Volume 3 Toward the mass gap — the questions that remain, and this book's results

Part 17 From the side of physics

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
17-01	How physics sees it	2
17-02	Mechanisms of confinement	15
17-03	The string picture	28
17-04	Large N	41
17-05	String theory and extra dimensions	53
17-06	How extra dimensions become invisible	65
17-07	Where gauge fields come from	77
17-08	Holography	89
17-09	The mass gap becomes geometry	101
17-10	Back to four dimensions	114
17-11	Supersymmetry and Seiberg–Witten	126
17-12	The sentence that plays the role of an equivalence principle	138
17-13	Where the way physics digs parts from mathematics	150

2026-10-01 · chapter Yang–Mills and the mass gap Part 17 From the side of physics asymptotic freedom, dimensional transmutation, confinement

How physics sees it — freedom at short distances, a single scale, the tube of confinement, and the premise that "the theory exists"

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 split the picture physics draws of pure Yang–Mills theory into three parts, asymptotic freedom, dimensional transmutation and confinement, and look at each with formulas and numbers. We read how the ratios of the lattice glueball masses connect with "there is a single scale". We check that the whole picture rests on the premise that "the theory exists", and that confinement and the mass gap are different claims in mathematics.

Prerequisite chapters — 4-07 Dimensional transmutation and Λ (and 4-06 before it). Easier with: 3-02 (gauge fields), 5-01 (the lattice and β), 4-11 (the definition of the gap). The vocabulary of Wilson loops and the lattice is defined on the spot as far as needed. Within Volume 3, this chapter is read after Part 16 (where things stand, and the questions that remain).

Said first — This is a chapter of pictures at the level of physics. "Confinement", "a single scale" and "the glueball mass" are not theorems of the four-dimensional continuum theory (label Physics). This book does not itself show either the mass gap or the construction of the continuum limit. For the construction of the continuum limit, Volume 2 reads the one published on 2026-11-10 (a theory satisfying the Wightman axioms; it does not show the mass gap). The lattice numbers are values from primary sources; only the ratios and conversions were recomputed on this terminal.

The order of this chapter

1. The question of physics and the question of mathematics — how it is, rather than whether it exists
2. Short distances: the coupling becomes weak — the minus sign of the β function

3. A single scale is born: dimensional transmutation — every mass is a constant multiple of Λ
4. Long distances: Wilson loops and the area law — the definition of confinement, and its linear case
5. The flux tube and the Lüscher term — the picture of the tube, and the correction at long distances
6. The physical face of the gap: glueballs — the lattice ratios $m/\sqrt{\sigma}$
7. Detour: where the unit MeV comes from — borrowing a unit
8. The premise that "the theory exists" — at which stage of the numbers it enters
9. Confinement and the gap are different claims — examples with only one of them, and the implication on the lattice
10. The tools of physics, and the shore of each — a map of Part 17
11. Detour: why the gap has a minimum — the picture of being squeezed from both sides — on the large side anti-screening, on the small side uncertainty

---

01

The question of physics and the question of mathematics

The prize problem asks "does the theory exist, and is there a gap in the spectrum?" (0-03). The questions of physics are "what is the mass of the lightest particle?" and "how does the force change as a pair of quarks is pulled apart?". Physics takes the theory as already existing and asks about its contents.

Jaffe and Witten list three properties that QCD (Yang–Mills theory with quarks) must have at the quantum level Known . (1) The mass gap: 'every excitation of the vacuum has energy at least Δ'. This is why the nuclear force is strong yet reaches only short distances. (2) Quark confinement: physical particle states are invariant under SU(3) transformations, and a single quark is not seen. (3) Chiral symmetry breaking.

Pure Yang–Mills has no quarks, so (3) does not arise. (1) is the body of the prize problem, and (2) is set apart in § 5 of the problem statement as a "natural extension". The picture of physics tries to explain (1) and (2) in one sheet, and is made of three parts: the coupling weakens at short distances (§ 02), a single scale is born (§ 03), and at long distances the flux is confined to a tube (§ 04, § 05).

02

Short distances: the coupling becomes weak

Under renormalisation, the coupling g changes its value with the scale μ at which we look (an energy; as a distance, roughly $1/\mu$) (4-04). The β function that fixes how it changes has the following form in the one-loop computation for pure SU(N) Known .

$$\beta(g) = \mu\frac{dg}{d\mu} = -b_0\, g^3 + O(g^5), \qquad b_0 = \frac{11N}{48\pi^2} \quad (1)$$

Because of the minus sign, raising μ (looking at shorter distances) makes g smaller. This is **asymptotic freedom** (4-06). The PDG review writes, with $\alpha = g^2/4\pi$, $\mu^2 d\alpha/d\mu^2 = -b'_0\alpha^2 - \cdots$ and $b'_0 = 11N/(12\pi)$ (the same thing Computed), and says 'The minus sign in Eq. (9.3) is the origin of "asymptotic freedom"'. With the first term alone, $d(1/g^2)/d\ln\mu = 2b_0$ is constant, so

$$\frac{1}{g^2(\mu)} = \frac{1}{g^2(\mu_0)} + 2b_0 \ln\frac{\mu}{\mu_0} \quad (2)$$

With SU(3) and $g^2(\mu_0) = 1$, we get $g^2 = 0.757$ at $\mu = 10\mu_0$, and still 0.342 at $10^6\mu_0$ Computed . The weakening is logarithmic. In the words of the problem statement, 'at short distances the field displays quantum behavior very similar to its classical behavior'. But (1) is a perturbative computation. That the continuum theory really behaves like this at short distances is one of the properties the problem statement requires of a solution.

03

A single scale is born: dimensional transmutation

Reading (2) in the direction of lowering μ, $1/g^2$ becomes 0 at a certain scale Λ.

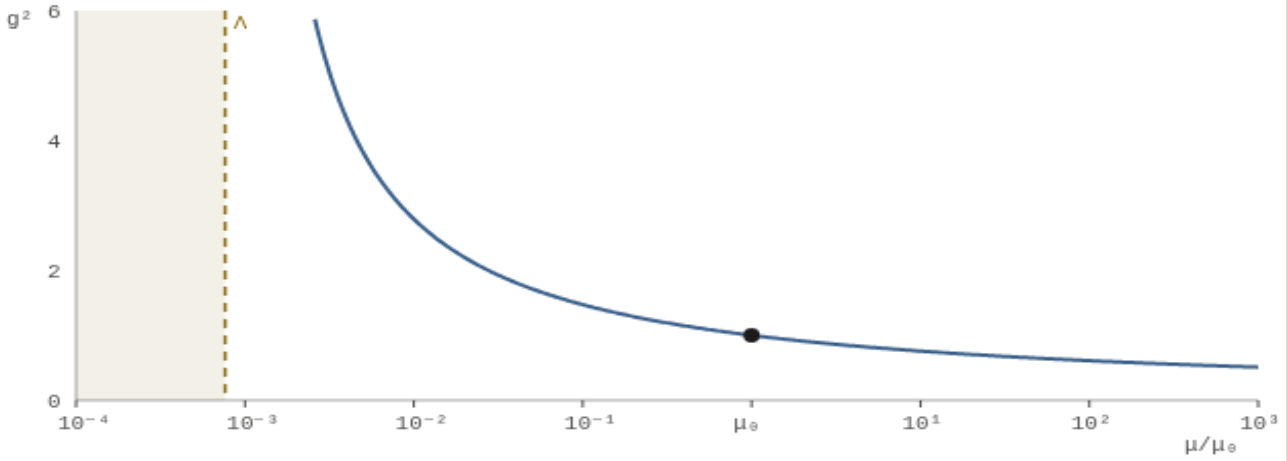
$$\Lambda = \mu\, \exp\!\Big(-\frac{1}{2b_0\, g^2(\mu)}\Big) \quad (3)$$

Substituting (2), the $\ln(\mu/\mu_0)$ cancels and Λ does not depend on μ (at three points along the flow it agrees up to double-precision rounding Computed). Classical four-dimensional Yang–Mills has no constant with the dimension of mass (g is dimensionless, 1-12). Even so, in the quantum theory a scale Λ appears. This is **dimensional transmutation** (4-07). The PDG explains Λ as 'a constant of integration, which corresponds to the scale where the perturbatively-defined coupling would diverge'. What breaks at $\mu = \Lambda$ is the one-loop formula, not the theory. From here physics concludes that every quantity with a dimension is a constant multiple of Λ Physics .

$$m_0{}^{++} = c_1\Lambda, \qquad \sqrt{\sigma} = c_2\Lambda, \qquad \Delta = c_3\Lambda \quad (4)$$

$m_0{}^{++}$ is the mass of the lightest glueball (§ 06), σ the string tension (§ 04), and c_i pure numbers fixed by N. This "single scale" looks obvious in physics as dimensional analysis, and is not proved in mathematics ("the diffusive-relaxation-time assumption" in the record of this book's study; 17-12, 23-01). Also, $\exp(-1/(2b_0 g^2))$ has all Taylor coefficients 0 at $g = 0$. If (4) is granted, Δ too appears at no order of the perturbation series. This runs the same way as the problem statement's 'the mass gap is not classically visible' (4-12).

Figure 1 Changing, with the knobs, the coupling $g^2(\mu_0)$ at the reference scale μ_0 and N moves the running coupling $g^2(\mu)$ (solid line) and the scale Λ where it diverges (dashed line). For SU(3), changing $g^2(\mu_0)$ from 0.5 to 1 moves Λ/μ_0 by a factor of about 1300 (three orders of magnitude)




```
SU(3), g²(μ₀) = 1.00 → Λ/μ₀ = 7.63×10⁻⁴ | increasing g² by 1% multiplies Λ by 1.074
```

Table 1 The one-loop Λ/μ_0, and the factor by which Λ changes when g^2 is increased by 1%

Group	$g^2(\mu_0)$	Λ/μ_0	Factor
SU(2)	0.5	4.45×10^{-10}	1.238
SU(2)	1.5	7.63×10^{-4}	1.074
SU(3)	1.0	7.63×10^{-4}	1.074
SU(3)	3.0	9.14×10^{-2}	1.024

Computed The numbers in this figure were computed on this machine. The curves are drawn on the spot from formulas (2) (3), and Table 1 evaluates the same formulas in Python 3 (for SU(3) at $g^2 = 0.5$, $\Lambda/\mu 0 = 5.83 \times 10^{-7}$).

The smaller the coupling, the exponentially smaller Λ is, and the more sensitive to changes. SU(2) at $g^2 = 1.5$ and SU(3) at $g^2 = 1.0$ give the same value because at one loop only Ng^2 (the 't Hooft coupling, 17-04) matters. The numerical value of Λ changes by a fixed factor depending on how g is defined, but the form of (4) is kept.

04

Long distances: Wilson loops and the area law

To measure the force in a theory without quarks, we place a pair of static colour charges from outside (a colour charge is a charge that acts as a source of the gauge field; 3-02). In Euclidean spacetime (4-02, One sign is different) take a rectangle of width R and length T; the trace of the matrix obtained by parallel transport once around it along the gauge field (the holonomy, 1-07) we call the **Wilson loop** $W_{R,T}$. The rectangle is the process "create the pair, leave it for a time T, annihilate it", and the expectation value decreases roughly as $e^{-V(R)T}$, so we define the potential by the following (Wilson 1974; the notation follows Chatterjee 2018 § 4) Known .

$$V(R) = -\lim_{T\to\infty} \frac{1}{T} \log\langle W_{R,T}\rangle \quad (5)$$

If $V(R) \to \infty$ as $R \to \infty$, the pair cannot move apart beyond a fixed distance. This is **confinement** (the definition in Chatterjee 2018 § 4). In particular, if $\langle W_{R,T}\rangle$ decreases as $e^{-\sigma RT}$ in the area RT (the **area law**), then $V(R) \approx \sigma R$ grows linearly. The coefficient σ in this linear case is the **string tension**. That it grows linearly in four-dimensional non-abelian Yang–Mills stands at the level of 'it is believed that V(R) grows like a multiple of R' (ibid. § 4) Physics .

On the lattice, which replaces spacetime by a mesh of points (5-01), the strength of the coupling is expressed by the **lattice inverse coupling** β (for the standard Wilson action $\beta = 2N/g^2$; a large β is weak coupling; this is not the β function of § 02). On the side where β is small (strong coupling) the area law is proved, but it holds for every group including `U(1)`, so it is no evidence of non-abelian confinement. The weak-coupling side for non-abelian groups in four dimensions 'remains elusive' (Chatterjee 2018; 5-02, 5-06) Known .

