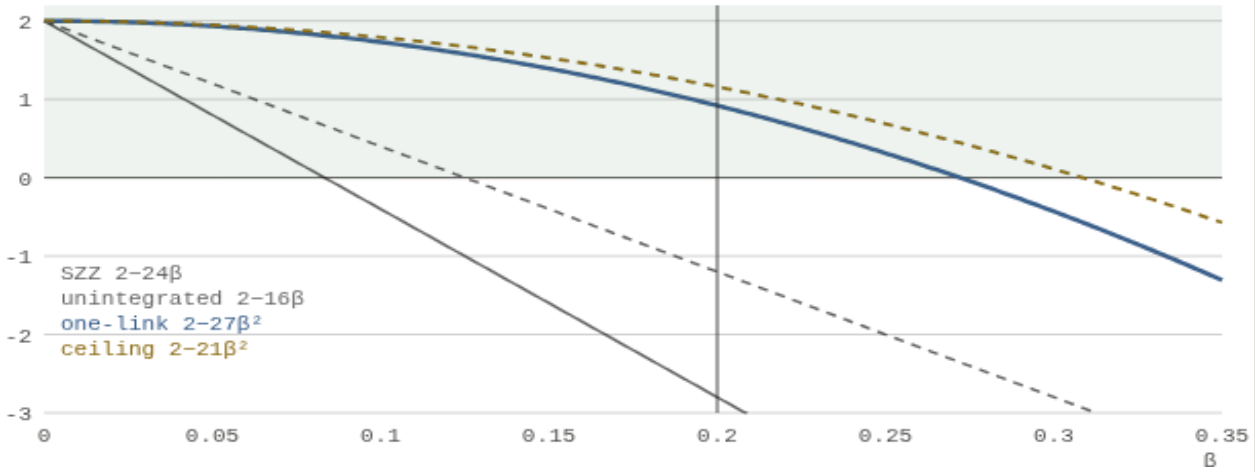
Combining (9) with $\mathrm{Ric} = 2$, the curvature of the marginal measure of the remaining links is at least $K' = 2 - 27\beta^2$, and the threshold is

$$\beta < \sqrt{2/27} = 0.27217 \qquad (\beta_{\mathrm{SZZ}} < 0.06804) \quad (11)$$

This is $4\sqrt{2/3} = 3.27$ times $1/12$, and 2.18 times the ceiling $1/8$ Computed. For d where a family exists, the threshold is $\sqrt{2/3}/(d-1)$, and the ratio 3.27 does not depend on d Paper (Lean covers only $d = 4$).

Figure 2 Four lower bounds $K(\beta)$ on the curvature. The knob and "Play" move β, and each value and whether $K >$

0 appear below.




```
β = 0.200, SZZ -2.800 (not usable), unintegrated -1.200 (not usable), one-link +0.920 (usable),
ceiling +1.160 (upper limit of this road)
```

Computed The numbers in this figure were computed on this machine (the formulas of Table 3). If JavaScript is disabled, Table 3 holds the content.

Table 3 Lower bounds on curvature and thresholds for four-dimensional $SU(2)$ Computed

Estimate	$K(\beta)$	Threshold	$\beta = 0.2$	Level
The counting of Shen–Zhu–Zhu	$2 - 24\beta$	0.0833	−2.80	Known
Best without integrating	$2 - 16\beta$	0.1250	−1.20	Paper ($\leq -16\beta$) + the numerical search of the paper
Integrating a complete family first	$2 - 27\beta^2$	0.2722	+0.92	Lean + Known
The ceiling of this road	$2 - 21\beta^2$	0.3086	+1.16	Exact computation by integer arithmetic in paper § 8 (not machine-checked; not recomputed on this machine; § 09)

The reach of the cluster expansion of Osterwalder–Seiler (1978), the strong-coupling classic, cannot be compared numerically in this chapter (source table: bibliographic data only). The crude Kotecký–Preiss estimate of paper § 9 gives $\beta \lesssim 0.014$ (Mayer form) to 0.06 (character expansion), the former of the same order as 5-04 Table 3. On the other hand, a necessary condition for the character expansion to converge absolutely allows β up to 0.4–0.8, and a careful expansion may reach further. **So the paper does not claim to have widened the strong-coupling range, and says only that "within the Bakry–Émery method the constant improves by a factor of 3.27".** The same dynamical method has also been used for the area law (Cao–Nissim–Sheffield) Known .

```
08
```

Detour: complete families stop at four dimensions — one line of coding theory

A section you may skip.