05

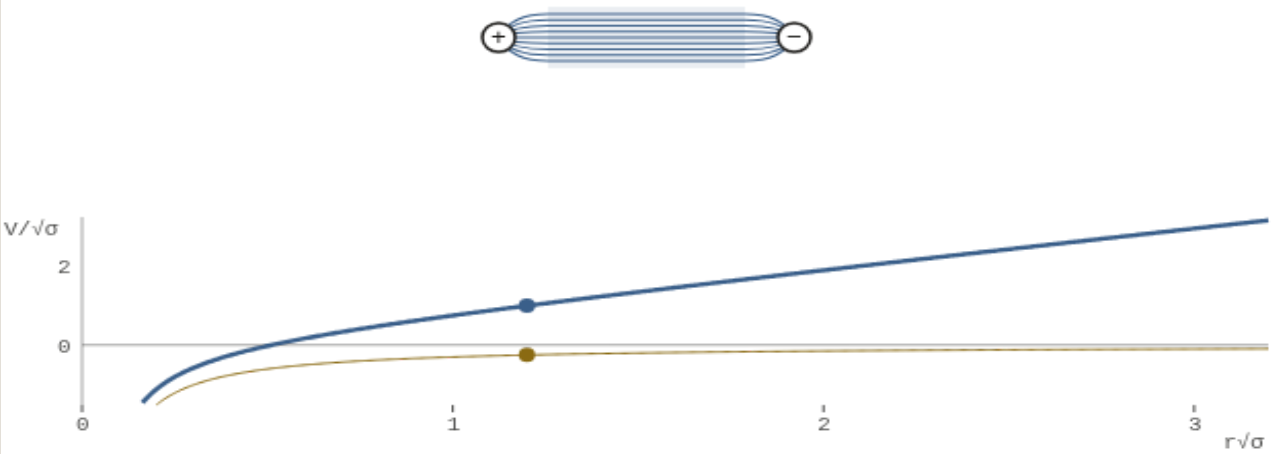
The flux tube and the Lüscher term

The picture of physics is the **flux tube**. The lines of force of electromagnetism spread out, and pulling the charges apart the energy approaches a constant value ($V \sim -c/R$). Under confinement the lines of force are squeezed into a tube of fixed thickness, and the energy is proportional to the length. As a metaphor, it is a string whose tension does not change as it is stretched. More precisely, the tube is not a fact derived from the theory but a picture that fits the lattice numbers; the effective theory of the fluctuations of the tube (17-03) takes the existence of the tube as its starting point and gives the correction at long distances as follows Physics .

$$V(r) = \sigma r + c_0 - \frac{\pi}{12\,r} + \cdots \quad (6)$$

$-\pi/(12r)$ is the Lüscher term, and its coefficient is held not to depend on the details of the theory ($-\pi(d-2)/24$ in d dimensions). c_0 is a non-universal constant. Putting (6) into the lattice length standard r_0 ($r_0^2 F(r_0) = 1.65$ for $F = dV/dr$, Sommer 1994) gives $r_0\sqrt{\sigma} = \sqrt{1.65 - \pi/12} = 1.178$, which differs by 1.6% from the lattice continuum value $1.160(6)$ (Athenodorou–Teper 2020) Computed . At r_0 the tube is still at short distance, so this is a rough guide that extends formula (6) to that distance. It agrees in outline, but the difference is about 3 times the lattice error, and it does not agree in detail.

Figure 2 Changing the distance r between the charges with the knob moves the lines of force above and the point on the energy V(r) below. The buttons switch between "electromagnetism (spreading)" and "confinement (tube)"



r√σ = 1.20 (about 0.49 fm), confinement: V/√σ = 0.98, electromagnetic sample: −0.25

r√σ	r (fm)	Confinement V/√σ	Electromagnetic sample V/√σ
0.3	0.12	−0.57	−1.00
0.51	0.21	0.00	−0.59
3.0	1.22	2.91	−0.10

Computed The numbers in this figure were computed on this machine. Confinement is $\sigma r - \pi/(12r)$ of (6) (without c_0), electromagnetism is the sample $-0.3/r$, and fm is converted with $\sqrt{\sigma} = 485$ MeV (§07). 0.51 is $\sqrt{\pi/12}$, where the two terms balance.

In QCD with quarks, signs that a long tube breaks have been reported on the lattice (SESAM 2005) Physics . Pure Yang–Mills has no quarks to create a pair midway and break the tube.

06

The physical face of the gap: glueballs

The particles of pure Yang–Mills are thought to be **glueballs**, colourless bound states made only of gluons. States are classified by J^{PC} (spin, parity, charge conjugation). 0^{++} is a state of spin 0 that changes sign neither under spatial inversion (parity) nor under the exchange of particle and antiparticle (charge conjugation). In physics the mass of the lightest 0^{++} is often called "the mass gap" Physics . Chatterjee 2018 explains the correlation length (2-08) as 'the reciprocal of the mass of the lightest glueball', and writes that the existence 'has not yet been experimentally verified'.

The gap of the problem statement is defined as a gap in the spectrum of the Hamiltonian (the problem statement calls its supremum 'the mass'), and does not presuppose the existence of particles. If the lowest excitation is an isolated one-particle state, the gap coincides with the mass of the lightest glueball (4-11). A proof that an isolated one-particle state exists is counted by the problem statement on the side of the "natural extensions".

Lattice computations replace spacetime by a mesh with lattice spacing a (the distance between lattice points) and produce the dimensionless ratio $m/\sqrt{\sigma}$.

Table 2 Glueball masses (continuum limit), and the ratios computed on this machine

J^{PC}	SU(2) $m/\sqrt{\sigma}$	SU(3) $m/\sqrt{\sigma}$	SU(∞) $m/\sqrt{\sigma}$	$m/m_{0^{++}}$: SU(2), SU(3), SU(∞)
0^{++}	3.781(23)	3.405(21)	3.072(14)	1, 1, 1
2^{++}	5.349(20)	4.894(22)	4.582(14)	1.415(10), 1.437(11), 1.492(8)
0^{-+}	6.017(61)	5.276(45)	4.711(25)	1.591(19), 1.549(16), 1.534(11)
0^{++*}	6.126(38)	5.855(41)	–	1.620(14), 1.720(16), –

Computed The errors of the ratios are a naive propagation treating them as independent, a rough guide that ignores correlations. SU(∞) is the value extrapolated in $1/N^2$. On the lattice rotation symmetry is reduced to the rotations of a cube, and spin 2 splits into two sets (E, T_2) that are measured separately. The 2^{++} of SU(∞) is the E value (4.578(11) for T_2; 5-08). 0^{++*} is an excited state.

For every N, 0^{++} is the lightest, and the ratios change only slowly with N (17-04). On the other hand, $m_{0^{++}}/\sqrt{\sigma}$ for SU(3) is 3.347(68) (Meyer 2005), 3.405(21) (Athenodorou–Teper 2020) and 3.607(87) (Lucini–Teper 2001), scattered more widely than the quoted errors. σ is obtained by subtracting the fluctuation correction from the mass of a closed tube. Formula (4) of Lucini–Teper is $am_l = a^2\sigma L - \pi/(3L)$, where m_l is the mass of a flux tube closed once around a periodic box, and L is the side of the box counted in lattice points. The authors note: 'in using eqn(4) we are assuming that we have linear confinement'. The continuum limit is an extrapolation assuming $m(a)/\sqrt{\sigma(a)} = m(0)/\sqrt{\sigma(0)} + c\,a^2\sigma$ (5-08).

07

Detour: where the unit MeV comes from

A section you may skip.

When we say "the lightest glueball is about 1.7 GeV", where does the MeV come from? Pure Yang–Mills has only one scale, Λ, and the unit is borrowed from outside. Athenodorou–Teper 2020 take $r_0\sqrt{\sigma} = 1.160(6)$ from pure Yang–Mills and $r_0 = 0.472(5)$ fm from lattice QCD with quarks, obtain $\sqrt{\sigma} = 485(6)$ MeV, and give 0^{++} as 1653(26) MeV. The recomputation on this machine gives $\sqrt{\sigma} = 484.96 \pm 5.7$ MeV and $0^{++} = 3.405 \times 485 = 1651 \pm 22$ MeV Computed . The paper states that the errors of its Table 18 are statistical only, while our ± 22 is a rough guide that includes the error of $\sqrt{\sigma}$; the meanings differ. The authors add: 'Strictly speaking this is of course not possible: the pure gauge theory and QCD … are two different theories'. What has meaning within pure Yang–Mills is the ratio 3.405.

08

The premise that "the theory exists"

The parts and numbers so far rest on the premise that four-dimensional pure Yang–Mills theory really does exist, as one theory. We list the stages of a lattice computation at which the premise enters.

Stage	What is done	Premise
Finite lattice	A finite-dimensional integral evaluated by the Monte Carlo method (5-07)	None (a well-defined probability measure)
Extracting σ	Subtracting the fluctuation correction of the tube	Linear confinement and the universality of the fluctuations
Infinite volume	Using only boxes with side at least $3/\sqrt{\sigma}$ (Lucini–Teper)	That the values do not change beyond that
Continuum limit	Extrapolating to $a \to 0$ in the form $c\,a^2\sigma$	That the limit exists, is unique, and is approached in this form

For physics this premise is natural, and the problem statement too writes that experiment and computers 'have given strong encouragement'. But the same paragraph continues 'there does not exist a convincing, whether or not mathematically complete, theoretical computation demonstrating any of the three properties in QCD', and the mathematical side is at the stage of 'nor even a precise definition' Known .

Put as a metaphor, physics draws a detailed map of land it is convinced exists, and mathematics is asked to show that the land exists. Precisely, the task is, for every compact simple group, to construct a non-trivial quantum Yang–Mills theory on $\mathbb{R}^4$ (with axioms at least as strong as those of Wightman or Osterwalder–Schrader; 4-09, 4-10) and to show $\Delta > 0$. A commonly taken road is to construct a Euclidean probability

measure with a regularisation and take the limit. A footnote of the problem statement says that weak existence, where only a subsequence converges, is not accepted unless the properties of the limit are shown separately. The numbers of physics, however precise, are "predictions made on the assumption of existence" (5-09, 7-10).

For the construction published on 2026-11-10, the committee's assessment is that it is proved in Lean that a theory satisfying the Wightman axioms can be constructed as a limit of Wilson's lattice gauge theory (Volume 2, 6-01). However, the infinite-volume limit is taken along a subsequence, and its uniqueness and the mass gap are not shown (6-04). The premise of the "continuum limit" row in the table above (that the limit exists and is unique) is still not fully met.

09

Confinement and the gap are different claims

In the picture of physics σ and m are two faces of the same Λ (formula (4)), and confinement and the gap are spoken of together. In mathematics they are separate claims.

There are examples where only one of them holds. For lattice gauge theory with finite groups, exponential decay of correlations at weak coupling is proved, while the absence of confinement is at the level of 'believed' (5-10). For compact `U(1)` in three dimensions the area law is proved at every coupling, and the continuum limit is a free scalar field (abstract of Göpfert–Mack 1982). In the converse direction, exponential decay under arbitrary boundary conditions on the lattice gives the area law for Wilson loops in representations on which the centre of the group acts non-trivially (Chatterjee 2021) `Known` . The table and the precise statements are in 17-02 § 02 (the master copy in this book).

This converse implication has a form close to "gap ⇒ confinement", but its hypothesis is a strong form of decay on the lattice. There is no theorem deriving confinement from a gap in the sense of the problem statement ($\Delta > 0$ in the continuum and in infinite volume). The variables physics mainly looks at (the tube and σ) and the target of the problem statement (Δ) are slightly offset.

10

The tools of physics, and the shore of each

Each tool of physics has a "shore" where its computations are under control. More precisely, a shore is a region near some limit ($g \to 0$, lattice $\beta \to 0$, $N \to \infty$ and so on) in which the errors of the computation can be controlled. The remaining chapters of Part 17 dig into the rows of the table.

Tool	Shore where the computation works	Level	Chapter
Perturbation theory	$g \to 0$. The gap does not show (granting (4))	Known	4-03–4-07
Strong-coupling expansion	The lattice inverse coupling β is small (strong coupling). Area law for every group including U(1)	Known	5-02
Lattice numerics	Finite a and boxes, and their extrapolation	Physics	5-08
Mechanisms of confinement, strings	Gauge-fixed variables / a long tube (confinement assumed)	Physics	17-02, 17-03
Large N, holography	$N \to \infty$ (holography also $g^2 N \to \infty$)	Physics	17-04, 17-08–17-10
Supersymmetry	Supersymmetric theories with matter fields	Physics	17-11

The theory of the problem statement — any compact simple group, finite N, no matter, the continuum $\mathbb{R}^4$ — is on none of the shores. The pictures of physics fill the space between the shores with the premise that "one theory connects them" (§ 08). The places where that premise leaves mathematics are gathered in 17-13 and 16-05.

11

Detour: why the gap has a minimum — the picture of being squeezed from both sides

This section may be skipped. It reassembles three parts of this chapter (§ 02 asymptotic freedom, § 03 dimensional transmutation, § 04 confinement) towards a single question: "why does the lightest excitation have a non-zero mass?" Physics This is a picture, not a proof. None of its steps is a theorem in the four-dimensional continuum theory (27-02).

In one sentence. Because the field binds itself with its own charge, it cannot become a long wave, only a short lump. The minimum appears where the size of an excitation is squeezed from both sides.

The large side is closed off. A massless free field (light) has no gap because a wave can be made however long the wavelength, and its energy $E = |p|$ goes to zero. A long wave needs the field to keep its direction aligned over long distances. In Yang–Mills the mediating field itself carries colour charge, so the fluctuations of the vacuum do not screen the field but anti-screen it, and the coupling grows stronger as the distance grows (§ 02, 4-06). Beyond the distance $1/\Lambda$ where the coupling becomes of order one, the field cannot keep its direction and falls apart. So waves with wavelength longer than $1/\Lambda$ cannot survive as states. The colour electric flux cannot spread out, and is bundled into a closed tube (§ 05).

The small side is closed off too. In a region smaller than $1/\Lambda$ the coupling is weak and the field is nearly free, but an excitation localised to size R cannot, by the uncertainty relation, have energy below about $\hbar c/R$ (2-05, 16-08 § 09). If $R < 1/\Lambda$ then $E > \Lambda$.

Squeezed from both sides. What keeps it from growing large is anti-screening, and what keeps it from becoming small is uncertainty, so the lightest excitation is a closed ring of electric flux of size about $1/\Lambda$. Its mass is of order Λ, and that is the gap (the glueball of § 06). In one word, the origin is self-interaction; in one word, the reason for the minimum is the squeeze from both sides.

The difference from U(1). The photon carries no charge, so the vacuum only screens, and the coupling stays weak at long distances. The electric flux spreads out, long waves are allowed, and the gap is zero (5-06). Non-commutativity takes effect at a single point, "whether the field acts on itself", and the one-loop $11/3 = -1/3 + 4$ is its face (26-03).

In the language of the strong-coupling lattice. On the Hamiltonian lattice (Kogut–Susskind), the Hamiltonian is the sum of an electric term $\frac{g^2}{2}\sum E^2$ and a magnetic term $\frac{1}{g^2}\sum(1-\operatorname{tr}U_\square)$, and at strong coupling the electric term wins. Electric flux sits on single links, and each link costs of order g^2. The smallest gauge-invariant excitation is a ring of electric flux going round one face, and its cost is a finite positive number. This is the gap at strong coupling, and at strong coupling with the lattice spacing fixed it is a theorem (5-02, 5-04). What asymptotic freedom says is that "at the distance $1/\Lambda$ the effective coupling is strong", so this picture becomes, as it is, the picture at long distances. The question that remains is that this picture does not break down as the lattice spacing goes to zero; that is the ultraviolet control row of the ledger (as existence, shown by the published construction; the form of this book's road is open) and the renormalisation comparison row (27-02, A-08).

What this chapter can say

What is said	What is not said
Known The one-loop b_0, the solution (2), and within it that Λ does not depend on μ (as a perturbative computation). The definition of the gap in the problem statement, and that the gap coincides with the mass of an isolated one-particle state if there is one. Physics A single scale, the area law and the tube, the Lüscher term, the values of Table 2.	None is a theorem of the four-dimensional continuum theory (the construction published on 2026-11-10 gives the existence of the theory, but does not show confinement or the mass gap; 6-04). Table 2 is an extrapolation presupposing that the limit exists, and σ passes through the tube correction.
Known The area law at strong coupling (every group), the decay of correlations at weak coupling for finite groups, centre symmetry and its implication for confinement on the lattice, the area law at every coupling for U(1) in three dimensions.	The area law on the weak-coupling side for non-abelian groups in four dimensions. Deriving confinement from a gap in the sense of the problem statement. That finite groups do not confine (the level at which Chatterjee 2021 writes 'believed').
Computed The ratios of Tables 1 and 2, Figures 1 and 2, the estimate 1.178 of $r_0\sqrt{\sigma}$, the recomputation in MeV.	Ratios and conversions of primary-source values and evaluations of one-loop formulas, not measurements. The errors of the ratios are rough guides ignoring correlations. 1.178 is a rough guide extending a long-distance formula to short distances.

What is said	What is not said
Physics The squeeze picture of § 11 (on the large side anti-screening, on the small side uncertainty, the minimum a ring of size $1/\Lambda$). That on the strong-coupling lattice the cost of a ring round one face is positive.	No step of the picture is a theorem in the four-dimensional continuum. That anti-screening removes long waves, that the ring is the smallest excitation, that the picture survives as the lattice spacing goes to zero.

Sources and reproduction

Item	Kind	Source / tools
The three properties, the short-distance requirement, the definition of the gap and 'the mass', the requirements of the problem (group, non-triviality, $\mathbb{R}^4$, axioms), the classification of the extension problems, the quotations from the text, the footnote excluding weak existence	Known	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory" (the problem statement of the Clay Mathematics Institute) § 1, § 2, § 4, § 5, § 6.6, footnote 2. Text checked.
The running of α, b_0', the minus sign, the explanation of Λ	Known	J. Huston, K. Rabbertz, G. Zanderighi, "Quantum Chromodynamics", Particle Data Group, Phys. Rev. D 110 (2024) 030001, § 9.1.1. Text checked.
The original papers on asymptotic freedom	Known	D. J. Gross, F. Wilczek, PRL 30 (1973) 1343. H. D. Politzer, PRL 30 (1973) 1346. Bibliographic data only (from the reference list of the problem statement).
Formula (5), the definition of confinement, the statements and quotations about the area law, strong coupling, the correlation length and glueballs	Known	S. Chatterjee, "Yang–Mills for probabilists", arXiv:1803.01950 § 4, § 5. Text checked. K. G. Wilson, PRD 10 (1974) 2445: bibliographic data only. Osterwalder–Seiler, Ann. Phys. 110 (1978) 440 and Göpfert–Mack, CMP 82 (1982) 545: abstracts (INSPIRE; the former on the infinite-volume limit at strong coupling and Wilson's confinement estimate, the latter on the Villain action and the description of the continuum limit; texts not read. The year of Göpfert–Mack is the Crossref publication year; INSPIRE gives 1981).
The definitions of the centre and centre symmetry, centre symmetry and confinement, 'believed' for finite groups	Known	S. Chatterjee, CMP (2021), arXiv:2006.16229. Definition 2.1, Theorems 2.2 and 2.4 and the quotation 'believed' checked in the text.
Decay of correlations at weak coupling for finite groups (four dimensions)	Known	A. Adhikari, S. Cao, arXiv:2202.10375, Ann. Probab. 53 (2025) 140–174 (bibliographic data from a search). Abstract and the opening of the text (restriction to four dimensions) checked.

Item	Kind	Source / tools
Table 2, $r_0\sqrt{\sigma}$, r_0, 485(6) MeV, 1653(26) MeV and the note on its error, the note on the conversion	Physics	A. Athenodorou, M. Teper, JHEP 11 (2020) 172, arXiv:2007.06422 / JHEP 12 (2021) 082, arXiv:2106.00364. Numbers and quotations checked in the text (Tables 17, 18, 23, 31, 34, formulas (8)(9)).
3.607(87), the extraction of σ (formula (4)) and the form of the extrapolation, the coefficient π/12 for Wilson loops	Physics	B. Lucini, M. Teper, JHEP 0106 (2001) 050, hep-lat/0103027 Table 5, § 4. Text checked.
The Lüscher term, the definition of r_0, 3.347(68)	Physics	M. Lüscher, NPB 180 (1981) 317, R. Sommer, NPB 411 (1994) 839, H. B. Meyer, hep-lat/0508002: bibliographic data only (Meyer's value via the table of Gürsoy–Kiritsis–Nitti, arXiv:0707.1349).
Signs of tube breaking in QCD	Physics	G. S. Bali et al. (SESAM), PRD 71 (2005) 114513, hep-lat/0505012. Abstract checked.
Table 1, Figure 1, the ratios of Table 2, the estimate of $r_0\sqrt{\sigma}$, the table of Figure 2, the recomputation in MeV	Computed	Python 3 on this machine (standard library, double precision). Errors propagated treating them as independent. $\hbar c = 197.3269804$ MeV·fm.
The picture of § 11 (squeeze)	Physics	Put into words by one of this book's authors (Shiori) in a dialogue on 2026-10-29. The primary sources for the parts are Gross–Wilczek and Politzer (1973, row above), J. Kogut, L. Susskind, PRD 11 (1975) 395 (the Hamiltonian lattice and electric flux at strong coupling; bibliographic data only), K. G. Wilson, PRD 10 (1974) 2445.

Next chapter to read: 17-02 Mechanisms of confinement — on to the explanations from physics of why the tube of § 05 forms (dual superconductivity, monopoles, centre vortices, Gribov–Zwanziger).

Mechanisms of confinement — reading four pictures by what each casts in the leading role, and where each is cut off from mathematics

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 leading roles and the support of the four pictures of confinement (dual superconductivity, abelian projection, centre vortices, Gribov–Zwanziger), and where each is cut off as a mathematical statement. That confinement and the gap are different sentences.

Prerequisite chapters — 17-01. Also 5-02 (Wilson loops), 4-08 (Gribov copies), 1-04 (the centre of a group).

Said first — All four pictures are physical pictures, and there is no theorem deriving confinement in the four-dimensional non-abelian theory. This chapter shows neither confinement nor the gap. The model of Figure 2 is a toy.

The order of this chapter

1. What we want to explain — the area law
2. Confinement and the gap are different sentences — the area law and exponential decay, a conditional implication
3. Dual superconductivity — turning the magnetic flux tube inside out (Figure 1)
4. Abelian projection and monopoles — singularities of diagonalisation / Detour: Elitzur's theorem
5. Centre vortices — membranes that multiply by a centre element
6. A model of a vortex gas — the area law and the perimeter law (Figure 2)
7. The Gribov region and the horizon — from the gauge-fixed side
8. The propagator and positivity — a correlation that would not go negative for a particle (Figure 3)
9. Setting the four pictures side by side — leading roles, support, what is missing

01

What we want to explain

"Confinement" is a word that comes from the observation that quarks cannot be extracted on their own. Pure Yang–Mills theory has no quarks, so we place two infinitely heavy test charges and ask about the energy of pulling them apart (17-01). On the lattice, the expectation of the trace of the matrix obtained by multiplying link variables along a closed path C is the Wilson loop (5-01); for an $R \times T$ rectangle $W \approx e^{-V(R)T}$, where $V(R)$ is the energy of the pair of charges.

$$W(C) = \left\langle \frac{1}{N} \operatorname{tr} \prod_{\ell \in C} U_\ell \right\rangle, \qquad V(R) = -\lim_{T\to\infty} \frac{1}{T} \ln W(R,T) \quad (1)$$

There are two ways a large loop can decrease.

$$W(R,T) \approx e^{-\sigma RT} \ \text{(area law)}, \qquad W(R,T) \approx e^{-c\,(2R+2T)} \ \text{(perimeter law)} \quad (2)$$

Under the area law $V(R) \approx \sigma R$, and the energy grows in proportion to the distance. σ is the **string tension**. Under the perimeter law V is bounded. Example: with $\sigma = c = 0.05$ and $R = T = 10$, $-\ln W$ is 5 and 2; with sides of 20 it is 20 and 4. To tell them apart we use the Creutz ratio, which is σ for a pure area law and 0 for the perimeter law Known .

$$\chi(R) = -\ln \frac{W(R,R)\,W(R-1,R-1)}{W(R,R-1)\,W(R-1,R)} \quad (3)$$

To ask about the "mechanism" is to ask **what in the vacuum** produces the area law. The area law at small β (the lattice inverse coupling; for $SU(N)$, $\beta = 2N/g^2$, smaller meaning stronger coupling) is a theorem (5-02), but it holds for every group including `U(1)`, so it is no explanation for the non-abelian case Known . The question remains on the side of large β, heading towards the continuum limit.

02

Confinement and the gap are different sentences

The mass gap is exponential decay of correlations (4-11), confinement is the area law, and as logic they are different claims. What the Clay problem statement requires is $\Delta > 0$; confinement is set apart as a "natural extension" Known .

Table 1 Presence of confinement (area law) and of the gap (exponential decay), and the level

Model	Area law	Exponential decay	Level

Model	Area law	Exponential decay	Level
Any group, small β	Yes	Yes	Theorem (5-02)
4-dimensional `U(1)`, large β	No	No (photon)	Theorem (5-06)
Finite group, large β	Believed not	Yes (in a weak sense)	Decay is a theorem (5-10)
3-dimensional compact `U(1)`	Yes	In physics the photon acquires a mass	Area law is a theorem (§ 03)
4-dimensional `SU(N)`, continuum limit	Expected	Expected	Unproved

On the lattice there is an implication in one direction. Chatterjee (2021) showed the following Known . If the representation forming the Wilson loop acts non-trivially on the centre of the group (§ 05) (the fundamental representation of `SU(N)` does), and centre symmetry is unbroken (that is, in slab-shaped regions of finite thickness, the measure is invariant under centre transformations whatever boundary condition is taken; Definition 2.1 of that paper), then for any measure on the infinite-volume lattice $\mathbb{Z}^d$ the following holds for some function V with $V(R) \to \infty$ as $R \to \infty$.

$$|\langle W(R \times T) \rangle| \leq e^{-V(R)\,T} \tag{4}$$

This means that V is unbounded, not that $\sigma > 0$. Assuming a strong form of the gap, exponential decay under arbitrary boundary conditions, yields even the area law (the Theorem from Uniform Decay to the Area Law; Theorem 2.4 of the same paper). The decay shown for finite groups is decay of two-point correlations under a specific boundary condition, weaker than the "decay under arbitrary boundary conditions" that this theorem assumes. So it does not contradict the finite-group row of Table 1 (§ 2.4 of the same paper). There is no theorem deriving confinement from a gap in the Clay sense.

03

Dual superconductivity: turning the magnetic flux tube inside out

A superconductor expels magnetic fields (the Meissner effect), and a magnetic field entering from the surface decays over the penetration depth λ_L. If forced through, it is bundled into thin tubes (Abrikosov vortices), and the flux of one tube is an integer multiple of $h/2e$ Known .

$$\nabla^2 \mathbf{B} = \mathbf{B}/\lambda_L^2, \qquad B(x) = B_0\, e^{-x/\lambda L} \tag{5}$$

A tube with energy density $B^2/2$, cross-section S and flux $\Phi = BS$ carries $E/\ell = \Phi^2/(2S)$ per unit length. In vacuum S can be widened to bring this close to 0, but in a

superconductor widening costs something, and E/ℓ stays at a positive constant. If a pair of monopoles were embedded, the energy should be proportional to the distance (a schematic estimate).