Complete families exist in $\mathbb{Z}^d$ only for $d \le 4$ Lean (`no_perfectZ_of_five`, with no periodicity assumption. On a box of period L it is equivalent to "$d \le 4$ and L even": `PerfectFamily.perfectK_one_iff`, $d \ge 2$). 23-02 § 02 shows it by counting vertices. The paper's argument on paper is coding theory. We prepare three words. A collection of strings of 0s and 1s of length n is a **binary code** and its strings are **words**; the **Hamming distance** of two strings is the number of coordinates where they differ, and the **minimum distance** is the smallest distance between two different words.

Inside a unit d-cube, the position of a family link of direction μ is fixed by the remaining $d-1$ coordinates (0 or 1), so they form a binary code C_μ of length $n = d-1$. The number of words comes from counting faces. There are 2^{d-2} faces of directions μ, ν, and each face contains exactly one family link, so $|C_\mu| + |C_\nu| = 2^{d-2}$; solving simultaneously for three directions gives $|C_\mu| = 2^{d-3}$. From the condition on the faces of three-dimensional sub-cubes, words can be told apart after deleting any two coordinates, that is, the minimum distance is at least 3. Balls of radius 1 around words at distance 3 or more do not overlap (the Hamming bound), so

$$2^{d-3}(1+n) \le 2^n \iff d \le 4 \qquad (d=4:\ 8=8, \quad d=5:\ 20>16) \ (12)$$

$d = 4$ is the case of equality: the repetition code of length 3, $\{000, 111\}$, fills $\{0,1\}^3$ exactly Computed . This "4" is a combinatorial number of the lattice, a coincidence unrelated to physical four-dimensional spacetime. For $d \ge 5$ only "at most one" families can be taken (density at most $1/d$ Lean), and the improvement drops to a factor of 1.72.

09

Limits and position of the method — 21 and 27, large β, SU(N), what is Lean

The four limits of paper § 8 are different faces of the fact that "a local curvature method can move the constant but not the shore".

(a) 27 and 21. An explicit configuration gives exactly $\lambda_{\min}(\mathrm{Hess}\,\Phi) = -21\beta^2$ (exact computation by integer arithmetic in $\mathbb{Z}[e^{i\pi/3}]$ in paper § 8; not machine-checked, not recomputed on this machine). The true value lies between 21 and 27, and the ceiling is $\sqrt{2/21} = 0.3086$.

(b) Large β. When β is large, integrating first makes things worse. $\lambda_{\min}$ at $\beta = 1$ is still -12.1 against -16 (with against without integration), but at $\beta = 2$ it reverses to $-43.1\beta^2$ (without integration -32β). At either value $K < 0$, and the criterion cannot be used (numerical calculations of the paper; not recomputed on this machine). As in § 03, the

criterion demands curvature at every configuration, and $\mathrm{Ric} = 2$ does not move with β, so the threshold does not exceed $O(1)$.

(c) It cannot be repeated. Integrating one more link makes the weight no longer of Bessel type Paper .

(d) $SU(2)$ only. "Tilting does not increase fluctuations" is already false for $U(2)$ with $M = \mathrm{diag}(s, 0)$. What the paper measured directly is mainly the star constant for $U(N)$. The threshold measured for $N = 2, 3, 4$ is nearly constant ($\beta_{\mathrm{SZZ}} \approx 0.093$), but for large N it is estimated as $\beta^* \sim 0.0962/\sqrt{N}$ (paper § 8), falling below the N-independent threshold of Shen–Zhu–Zhu for $N \gtrsim 21$. In the 't Hooft limit the gain disappears (numerical calculations and asymptotic estimates of the paper; not recomputed on this machine).