Maxwell's equations with a magnetic current j^m added are symmetric under exchanging electric and magnetic ($\tilde{F}$ is the dual; only here do we use real-time notation) Known .

$$\partial^\mu F_{\mu\nu} = j^e_\nu, \quad \partial^\mu \tilde{F}_{\mu\nu} = j^m_\nu; \qquad \mathbf{E} \to \mathbf{B},\ \mathbf{B} \to -\mathbf{E},\ j^e \to j^m,\ j^m \to -j^e \tag{6}$$

If the vacuum is a **condensate of colour-magnetic monopoles**, what is expelled is the colour electric field, the colour electric flux forms a tube, and quarks are confined. This is the picture of dual superconductivity, said to have been proposed independently by 't Hooft (a 1975 talk) and Mandelstam (1976) Physics . Nambu (1974) is regarded as its forerunner. More precisely, the monopoles are not fields appearing in the action but objects defined only once a gauge is chosen (§ 04), and there is as yet no way to write "condensation" as a statement about the measure.

Figure 1 Left: in vacuum the lines of force spread out. Right: in a vacuum where monopoles have condensed, the electric flux forms a tube (schematic).



Vacuum (E/ℓ → 0) Condensed vacuum (E/ℓ = σ)

Di Giacomo et al. (2000, `SU(2)`) measured an order parameter for monopole condensation on the lattice and reported that "confinement appears as a transition to a dual superconductor" Physics . The context is the deconfinement transition at finite temperature, and the mechanism was not shown. Neither step of "condensation ⇒ area law ⇒ gap" is proved for the four-dimensional non-abelian case.

Examples where the mechanism can be computed in an abelian model

In three-dimensional compact `U(1)` monopoles act as instantons; Polyakov (1977) obtained the area law semiclassically, and Göpfert–Mack proved $\sigma > 0$ at every coupling (on the lattice with the Villain action) Known . But the abstract of the same work also states that the continuum limit is a free scalar field and that the string tension in physical units goes to ∞ (text not read). From this mechanism "the masses and the string tension are fixed by a single scale" (17-12) does not follow. In four dimensions, for an $\mathcal{N} = 2$ supersymmetric theory broken to $\mathcal{N} = 1$ by a mass term,

Seiberg–Witten (1994) described confinement by monopole condensation (17-11) Physics .

04

Abelian projection and monopoles

't Hooft (1981) found monopoles, absent from the action, as **singularities of the choice of gauge**. Choose a quantity $X(x)$ that transforms as $X \mapsto gXg^{-1}$ (for example F_{12}) and diagonalise it at each point.

$$g(x)\,X(x)\,g(x)^{-1} = \operatorname{diag}\big(\varepsilon_1(x),\dots,\varepsilon_N(x)\big) \quad (7)$$

ε_i are the eigenvalues. The remaining symmetry is $U(1)^{N-1}$, and the diagonal gluons are the "photons". At points where eigenvalues coincide the diagonalisation is not determined, and these are the monopoles Physics . Example: in `SU(2)` with $X = \sum_a X^a \sigma^a/2$ the eigenvalues are $\pm|\boldsymbol{X}|/2$, and they coincide at $\boldsymbol{X}=0$, which is 3 real conditions. In three dimensions this is a point, in four a line (a world line) (the same for general N, the detour of 1-01). If $\boldsymbol{X} \approx \underline{x}$ near the zero, then $\boldsymbol{X}/|\boldsymbol{X}|$ covers the sphere surrounding the zero once, and that number of times is the magnetic charge Known .

On the lattice, the maximal abelian gauge (MAG, Kronfeld et al. 1987) brings the links close to diagonal.

$$R[U] = \sum_{x,\mu} \operatorname{tr}\left[\,U_\mu(x)\,\sigma^3\,U_\mu(x)^\dagger\,\sigma^3\,\right] \;\to\; \max \quad (8)$$

No method is known for reliably finding the global maximum, and in practice local maxima (Gribov copies) are used. Suzuki–Yotsuyanagi (1990) reported that the string tension of the projected links accounts for most of the original (abelian dominance), and Greensite's review (2003), citing $(92 \pm 4)\%$ for `SU(2)` (Bali et al. 1996), writes that "abelian dominance is not yet a proof of confinement by monopoles" Physics .

Detour: Elitzur's theorem

A subsection you may skip.

A gauge transformation g at the starting point of a link U_ℓ alone gives $U_\ell \to gU_\ell$ (the other links touching that point change too, but the action and the measure are invariant). So $\langle U_\ell\rangle = g\langle U_\ell\rangle$ holds for every g, and averaging over the Haar measure (5-01) gives

$$\langle U_\ell\rangle = \Big(\int_{SU(N)} g\,dg\Big)\langle U_\ell\rangle = 0 \quad (9)$$

Elitzur's theorem (1975) states that, even in infinite volume, the expectation of a local quantity that is not gauge invariant stays 0 Known . Unlike a global symmetry (2-09), a gauge transformation at one point meets no barrier. A monopole defined by fixing a gauge is a function of the configuration once the gauge-fixing procedure is included. But since the maximum is not unique (Gribov copies), it is not determined as a single function, nor is it a local quantity.

05

Centre vortices: membranes that multiply a loop by a centre element

The **centre** of SU(N) is $Z_N = \{e^{2\pi i k/N} I\}$, which commutes with every element. A centre element z acts in each representation as z^k, and this $k \pmod N$ is the **N-ality**. The fundamental representation has 1 and the adjoint 0 Known .

The starting point is the lattice observation that the string tension of large loops is fixed by the N-ality alone, and is 0 for the adjoint (screened by the gluons themselves) (Greensite's review § 4.3) Physics . A centre vortex is a line in three dimensions and a surface in four, and multiplies a loop linked with it by one centre element (-1 for SU(2)). Adjoint loops do not notice the vortex. More precisely, the thinnest vortex on the lattice is a configuration in which the links piercing some region are all multiplied by a centre element, and only the plaquettes on the rim of the region change (inside, the centre elements cancel). This rim is the vortex, and only loops linked with the rim pick up the centre element. Divide the plane into squares of area $\Sigma_{\min}$; if the $Z(C_i)$ picked up by the loops of the squares are nearly independent, a large loop splits into a product and obeys the area law (formulas (94)–(96) of the same review).

$$\langle W_k(C) \rangle \approx e^{-\sigma(k)\,\Sigma}, \qquad \sigma(k) = -\ln \langle Z(C_i)^k \rangle / \Sigma_{\min} \quad (10)$$

Confinement becomes the question of whether the vortex surfaces spread through the volume (percolate). Numerically, Del Debbio–Faber–Greensite–Olejník (1997) reported that projected vortices carry most of the string tension (centre dominance), and de Forcrand–D'Elia (1999) that removing the vortices removes confinement Physics . As a theorem, Tomboulis–Yaffe (1985, SU(2)) bounded the Wilson loop from above by a function of the vortex free energy Known . The vortex free energy is $-\ln$ of the ratio of the partition function $\mathcal{Z}_{\mathrm{tw}}$ with a centre twist put into the boundary condition in one direction to the partition function $\mathcal{Z}$ without it, that is $F_v = -\ln(\mathcal{Z}_{\mathrm{tw}}/\mathcal{Z})$, and it is gauge invariant. In the same family of gauge-invariant quantities is the 't Hooft loop (an operator inserting a centre vortex along a loop). Formula (4) is also a theorem bounding Wilson loops from centre symmetry. Tomboulis (2007) claimed a proof at every coupling for SU(2), and Ito–Seiler point out a gap in the argument.

Vortices defined by projection depend on the gauge. The hypotheses of the conditional theorems concentrate on a single point, that vortices condense at weak coupling in four dimensions (gauge-invariantly, decay of the vortex free energy), and this is not proved.

For how this book uses this, see 26-08 § 03 (where the average over vortex containers and the Migdal–Kadanoff interpolation stopped in the literature).

06

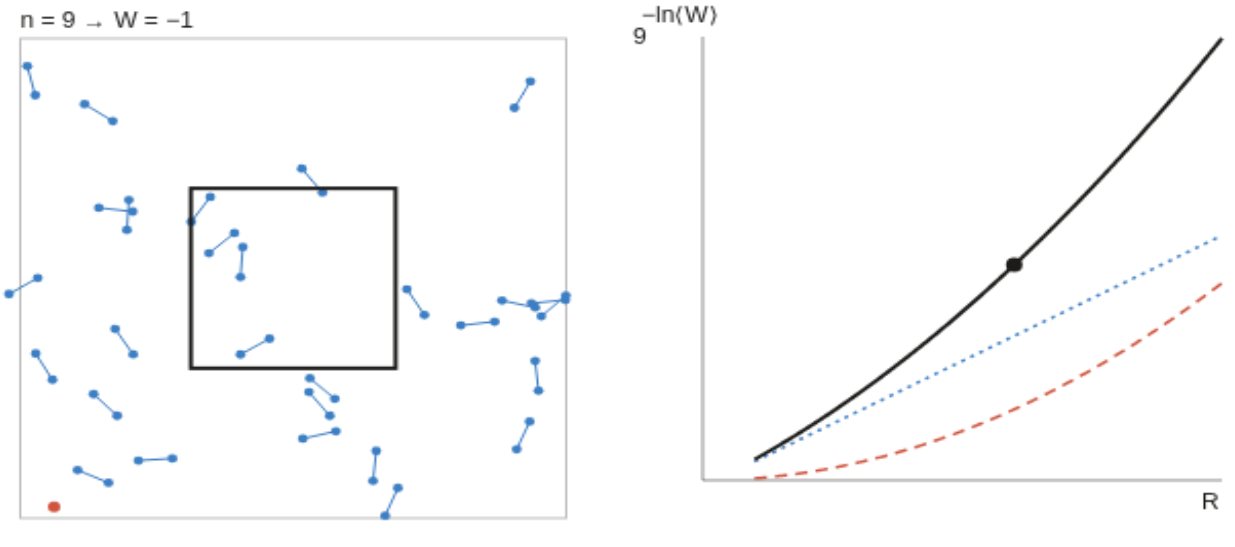
A model of a vortex gas: counting the area law and the perimeter law

Scatter over a plane the points where `SU(2)` vortices pierce it, and with n the number of points inside the loop set $W = (-1)^n$. There are two kinds of points. (a) Independent points of density ρ_f (cross-sections of extended vortices). (b) Pairs of points at distance d, with point density ρ_p (small vortices piercing the plane twice). A pair changes the sign only when it straddles the edge. For an $I \times J$ rectangle the following holds exactly (averaging over directions the area of the symmetric difference of two translated rectangles) Computed .

$$-\ln\langle W(I,J)\rangle = 2\rho_f\, IJ + 2\rho_p\Big[\frac{2d}{\pi}(I+J) - \frac{d^2}{\pi}\Big] \qquad (I,J \geq d) \quad (11)$$

The first term comes from the number n of independent points following a Poisson distribution with mean $\mu = \rho_f IJ$, so that $\mathbb{E}[(-1)^n] = \sum_n e^{-\mu}\mu^n(-1)^n/n! = e^{-2\mu}$ (1-08). The first term is the area law and the second the perimeter law, and the Creutz ratio is $2\rho_f$. However many small vortices there are, they contribute nothing to the string tension.

Figure 2 Changing R, ρ_f and ρ_p changes the sample sheet on the left (red = independent, blue = pairs) and, on the right, −ln⟨W⟩ (solid line, formula (11)), the area term (red dashed) and the perimeter term (blue dotted).



```
R = 6、ρf = 0.020、ρp = 0.20 : −ln⟨W⟩ = 1.44 (area) + 2.93 (perimeter) = 4.37. Creutz ratio = 0.040.
This sheet has n = 9
```

Computed The curves are computed on the spot from formula (11), and the points on the left are one random sample. Formula (11) was checked on this machine against a Monte Carlo (MC) of one million samples (Table 2).

Table 2 −ln⟨W(R,R)⟩: formula (11) and MC (lattice spacing 1, d = 1; in brackets the MC standard error in the last digit; "—" means it cannot be measured)

R	Independent 0.05	MC	Pairs 0.2	MC	Mixed	MC
2	0.400	0.399(1)	0.891	0.890(2)	1.051	1.052(3)
4	1.600	1.596(5)	1.910	1.920(7)	2.550	2.546(13)
6	3.600	3.574(36)	2.929	2.929(19)	4.369	4.255(70)
8	6.400	—	3.947	3.921(50)	6.507	—
Creutz ratio	0.100		0.000		0.040	

Mixed means $\rho_f = 0.02$, $\rho_p = 0.2$. The MC standard error grows as $-\ln W$ grows, and is at most 0.02 for $-\ln W \le 3$. Mixed at $R = 6$ has error 0.070, and its difference 0.11 from the exact formula is about 1.6 standard errors. $e^{-6.4} \approx 0.0017$ is about the size of the noise 10^{-3} of an average of ± 1, and cannot be measured. This difficulty is the same as in real lattice computations (5-07). For `SU(3)`, multiplying by $e^{\pm 2\pi i/3}$ at each piercing gives the string tension $1.5\rho_f$ for $k = 1, 2$ and 0 for the adjoint Computed .

For how this book uses this, see 25-03 § 03 (the identity by which fixing the frame makes the container signs independent).

07

The Gribov region and the horizon: from the gauge-fixed side

The fourth picture reads confinement from the shape of the gauge-fixed propagator. The Landau gauge surface $\partial_\mu A_\mu = 0$ can meet a single orbit more than once (4-08), so Gribov (1978) restricted the integral to the region Ω where the Faddeev–Popov operator $\mathcal{M}$ is positive. Its boundary is the Gribov horizon.

$$\mathcal{M}^{ab} = -\partial_\mu\left(\partial_\mu \delta^{ab} + g f^{acb} A^c_\mu\right), \qquad \Omega = \{A : \ \partial_\mu A^a_\mu = 0,\ \mathcal{M} > 0\} \quad (12)$$

"$\mathcal{M} > 0$" means that the eigenvalues of $\mathcal{M}$ (the spectrum of a self-adjoint operator, 1-10) are all positive. At $A = 0$, $\mathcal{M} = -\partial^2$, which is positive apart from constants. As A grows the smallest eigenvalue goes down, and where it reaches 0 is the horizon. On the horizon an infinitesimal gauge transformation in the direction of the zero eigenvalue preserves the gauge condition, so copies arbitrarily close appear.

What is certain as mathematics goes only as far as Dell'Antonio–Zwanziger (1991): "every orbit passes inside the horizon" Known . Copies remain inside Ω too, and no method is known for handling a copy-free region analytically (Zwanziger claims that the copies inside do not affect expectation values Physics).

08

The propagator and positivity

Zwanziger (1989) wrote the restriction to Ω as a local action. The propagator of its tree approximation (lowest order, without loop contributions, 4-03) is on the left below, and the version with condensates added (RGZ, Dudal et al. 2008) on the right Physics .

$$D_{\mathrm{GZ}}=\frac{p^2}{p^4+\lambda^4},\qquad D_{\mathrm{RGZ}}=\frac{p^2+M^2}{p^4+(M^2+m^2)\,p^2+\lambda^4+M^2m^2} \tag{13}$$

$\lambda^4=2g^2N\gamma^4$, and γ is fixed by the horizon condition. Here λ is a mass scale, different both from the penetration depth λ_L of § 03 and from the large-N $\lambda=g^2N$. M^2 and m^2 are mass parameters added in RGZ. The GZ form is 0 at $p=0$ and its poles are complex, $p^2=\pm i\lambda^2$, quite unlike $1/(p^2+m_0^2)$ for a free particle of mass m_0.

If the gluon were a physical particle, the propagator could be written as a superposition with positive density ρ (the Källén–Lehmann representation), and the correlation in the time direction $C(t)$ could not be negative (formulas (159)(160) of Dudal et al. 2008, 4-10). ω is the mass being superposed.

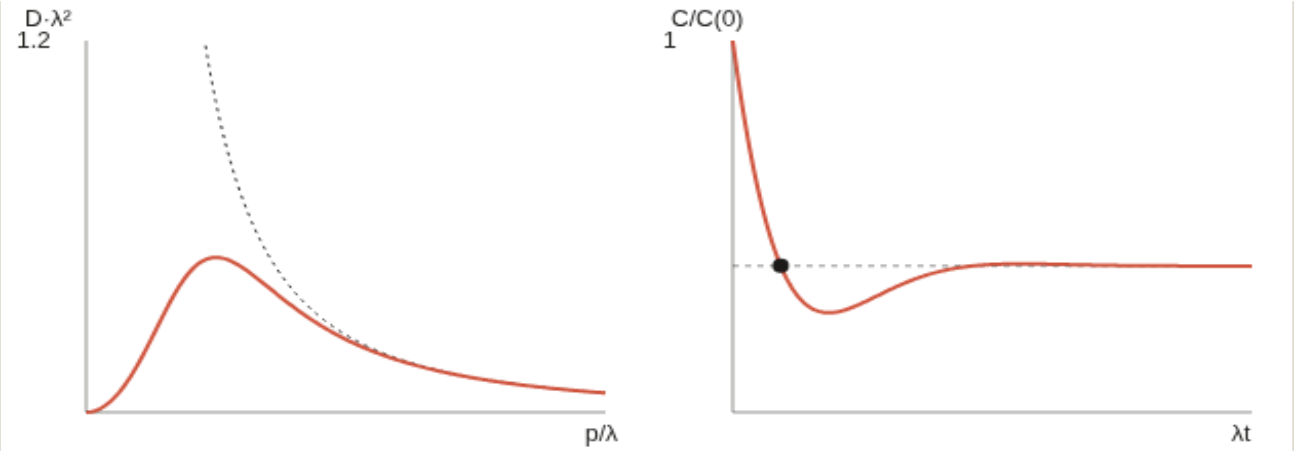
$$D(p^2)=\int_0^\infty d\omega^2\,\frac{\rho(\omega^2)}{p^2+\omega^2},\qquad C(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-ipt}D(p^2)\,dp=\int_0^\infty d\omega\,\rho(\omega^2)\,e^{-\omega t} \tag{14}$$

At $m=0$, computing by complex integration (residues) gives a closed form (it can also be checked by numerical integration; bottom of Figure 3). For $M^2<2\lambda^2$, setting $\cos\varphi=M^2/(2\lambda^2)$ Computed

$$C(t)=\frac{e^{-\lambda t\cos(\varphi/2)}}{2\lambda\sin\varphi}\,\sin\!\left(\frac{3\varphi}{2}+\lambda t\sin\frac{\varphi}{2}\right) \tag{15}$$

and it first becomes negative at $t_*=(\pi-3\varphi/2)/(\lambda\sin\frac{\varphi}{2})$. For GZ, $t_*=\pi/(2\sqrt{2}\,\lambda)\approx 1.111/\lambda$. For $M^2>2\lambda^2$ it splits into real poles, $D=A/(p^2+r_1)+B/(p^2+r_2)$ $(0<r_1<r_2)$, and although B is negative (ρ is not positive), $C(t)\geq e^{-\sqrt{r_2}\,t}C(0)>0$ and it does not become negative. This test works in one direction only.

Figure 3 Changing M^2/λ^2 (m = 0) changes $D(p^2)$ on the left (dotted line $1/p^2$) and $C(t)/C(0)$ on the right. 0 is GZ; the black dot is the first zero.



M²/λ² = 0.0 (GZ) : D(0)·λ² = 0.0₀ The poles are complex conjugates. Negative from t* = 1.111/λ.

Computed The curves are computed on the spot from formulas (13)(15) and the partial fractions for real roots. The closed formula was checked on this machine against numerical integration, with differences at most 2×10^{-11}.

Table 3 Tree approximation of RGZ at m = 0 (λ = 1)

M²	D(0)	Poles	C(0)	Zero t*
0 (GZ)	0	Complex conjugate	0.354	1.111
1.0	1.0	Complex conjugate	0.577	3.142
1.9	1.9	Complex conjugate	0.734	16.86
2.5	2.5	Real (B = −0.333)	0.825	No
4.0	4.0	Real (B = −0.077)	1.021	No

The lattice did not agree with the original prediction $D(0) = 0$. Cucchieri–Mendes (SU(2), 128^4, about $(27\ \text{fm})^4$; 1 fm = 10^{-15} m) report in their abstract that at zero momentum it "seems to show a finite non-zero value", and Bogolubsky et al. (SU(3)) point the same way Physics . RGZ is built to reproduce this Physics . BRST symmetry (the symmetry that remains after gauge fixing, 4-08) is softly broken in GZ. Soft breaking means that the breaking term has low dimension and has no effect at short distances. This breaking is not a price paid by RGZ but a property of the horizon term of GZ itself Physics .

Seam

The propagator depends on the gauge. The target of the problem statement is correlations of gauge-invariant local fields such as $\mathrm{tr}\, F_{\mu\nu}F_{\rho\sigma}$, and there is no theorem deriving their gap from the infrared shape of the propagator. Nor can a continuous global gauge fixing be made (Singer, 3-04).

09

Setting the four pictures side by side

Table 4 The leading role, support and what is missing in the four pictures

Picture	Leading role	Gauge dependence	Support	What mathematics can say	What
Dual superconductivity	Monopole condensation	Through the definition of monopoles	Numbers for the order parameter	Area law for 3-dimensional compact `U(1)`	Cond ⇒ ar gap
Abelian projection	MAG monopoles	Strong	Abelian dominance	— (Elitzur's theorem constrains it)	A def a uni deter gaug invar quan
Centre vortices	Vortices / their free energy	The former only	Centre dominance, vortex removal	Tomboulis–Yaffe, formula (4)	Vorte cond at we coup dime
GZ	Propagator	By definition	Lattice propagators	Orbits pass inside the horizon	A bri gaug invar corre

Every leading variable becomes visible only once a gauge is chosen, and the gauge-invariant core is only the vortex free energy and the inequalities around it. What every picture explains is σ; Δ is merely expected as a consequence (§ 02). None of the pictures yields "a single scale". The seams are sorted out in 16-05, obstacles other than divergence in 7-11.

What this chapter can say

What is said	What is not said
Known The area law at small β. On the lattice, a strong form of the gap ⇒ the area law	The area law in the continuum limit of the four-dimensional non-abelian theory. An implication between the Clay gap and confinement
Computed Agreement of the exact formula (11) for the vortex gas with MC, Creutz ratio $= 2\rho_f$	That this model describes the Yang–Mills vacuum (the distribution of vortices is assumed)
Computed The closed form (15) of $C(t)$ in the tree approximation of RGZ ($m = 0$)	Deriving the gap of gauge-invariant correlations from the propagator

What is said	What is not said
Physics Lattice numbers for monopole condensation, abelian dominance, centre dominance, vortex removal and $D(0) \not= 0$	Deciding which mechanism is right (the numbers depend on the gauge)

Sources and reproduction

Item	Kind	Source / tools
Formulas (1)(3)(6)(8)(10), N-ality, $(92 \pm 4)\%$, attribution of dual superconductivity	Review	Greensite, Prog. Part. Nucl. Phys. 51 (2003) 1, hep-lat/0301023 (text)
Wilson loops	Primary	Wilson, PRD 10 (1974) 2445 (abstract)
Confinement is a "natural extension" (§ 02)	Primary	Jaffe–Witten, the Clay problem statement (text)
Dual superconductivity	Primary	't Hooft, EPS 1975 talk (published 1976) (bibliographic data only). Mandelstam, Phys. Rep. 23 (1976) 245 (bibliographic data only). Nambu, PRD 10 (1974) 4262 (bibliographic data only)
3-dimensional compact `U(1)`	Primary	Polyakov, NPB 120 (1977) 429 (bibliographic data only). Göpfert–Mack, CMP 82 (1982) 545 (abstract; the description of the continuum limit checked in the INSPIRE abstract, text not read; the year is the Crossref publication year, INSPIRE gives 1981)
Lattice numbers for monopole condensation, SW	Primary	Di Giacomo et al., PRD 61 (2000) 034503 (abstract). Seiberg–Witten, NPB 426 (1994) 19 (abstract)
Abelian projection, MAG, abelian dominance	Primary	't Hooft, NPB 190 (1981) 455 (abstract). Kronfeld–Schierholz–Wiese, NPB 293 (1987) 461 (bibliographic data only). Kronfeld et al., PLB 198 (1987) 516 (bibliographic data only). Suzuki–Yotsuyanagi, PRD 42 (1990) 4257 (bibliographic data only). Bali et al., PRD 54 (1996) 2863 (bibliographic data only)
Elitzur's theorem	Primary	Elitzur, PRD 12 (1975) 3978 (bibliographic data only)
Centre vortices	Primary	Del Debbio et al., PRD 55 (1997) 2298 (abstract). de Forcrand–D'Elia, PRL 82 (1999) 4582 (abstract). Tomboulis, arXiv:0707.2179 (abstract). Ito–Seiler, arXiv:0711.4930 (abstract). Tomboulis–Yaffe, CMP 100 (1985) 313 (bibliographic data only)

Item	Kind	Source / tools
Formula (4), Table 1	Primary	Chatterjee, arXiv:2006.16229 (text; Definition 2.1, Theorems 2.2 and 2.4, § 2.4). Chatterjee, arXiv:1803.01950 (text). Fröhlich–Spencer, CMP 83 (1982) 411 (bibliographic data only). Adhikari–Cao, arXiv:2202.10375 (abstract). Osterwalder–Seiler, Ann. Phys. 110 (1978) 440 (bibliographic data only). Shen–Zhu–Zhu, arXiv:2204.12737 (abstract)
Formulas (12)–(14)	Primary	Dudal et al., PRD 78 (2008) 065047, arXiv:0806.4348 (text; formulas (2)(3)(58)(59)(159)(160)(207)). Gribov, NPB 139 (1978) 1 (bibliographic data only). Zwanziger, NPB 323 (1989) 513 (bibliographic data only)
The horizon, copies inside, lattice propagators	Primary	Dell'Antonio–Zwanziger, CMP 138 (1991) 291 (bibliographic data only). Zwanziger, PRD 69 (2004) 016002 (abstract). Cucchieri–Mendes, arXiv:0710.0412, PoS LAT2007:297 (abstract). Bogolubsky et al., arXiv:0901.0736 (abstract)
Formulas (11)(15), Tables 2 and 3	Computed	python3 (numpy, scipy): direction average of the area of symmetric differences and MC of one million samples (with standard errors) / checking the closed form against numerical integration, the minimum of $C(t)/C(0)$

Next chapter to read: 17-03 The string picture — treating the squeezed tube as a single string and looking at its fluctuations.