What is Lean. The Lean theorems are (7), the marginal density $e^{-\Phi}$, the three factors, (9), and $d \leq 4$ for families. $\mathrm{Ric} = 2$ and the criterion are standard facts Known , and the step to exponential decay of correlations and uniqueness (a transfer of the argument of Shen–Zhu–Zhu) is an **outline** with constants Paper . That is, the lattice-unit mass gap at $\beta < 0.2722$ is not a Lean theorem. Within the range the paper searched the same construction was not found, but the paper writes that it may be known in the context of variance reduction for Monte Carlo methods.

The curvature method works uniformly in a world with $C_P = O(1)$, where the correlation length is a few lattice spacings, the side opposite what is wanted in the continuum limit (23-01). However much this estimate is improved, it comes no closer to the mass gap of the problem statement.

What this chapter can say

Matter	Said	Not said
(2)–(4), (7)	The criterion, one-link integration Known	Details of the criterion (original: bibliographic data only)
Table 1	The single-link gap grows with κ, while the criterion covers only $\kappa < 2$ Computed	The gap of the whole lattice
§ 04	The conversion of $\beta < 1/12$ Known , the ceiling $1/8$ without integration Paper	The reach of the cluster expansion (original: bibliographic data only)

Matter	Said	Not said
(9), (11)	Lower bound $-27\beta^2$ Lean , curvature threshold 0.2722	The best constant. The mass gap is an outline on paper. It cannot be said that the range is widened
§ 08	Families for $d \le 4$ Lean	Relation to physical four dimensions (none)
The whole chapter	An example on the strong-coupling side at fixed lattice spacing	Shows nothing about the continuum limit or the mass gap of the problem statement

Sources and reproduction

Item	Kind	Source / tools
The theorems, constants and limits of § 04– § 09	Text § 1– § 11 (draft, not submitted, not refereed)	The paper on this site “One-link integration before Bakry–Émery for SU(2) lattice Yang–Mills” (version of 2026-09-25, list of papers)
The condition and quotation in § 04	Abstract (the numbering in the text is as cited by the paper above)	H. Shen, R. Zhu, X. Zhu, Commun. Math. Phys. 400 (2023) 805–851, arXiv:2204.12737
The criterion (2)	Bibliographic data only	D. Bakry, M. Émery, Diffusions hypercontractives, Séminaire de Probabilités XIX (1985) 177–206
That there is nothing below the -16β of § 04, and the values in § 09 (a)(b)(d) ($-21\beta^2$, $-12.1\beta^2$, $-43.1\beta^2$, the counterexample for $U(2)$, the thresholds for $N = 2, 3, 4$)	Calculations of the paper (not recomputed on this machine)	§ 8 of the paper above (numerical search, integer arithmetic; not machine-checked)
The Kotecký–Preiss estimate in § 07	A crude estimate of the paper	§ 9 of the paper above
The area law in § 07	Abstract	S. Cao, R. Nissim, S. Sheffield, Dynamical approach to area law for lattice Yang–Mills, arXiv:2509.04688
The strong-coupling classic	Bibliographic data only	K. Osterwalder, E. Seiler, Ann. Phys. 110(2) (1978) 440–471
Turán type $F_1^2 \ge F_0 F_2$	Abstract	Á. Baricz, S. Ponnusamy, Proc. Amer. Math. Soc. 141(2) (2013) 523–532, arXiv:1010.3346

Item	Kind	Source / tools
Tables 1–3, Figures 1 and 2, the ratio in § 06, (11)(12)	Computed on this machine	Python 3, NumPy, SciPy: gap_s3.py (tridiagonal eigenvalue problems for $l = 0, 1, 2$, and the numerical integral of the weight of the region of negative curvature in Table 1), onelink_series.py (coefficients of $F_1^2 - F_0 F_2$ checked in rationals up to $n \leq 30$), star_check.py, thresholds.py

Next chapter to read: 6-00 Entrance — the promise of this volume. To read alongside it: 7-01 What constructive field theory is (between the two shores). The families are used again in 23-02. For Lean, the guide and the list.