The string picture — treating the tube of confinement as a single string, a universal correction and a shape come out of its fluctuations

> 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 follow how the Lüscher term $-\pi(d-2)/(24R)$ comes out of the zero-point energy of the transverse fluctuations of the tube, and why its coefficient does not depend on the details of the theory. With formulas, moving figures and computations we look at the thickness of the tube, the Nambu–Goto levels, the "rigid side" and the "rough side" of the lattice string, and the shape selected by a tension that depends on direction (the Wulff shape).

Prerequisite chapters — 17-01 How physics sees it (Wilson loops, the area law, the string tension σ). The zero-point energy (2-06) and the strong-coupling expansion (5-02) are written on the spot as far as needed.

Said first — This is a chapter at the level of physics Physics . The string picture takes as its starting point that a long tube exists (confinement), and in four-dimensional non-abelian Yang–Mills that is not a theorem. This book does not itself show confinement, the mass gap or the construction of the continuum limit (the published construction of the continuum limit is read in Volume 2).

The order of this chapter

1. Treating the tube as a single string — the weight of the area
2. The Nambu–Goto action — expanding the fluctuations
3. The zero-point energy and the Lüscher term — open strings and closed strings
4. Detour: 1+2+3+… and −1/12 — the part that does not depend on how we cut off
5. The tube thickens logarithmically — a fluctuating string
6. The Nambu–Goto levels and short tubes — the range of universality
7. The lattice string is rigid, and roughens somewhere — the side with theorems
8. The Wulff shape — a tension depending on direction, and the shape

9. The droplet of the two-dimensional Ising model — an exactly known tension, and from a square to a circle
10. What the string picture hands to mathematics, and what it does not — what is fixed under an assumption, and the assumption itself

---

01

Treating the tube as a single string

The flux tube of 17-01 § 05, if its thickness is much smaller than the distance R of the pair, is from afar a single line. Extended in the direction of Euclidean time, the line sweeps out a surface (the **world sheet**). The string picture is an effective theory that describes long distances by the fluctuations of this surface Physics . To a surface S bounded by the rectangle $R \times T$ of the Wilson loop we give a weight fixed by its area $A(S)$.

$$\langle W_{R,T} \rangle \approx \sum_{S:\ \partial S = \text{rectangle}} e^{-\sigma A(S)} \quad (1)$$

With the minimal surface alone this is $e^{-\sigma RT}$, the area law. The question is what the fluctuations add. As a metaphor, it is the trembling of a stretched membrane. More precisely, (1) is not a formula derived from Yang–Mills but a starting point under the assumption that "a long tube exists, and its low excitations are only transverse fluctuations" (Athenodorou–Bringoltz–Teper 2011 § 2; ABT below).

At strong coupling on the lattice (5-02) a formula close to (1) holds, and the weight u of one plaquette plays the role of $e^{-\sigma 0}$ Known . But the area law also comes out for `U(1)`, so it is no evidence of non-abelian confinement (Chatterjee 2018 § 4, 5-06).

02

The Nambu–Goto action

An action proportional to the area is called the **Nambu–Goto action** (the name is said to be after Nambu and Goto). We describe the surface by coordinates t, x and a perpendicular displacement $h(t,x) \in \mathbb{R}^{d-2}$ (static gauge; d is the dimension of spacetime). The area of a surface $X(t,x)$ is given by the surface-area formula learnt in multivariable calculus, $\iint \sqrt{|X_t|^2 |X_x|^2 - (X_t \cdot X_x)^2}\, dt\, dx$ ($X_t = \partial_t X$, $X_x = \partial_x X$). Putting in $X = (t, x, h)$ gives the following.

$$A = \int_0^T dt \int_0^R dx \, \sqrt{1 + |\partial_t h|^2 + |\partial_x h|^2 + |\partial_t h|^2 |\partial_x h|^2 - (\partial_t h \cdot \partial_x h)^2} \tag{2}$$

If the fluctuations are gentle, expanding with $\sqrt{1+z} \approx 1 + z/2$ gives

$$\sigma A = \sigma R T + \frac{\sigma}{2} \iint \left(|\partial_t h|^2 + |\partial_x h|^2 \right) dt \, dx + O\big((\partial h)^4\big) \tag{3}$$

The second term is a massless free field with $d-2$ components, of the same form as the vibrating string of 2-02 (with $h = 0$ at the ends). For a long string the higher terms are suppressed by powers of $1/(\sigma R^2)$ (ABT § 2.1).

Example: the increase in length of a ripple $h = A_0 \sin(\pi x / R)$ is $\pi^2 A_0^2 / (4R)$ in the quadratic approximation. With $R = 1$ fm and $A_0 = 0.2$ fm this is 0.099 fm (0.092 fm exactly). Multiplying by $\sigma = 1192$ MeV/fm ($\sqrt{\sigma} = 485$ MeV, 17-01 § 07) gives about 110 MeV Computed .

03

The zero-point energy and the Lüscher term

The second term of (3) is the action of a Euclideanised free field. When T is long, $\langle W_{R,T} \rangle \approx e^{-V(R) T}$, so the ground-state energy of the Hamiltonian of the same field gives $V(R)$ (4-02). The modes of this field are $k_n = n\pi/R$ ($n = 1, 2, \ldots$). Each mode is a harmonic oscillator of frequency k_n, and even in its lowest state it has the zero-point energy $k_n/2$ (2-06; $\hbar = c = 1$). Summing over the $d-2$ components gives the following.

$$E_0(R) = \sigma R + \frac{d-2}{2} \sum_{n=1}^{\infty} \frac{n\pi}{R} \tag{4}$$

The sum diverges. Fluctuations finer than a length a are outside the effective theory, so we cut off with $e^{-k_n a}$.

$$\sum_{n \geq 1} n \, e^{-\varepsilon n} = \frac{e^{-\varepsilon}}{(1 - e^{-\varepsilon})^2} = \frac{1}{\varepsilon^2} - \frac{1}{12} + \frac{\varepsilon^2}{240} - \cdots, \qquad \varepsilon = \frac{\pi a}{R} \tag{5}$$

Putting (5) into (4), the first term is $(d-2) R / (2\pi a^2)$, proportional to R, and is absorbed into the tension. What remains is the static potential.

$$V(R) = \sigma R + c_0 - \frac{\pi (d-2)}{24 R} + \cdots \tag{6}$$

c_0 is a non-universal constant. At $d = 4$ this is $-\pi/(12R)$, the **Lüscher term** of 17-01 formula (6) (Lüscher–Symanzik–Weisz 1980, Lüscher 1981). It is a coefficient derived

within the effective theory of long tubes, not a mathematical theorem Physics . The $1/R$ term is a finite part that does not diverge, so its coefficient does not depend on a. With $R = 1$ and $a = 0.2,\ 0.1,\ 0.05$, the remainder is $-0.2567,\ -0.2605,\ -0.2615$, approaching $-\pi/12$ Computed .

Closed strings and the Casimir energy of a cylinder

For a tube wound around a direction of period L (flux going round a periodic boundary), $k_n = 2\pi n/L$ occurs in both directions, positive and negative, and the same computation gives the following (ABT formula (9)).

$$E_0(L) = \sigma L - \frac{\pi(d-2)}{6L} + \cdots \quad (7)$$

This is the Casimir energy of a free field on a cylinder. Lucini–Teper 2001 extract the lattice σ with formula (4) of their paper, $am_l = a^2\sigma L - \pi/(3L)$, and note: 'in using eqn(4) we are assuming that we have linear confinement' (5-08). A rough guide: at $R = 1$ fm, $\pi/(12R) \approx 52$ MeV is 4.3% of σR, and at $L\sqrt{\sigma} = 3$ (the lower limit of Lucini–Teper; about 1.22 fm converting with $\sqrt{\sigma} = 485$ MeV, while Lucini–Teper themselves write $\simeq$1.3 fm) the correction in (7) is 11.6% of the leading term Computed .

04

Detour: 1+2+3+… and −1/12

A section you may skip.

The sum in the slogan "$1 + 2 + 3 + \cdots = -1/12$" diverges. The precise content is the following formula. For a smooth cutoff f ($f(0) = 1$, decreasing fast), as $\varepsilon \to 0$,

$$\sum_{n \geq 1} n\, f(\varepsilon n) = \frac{C_f}{\varepsilon^2} - \frac{1}{12} + O(\varepsilon^2), \qquad C_f = \int_0^\infty x f(x)\, dx \quad (8)$$

The coefficient C_f of the divergent term changes with the cutoff, and $-1/12$ does not. The reason is the **Euler–Maclaurin formula**. It writes the difference between a sum and an integral in terms of derivatives at the end point, and when g and its derivatives decrease fast at infinity it takes the following form ($B_2 = 1/6$ is a Bernoulli number).

$$\sum_{n \geq 0} g(n) = \int_0^\infty g(x)\, dx + \tfrac{1}{2} g(0) - \tfrac{B_2}{2!}\, g'(0) - \cdots$$

For $g(x) = x f(\varepsilon x)$ we have $g(0) = 0$, $g'(0) = f(0) = 1$, and the integral is C_f/ε^2, so the constant term is $-B_2/2 = -1/12$. The remaining terms are powers ε^2 and higher. The analytic continuation of the zeta function, $\zeta(-1) = -1/12$, is the same number Known .

Computed Table 1 The sum minus C_f/ε^2 (exponential $f=e^{-x}$, $C_f=1$; Gaussian $f=e^{-x^2}$, $C_f=1/2$)

ε	Exponential	Gaussian
0.1	−0.0832917	−0.0834169
0.01	−0.0833329	−0.0833342
→ 0	−0.0833333	−0.0833333

The part that changes with the cutoff is absorbed into the tension, and only $-1/12$ remains in the coefficient of $1/R$ (for divergent series see also 4-12, 7-12).

05

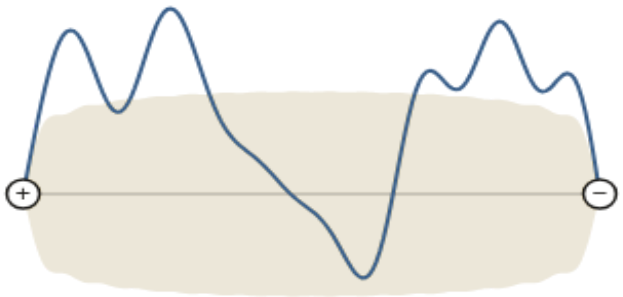
The tube thickens logarithmically

The modes q_n of (3) are oscillators with mass σ and frequency k_n, and the variance in the ground state is $1/(2\sigma k_n)$. At the midpoint of $h=\sum_n \sqrt{2/R}\, q_n \sin k_n x$ only odd n survive, and $\langle h^2\rangle = \sum_{n\ \mathrm{odd}} 1/(\sigma\pi n)$. Summing up to $n \lesssim R/a$ gives a logarithm (Lüscher–Münster–Weisz 1981; Gliozzi–Pepe–Wiese 2010 formula (3)). It is a formula derived within the effective theory of long tubes Physics .

$$w^2\left(\frac{R}{2}\right) = \left\langle |h(R/2)|^2 \right\rangle \approx \frac{d-2}{2\pi\sigma}\ln\frac{R}{R_0} \quad (9)$$

What is universal is the coefficient of $\ln R$; R_0 is not universal. The thickness grows without bound, but $w/R \to 0$, so from afar it stays a thin line.

Figure 1 Changing the distance R with the knob changes the number of modes on the string and the width of the fluctuation (the band). "Play" makes each mode oscillate with its zero-point size



1/√σ

R√σ = 4.00 (1.63 fm), modes: 16, midpoint w²σ = 0.644 (ln N+γ+ln2)/(2π) = 0.643

Computed The numbers in this figure were computed on this machine. Cutoff $a = 0.25/\sqrt{\sigma}$, the band is $\pm w$, one transverse component. The oscillation is a classical picture, not the quantum fluctuation itself.

Table 2 Variance of the free field at the midpoint of a lattice strip (width R, h=0 at both ends, infinite time)

(one component, in units of $1/\sigma$)

R (lattice spacings)	$\sigma\langle h^2\rangle$	Increase when R is doubled
8	0.516070	0.111626
1024	1.288650	0.110318
Prediction of (9)	–	ln2/(2π) = 0.110318

Computed Summed exactly with the closed form $1/\sqrt{m^2(m^2+4)}$ for each mode. The increase agrees with (9) to 6 digits.

At $d=4$ the coefficient $1/(\pi\sigma)$ is 0.053 fm^2, and doubling R increases w^2 by only 0.037 fm^2 Computed . The logarithmic growth has been confirmed on the lattice for `SU(2)` in (2+1) dimensions (Gliozzi–Pepe–Wiese 2010) Physics .

06

The Nambu–Goto levels and short tubes

The levels of a closed string with no longitudinal momentum can be written in closed form in Nambu–Goto (with the sums of the left- and right-moving excitations both equal to n; ABT formula (21)). It is a formula that comes from naive light-cone quantisation (Aharony–Komargodski cite, as the source of this method, Arvis 1983, which treated a string with fixed ends).

$$E_n(L)=\sqrt{\sigma^2L^2+8\pi\sigma\left(n-\frac{d-2}{24}\right)} \quad (10)$$

Expanding gives $\sigma L+\frac{4\pi}{L}\left(n-\frac{d-2}{24}\right)+\cdots$, and $n=0$ returns to (7). The ground state becomes imaginary at $\sigma L^2<\pi(d-2)/3$, which in $d=4$ is $L_H\sqrt{\sigma}=\sqrt{2\pi/3}=1.447$. If the periodic direction is regarded as time, $1/L$ is the temperature (4-02). The number of string states grows exponentially with energy, so above some temperature the sum of the partition function stops converging. This point is called the **Hagedorn transition** (for the growth of the states, see the detour of 17-05) Physics . ABT write that L_H can be regarded as 'the Hagedorn/deconfining transition in the Nambu-Goto model', and go on to say that pure gauge theories at large N do not reach this length. As the length shrinks, a first-order deconfinement transition occurs before it, at $l_c\sqrt{\sigma}=\sqrt{\sigma}/T_c\simeq 1.65$ (the l_c below Table 3).

Table 3 Energies of the closed string ($d=4$, in units of $\sqrt{\sigma}$)

L√σ	σL−π/(3L)	Ground state of (10)	Error of the approximation	n=1 of (10)	E_1-E_0
1.6	0.9455	0.6824	+38.6%	5.0595	4.377
3.0	2.6509	2.6279	+0.88%	5.6602	3.032
6.0	5.8255	5.8229	+0.04%	7.6836	1.861

Computed By formula (10). The length $l_c\sqrt{\sigma}$ at which the tube disappears on the lattice is the product of $a\sqrt{\sigma}$ and l_c/a in Table 3 of ABT: SU(3) 1.544, 1.559, SU(5) 1.636, SU(6) 1.676. All are slightly longer than L_H, and the two at larger N lie on either side of ABT's 1.65.

How far does universality go? According to Aharony–Komargodski 2013 'all the terms up to (and including) order 1/L^5 are universal': the ground state agrees with (10) up to $1/L^5$, the excitations for $d > 3$ deviate universally at $1/L^5$, and theory-specific terms enter from $1/L^7$. These are results derived within the effective theory of long tubes Physics . Of the quantisation in (10) they also write 'There is no rigorous derivation'. ABT measured about 20 levels of closed tubes on the lattice for four-dimensional SU(N), and reported that many are well described by (10) down to short lengths where the $1/L$ expansion does not converge. Some states with $J^P = 0^-$ (J is the angular momentum about the axis of the tube, P the sign under reflection) deviate strongly Physics .

In the last column of Table 3, the first excitation $3.03\sqrt{\sigma}$ of the string at $L\sqrt{\sigma} = 3$ is of the same order as the 0^{++} glueball of SU(3), $3.405\sqrt{\sigma}$ (17-01 Table 2) Computed (this does not imply a correspondence). Turning "there is a single scale" into a mathematical sentence is the subject of 17-12.

07

The lattice string is rigid, and roughens somewhere

On the strong-coupling side of the lattice, shifting part of the surface by one lattice step adds plaquettes, each carrying a small weight u. Transverse fluctuations are suppressed exponentially, and the surface stays **rigid** (flat). In the world of (9) the variance diverges logarithmically, and the surface is **rough**.

For crystal surfaces there is a **roughening transition** between the two. Itzykson–Peskin–Zuber 1980 argued that in four-dimensional lattice gauge theory too, roughening occurs at a finite coupling without any phase transition of the vacuum, and that $\sigma(\beta)$ has an essential singularity there. An essential singularity here is a singularity of the kind where the function and all its derivatives stay finite, but the Taylor expansion no longer represents the function at that point. An example is the function set to e^{-1/x^2} for $x > 0$ and 0 for $x \le 0$; its Taylor series at $x = 0$ has all coefficients 0. However far the series is extended, this point cannot be seen. IPZ write 'The roughening transition is a natural barrier to the extrapolation of strong-coupling series for the string tension'. For SU(2), β_r is about 2 (with a large error), and the authors note that their four-dimensional argument is 'for the most part, intuitive' Physics .

The three-dimensional Z(2) gauge theory is **dual** to the three-dimensional Ising model. Dual means that under a change of variables the partition functions of the two

models agree (up to simple factors) (the detour of 2-09; IPZ cite Wegner 1971). Under this correspondence the surface of a Wilson loop becomes an interface of the Ising model (IPZ formula (8)) Known . Measuring the gauge coupling by $t = \tanh \beta$ (in terms of the temperature of the dual Ising model, $t = e^{-2J/T}$; IPZ formula (7)), the phase transition is at $t_c = 0.6418$. The roughening point, by the series estimate IPZ cite, is $t_r \approx 0.46$, before the phase transition Physics . The effective string theory applies only on the rough side (Gliozzi–Pepe–Wiese 2010) Physics .

Key point

Physics The theorem on the area law at strong coupling (Osterwalder–Seiler 1978, 5-04) is about the rigid side; the Lüscher term and the logarithmic thickness are about the rough side. Physics holds that between them lies a singularity of $\sigma(\beta)$ beyond which the strong-coupling series cannot be extended.

08

The Wulff shape

On the rigid side of § 07, the tension also depends on direction. The edge of a Wilson loop at angle θ to the lattice axes becomes a staircase, and the minimal surface has $RT(|\cos\theta| + |\sin\theta|)$ plaquettes. At leading order in strong coupling this gives the following.

$$\sigma(\theta) = \sigma_0 \big(|\cos\theta| + |\sin\theta| \big) \quad (11)$$

At 45° it is $\sqrt{2}$ times larger Computed . In the continuum limit rotation symmetry is expected to return and the tension to become constant Physics .

What shape does a region of fixed area (a two-dimensional droplet) take when it is enclosed by a line whose tension depends on direction? With the tension $\tau(\mathbf{n})$ for each normal $\mathbf{n}$, we minimise $\oint \tau \, ds$ with the enclosed area fixed. The answer is given by the construction of Wulff 1901 (bibliographic data for the proof that the construction gives the minimum is in the table of sources) Known .

$$\mathcal{W}_\tau = \bigcap_{|\mathbf{n}|=1} \{ \mathbf{x} : \ \mathbf{x}\cdot\mathbf{n} \le \tau(\mathbf{n}) \}, \qquad \min_{\text{area}=A} \oint \tau \, ds = 2\sqrt{|\mathcal{W}_\tau| \, A} \quad (12)$$

If the tension is constant the answer is a circle; for $\tau = |n_1| + |n_2|$ of (11) it is the square $\{\max(|x|,|y|) \le 1\}$. The total at area 1 is 4 for the square and $8/\sqrt{\pi} = 4.51$ for a circle Computed .

09

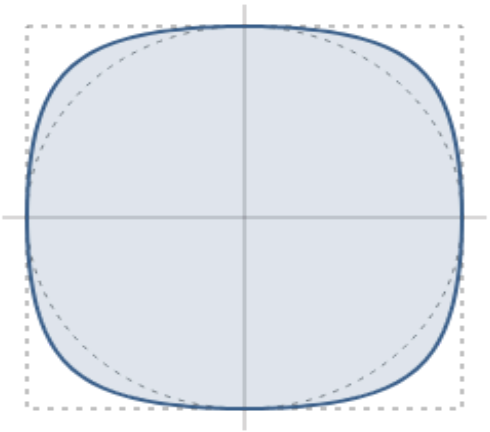
The droplet of the two-dimensional Ising model

An example where the direction-dependent tension is known exactly is the Ising model on the square lattice (coupling J, 2-09). The $\beta = 1/T$ of this section is the inverse temperature ($k_B = 1$; not the β of lattice gauge theory). The equilibrium shape of an island of the minority (a droplet) in the low-temperature phase is the following curve (the first exact analysis is attributed to Rottman–Wortis 1981) Known .

$$\cosh(\beta x) + \cosh(\beta y) = \sinh(2\beta J) + \frac{1}{\sinh(2\beta J)} \quad (13)$$

(13) is the boundary of $\mathcal{W}_\tau$ in (12) itself. Its size too is as in (12): the extent $\max \mathbf{x}\cdot\mathbf{n}$ in the direction $\mathbf{n}$ equals the tension $\tau(\mathbf{n})$ (in units of J). So x, y are not positions but coordinates measured in units of energy. The intercept with the x axis (the tension in the axis direction) agrees to 5 digits with $2J + \beta^{-1}\ln\tanh\beta J$ from Onsager's exact solution Computed . As $T \to 0$, (13) becomes a square, and the tension is the same staircase count as in (11). As $T \to T_c$ the right-hand side approaches 2, and for $u = \beta x$, $v = \beta y$ we have $\cosh u + \cosh v \approx 2 + (u^2 + v^2)/2$, so it becomes a circle.

Figure 2 Changing the temperature T/T_c with the knob changes the equilibrium shape of the two-dimensional Ising droplet (filled) from a square to a circle. The tension along the axis is normalised to 1; the dotted lines are the square and the circle



```
T/T_c = 0.30  τ(0°) = 1.9278 J (Onsager's formula 1.9278)  τ(45°)/τ(0°) = 1.1196 (square 1.4142, circle 1)
```

Table 4 Equilibrium shapes of the two-dimensional Ising model. The isoperimetric ratio is $4\pi\,\text{area}/\text{perimeter}^2$; the last column is the ratio of the total tension to that of a circular droplet of the same area

T/T_c	τ(45°)/τ(0°)	Isoperimetric ratio	Wulff/circle
→ 0	1.4142	0.7854	0.8862
0.2	1.1984	0.9406	0.9698
0.4	1.0690	0.9917	0.9958

T/T_c	τ(45°)/τ(0°)	Isoperimetric ratio	Wulff/circle
0.8	1.0035	1.0000	1.0000

Computed The numbers in this figure and table were computed on this machine ((13) solved by bisection for each angle).

At low temperature the diagonal tension falls first. A stepped edge has a number of arrangements (entropy), and the corners round off from there. Approaching T_c the correlation length becomes much longer than the lattice spacing, the orientation of the lattice is forgotten, and the shape becomes a circle. This is how the expectation that rotation symmetry returns in the continuum limit (5-05) looks in an exactly tractable model.

That the droplet takes the Wulff shape in the low-temperature two-dimensional Ising model has been shown from the microscopic model (Dobrushin–Kotecký–Shlosman 1992, 1993) Known .

The Wulff problem applies directly to droplets of a two-dimensional phase with fixed area like this one. A closed ring of flux in Yang–Mills has no constraint fixing its area, and left alone it shrinks. Still, a closed tube on the lattice must turn corners, so its energy depends on its shape through $\sigma(\theta)$ Physics . How it is used in the record of this book's study is in 23-02.

10

What the string picture hands to mathematics, and what it does not

What it hands over

Physics If a long tube is assumed to exist, the Lüscher coefficient, the coefficient of $\ln R$ in the thickness, and the form of the levels up to $1/L^5$ are fixed independently of the details.

Known For the low-temperature two-dimensional Ising model, the relation between the direction-dependent tension and the shape of the droplet is a theorem.

What it does not hand over

That a tube exists (the area law at large β) 'remains elusive' for the four-dimensional non-abelian theory (Chatterjee 2018). The lattice σ is extracted by subtracting the string correction, and Lucini–Teper, fitting only two points, note: 'we cannot claim to have evidence for both the 1/L functional form and for its coefficient'.

The string picture is about massless transverse fluctuations, and the glueball masses do not come out of it directly. ABT examine the possibility that the $J^P = 0^-$ levels that deviate from (10) represent 'the massive modes associated with the internal structure

of the flux tube' Physics . Since the only length scale is $1/\sqrt{\sigma}$, $m \sim \sqrt{\sigma}$ looks natural, but it is not derived (17-01 § 09, 17-12). The view as a sum over surfaces appears in the next chapter, 17-04, as the $1/N$ expansion.

What this chapter can say

What is said	What is not said
Physics The universality, derived within the effective theory that assumes a long tube exists, of the Lüscher coefficient, the coefficient of $\ln R$ in the thickness, and the levels up to $1/L^5$	That a long tube exists in four-dimensional non-abelian Yang–Mills (the area law at large β)
Computed That differences of regularisation enter only the term proportional to R (Table 1). The increase in thickness of the lattice free field (Table 2)	A theorem on the thickness of the Yang–Mills tube
Computed The difference between (10) and the Lüscher approximation, L_H, l_c (Table 3)	That (10) is exact for Yang–Mills
Known The Wulff construction and its justification for the low-temperature two-dimensional Ising model. Computed Table 4	That $\sigma(\theta)$ becomes constant in the four-dimensional continuum limit
Known The area law at strong coupling (Osterwalder–Seiler). The duality between the three-dimensional `Z(2)` gauge theory and the Ising model. $-1/12$ by Euler–Maclaurin	Extending the strong-coupling series beyond the roughening point
Physics The roughening transition of four-dimensional lattice gauge theory (the argument of IPZ), and the series estimate of t_r for three-dimensional `Z(2)`	A proof of the existence and location of the roughening transition

Sources and reproduction

Item	Kind	Source / tools
§ 01, (7)(10), l_c, the Hagedorn sentence and $l_c\sqrt{\sigma} \simeq 1.65$, $J^P = 0^-$	Primary	Athenodorou–Bringoltz–Teper, JHEP 1102 (2011) 030, arXiv:1007.4720 (text)
The order of universality, the caveat on (10), that (10) comes from naive light-cone quantisation	Primary	Aharony–Komargodski, JHEP 05 (2013) 118, arXiv:1302.6257 (text)

Item	Kind	Source / tools
The original sources of the Lüscher term and (10)	Primary	Lüscher–Symanzik–Weisz, NPB 173 (1980) 365 (bibliographic data only). Lüscher, NPB 180 (1981) 317 (bibliographic data only). Arvis, PLB 127 (1983) 106 "The Exact q Anti-q Potential In Nambu String Theory" (bibliographic data only; taken from reference [19] of Aharony–Komargodski). The original source of the name Nambu–Goto not checked
Formula (9), that effective string theory applies on the rough side	Primary	Lüscher–Münster–Weisz, NPB 180 (1981) 1 (bibliographic data only). Gliozzi–Pepe–Wiese, PRL 104 (2010) 232001, arXiv:1002.4888 (text)
Extraction of the lattice σ, $L\sqrt{\sigma} \geq 3$	Primary	Lucini–Teper, JHEP 0106 (2001) 050, hep-lat/0103027 § 4.1 (text). The original reads 'r $\geq 3/\sqrt{\sigma} \simeq 1.3$ fm'. The 1.22 fm in the text is the conversion with $\sqrt{\sigma} = 485$ MeV (the same as 17-01)
The roughening transition, $t = \tanh\beta$, t_c, t_r, duality	Primary	Itzykson–Peskin–Zuber, PLB 95 (1980) 259 (text). $t = \tanh\beta = e^{-2/T}$ is formula (7) of that paper, $t_c = 0.6418$ is in its text. $t_r \approx 0.46$ is the value $t_r \approx 0.4593$ from the series analysis of Weeks et al. that IPZ cite at formula (13) (an estimate as of 1980; the original sources of Weeks et al. are references [7][9] of IPZ, one an unpublished communication; not checked). In IPZ's own quantity, formula (19) gives $t_r \approx 0.47$. Duality: Wegner, J. Math. Phys. 12 (1971) 2259 (bibliographic data only; from the reference list of IPZ)
The status of the area law	Primary	Chatterjee arXiv:1803.01950 § 4 (text). Osterwalder–Seiler, Ann. Phys. 110 (1978) 440 (bibliographic data only)
Formula (12), the microscopic justification of the Wulff shape	Primary	Wulff, Z. Kristallogr. 34 (1901) 449–530 (bibliographic data only; taken from the reference list of Fujimoto–Otsuka). The original source of the proof that the construction gives the minimum not checked. Explanation of the construction: Günther–Rikvold–Novotny, Physica A 212 (1994) 194, cond-mat/9406114 (text), and the review cited there, Rottman–Wortis, Phys. Rep. 103 (1984) 59 (bibliographic data only). Dobrushin–Kotecký–Shlosman, Wulff Construction: A Global Shape from Local Interaction, AMS Transl. Math. Monographs 104 (1992). The same authors' short paper A microscopic justification of the Wulff construction, J. Stat. Phys. 72 (1993) 1–14 (both bibliographic data only; checked on Crossref)
Formula (13)	Primary	Rottman–Wortis, PRB 24 (1981) 6274 (bibliographic data only; the attribution "the first exact analysis" is from the introduction of Fujimoto–Otsuka arXiv:1906.01925). Also Avron–van Beijeren–Schulman–Zia, J. Phys. A 15 (1982) L81, Zia–Avron, PRB 25 (1982) 2042 (bibliographic data only; listed by the same introduction as 'see also'). The shape checked against Onsager's tension
Tables 1–4, Figures 1 and 2, the Computed items in the text	Computed	python3 (standard library): regularised sums / the lattice free field / (10) and Table 3 of ABT / bisection for (13)

Next chapter to read: 17-04 Large N — a view that organises the sum over surfaces by $1/N$.

Large N — sending the number of colours to infinity sorts Feynman diagrams by surfaces

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 as N grows, diagrams are sorted by genus into powers of N, and only planar diagrams survive ('t Hooft 1974). That the sum of diagrams becomes a sum over random surfaces, and fluctuations disappear. A phase transition that appears only at N=∞ on the two-dimensional lattice, and the N-dependence on the four-dimensional lattice.

Prerequisite chapters — 4-03 Perturbation theory and Feynman diagrams. 3-02, 7-02 and 17-01 are written on the spot as far as needed.

Said first — Large N is an expansion at the level of physics. The prize problem requires finite N, and no road back from N→∞ is known. Mathematically established results are limited to the strong-coupling lattice or to two dimensions. This book does not itself show either the mass gap or the construction of the continuum limit (the published construction of the continuum limit is read in Volume 2).

The order of this chapter

1. The idea of varying N — 1/N in place of a small number
2. The double-line notation — a line is two indices
3. 't Hooft's counting — the power of N is fixed by the genus
4. The genus expansion of a Gaussian matrix — actually counting (Detour: Catalan numbers and the semicircle)
5. Random surfaces — towards the string picture
6. The master field — fluctuations disappear as 1/N
7. The Gross–Witten transition — a kink only at N=∞
8. The four-dimensional lattice — ratios move slowly with N
9. What can be said, and what cannot — the shores of strong coupling and two dimensions

01

The idea of varying N

Four-dimensional pure Yang–Mills has no small number to expand in. The mass scale Λ has the form $e^{-1/(2b_0 g^2)}$ ($b_0 = 11N/(48\pi^2)$, formula (1) of 17-01) and shows up at no power of g (17-01 § 03). 't Hooft (1974) treated the size N of the group as a knob and considered the following limit (the group in the original paper is U(N)) Known .

$$N \to \infty, \qquad \lambda = g^2 N \ \text{ fixed} \quad (1)$$

According to the abstract, 'only planar diagrams with the quarks at the edges dominate', and 1/N corresponds to the coupling of the string models of the time. As a metaphor, the idea is to see $N = 3$ as "$N = \infty$ with a correction of order $1/9$"; more precisely, how small the correction is has to be checked with lattice numbers (§ 08). The problem statement calls the duality conjecture, that through the 1/N expansion a gauge theory might be equivalent to a string theory with coupling $1/N$, 'A prominent speculation', and writes that such a description could be 'perhaps a good starting point for a rigorous proof (for sufficiently large N)' Known . But the requirement is finite N (0-03).

02

The gauge field is a matrix: the double-line notation

The gauge field A_μ is an $N \times N$ Hermitian matrix at each point (3-02), and we regard the two indices of a component ${A_\mu}^i{}_j$ as the two ends of colour lines. The action is (Euclidean)

$$S = \frac{1}{2g^2}\int \operatorname{tr} F_{\mu\nu}F_{\mu\nu}\, d^4x = \frac{N}{2\lambda}\int \operatorname{tr} F_{\mu\nu}F_{\mu\nu}\, d^4x, \qquad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - i[A_\mu, A_\nu].$$

The trace of a product joins indices into a ring, as in $\operatorname{tr}(ABC) = A^i{}_j B^j{}_k C^k{}_i$, and the three-point and four-point vertices of (2) are all of this form. The colour part of the propagator (4-03) for U(N) is

$$\langle A^i{}_j\, A^k{}_l \rangle \ \propto\ \frac{\lambda}{N}\,\delta^i{}_l\,\delta^k{}_j. \quad (3)$$

One delta is one index line, so a gluon line is two lines of opposite orientation, a thin ribbon. This is the **double-line notation**. In the continuum action (at the level of the Lie algebra) the direction of the identity matrix drops out of the commutators, and its cross terms with the traceless part vanish too, so the U(N) of pure gauge theory is

SU(N) plus one **free U(1) field that does not interact** Known . As groups, U(N) does not split into SU(N)×U(1) (the identity matrix times $e^{2\pi ik/N}$ lies in both), and on the lattice U(N) (§ 07) this separation does not hold.

03

't Hooft's counting: the power of N is fixed by the genus of a surface

From (2)(3), each vertex carries N/λ, each propagator λ/N, and each closed index ring $\sum_i \delta^i{}_i = N$. A vacuum diagram with V vertices, E propagators and F index rings is proportional to

$$\left(\frac{N}{\lambda}\right)^{V}\left(\frac{\lambda}{N}\right)^{E} N^{F} = N^{\,V-E+F}\,\lambda^{\,E-V} \tag{4}$$

Gluing a disc to each index ring gives an oriented closed surface, and $V-E+F$ becomes the **Euler characteristic** $\chi = 2-2h$, fixed by the surface alone (h is the number of handles = the **genus**; for a tetrahedron on a sphere $4-6+4=2$). Therefore

$$\ln Z = \sum_{h\geq 0} N^{\,2-2h}\, f_h(\lambda) \tag{5}$$

where f_0 is the sum of diagrams that can be drawn in the plane (**planar diagrams**), and the corrections go in steps of $1/N^2$ Known . A quark line is a single index line, and opens a boundary in the surface, multiplying by $1/N$. (5) is a rearrangement of the perturbation series; convergence is a separate question (4-12). The reason for fixing λ lies in the β function (the formula describing how the coupling constant changes with the scale μ at which we look; 17-01 § 02). Writing it in terms of λ with the b_0 of § 01, N disappears Computed .

$$\mu\frac{d\lambda}{d\mu} = -\frac{11}{24\pi^2}\,\lambda^2, \qquad \Lambda = \mu\,\exp\!\left(-\frac{24\pi^2}{11\,\lambda(\mu)}\right) \tag{6}$$

The left-hand side is how λ changes when the scale μ is changed slightly, and since the right-hand side is negative, λ is smaller at shorter distances (larger μ). Fixing λ fixes Λ too, and on the lattice this choice gives a smooth limit (Lucini–Teper) Physics .

04

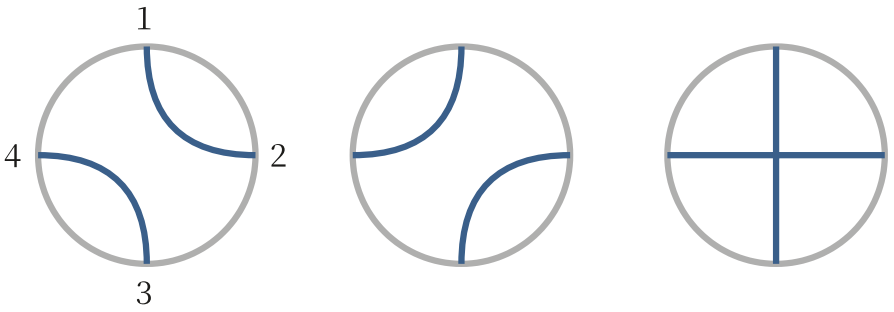
Counting: the genus expansion of a Gaussian matrix

The counting (4) can be checked in a model where spacetime is shrunk to a single point. Take one Hermitian matrix M distributed according to a Gaussian (1-08) (dM is the Lebesgue measure on the real components).

$$d\mu(M) \;\propto\; e^{-\frac{N}{2}\operatorname{tr} M^2}\, dM, \qquad \langle M_{ij} M_{kl}\rangle = \frac{1}{N}\,\delta_{il}\,\delta_{jk} \tag{7}$$

The expectation of a product of Gaussian variables is the sum over ways of pairing them (Wick's theorem, 2-07). The pairings of $\frac{1}{N}\operatorname{tr} M^4$ are the three of Figure 0: (a)(b) give $\frac{1}{N}\cdot N^3 \cdot N^{-2} = 1$ and (c) gives N^{-2}.

Figure 0 The three ways of joining legs 1–4 of $\operatorname{tr} M^4$ in pairs. (a)(b), where the strings do not cross, have 3 index rings; (c), where they cross, has 1 (it cannot be drawn on a sphere and needs a torus)



(a) 3 rings: 1 (b) 3 rings: 1 (c) 1 ring: $1/N^2$

In general

$$\left\langle \tfrac{1}{N}\operatorname{tr} M^4 \right\rangle = 2 + \frac{1}{N^2}, \qquad \left\langle \tfrac{1}{N}\operatorname{tr} M^{2k} \right\rangle = \sum_{g\ge 0} \varepsilon_g(k)\, N^{-2g} \tag{8}$$

where $\varepsilon_g(k)$ is the number, among the $(2k-1)!!$ pairings, of those of genus g (with $k+1-2g$ index rings). The coefficients can be obtained from the recursion of Harer–Zagier (1986).

$$(k+1)\,\varepsilon_g(k) = (4k-2)\,\varepsilon_g(k-1) + (k-1)(2k-1)(2k-3)\,\varepsilon_{g-1}(k-2) \tag{9}$$

Table 1 $\varepsilon_g(k)$ (full enumeration of pairings; agrees with recursion (9))

k	g=0	g=1	g=2	g=3	Total
4	14	70	21		105
5	42	420	483		945
6	132	2310	6468	1485	10395
7	429	12012	66066	56628	135135

Figure 1 Varying N and k changes the share of genus-g diagrams in $\langle \frac{1}{N}\operatorname{tr} M^{2k}\rangle$. Raising N leaves only the planar diagrams (g=0)



k = 4 N = 2.00 ⟨(1/N)tr M^8⟩ = 32.81, planar part 14 (42.7%)

Computed The numbers of Table 1 and Figure 1 were computed on this machine (for $k \le 7$, full enumeration checked against the recursion). If JavaScript is disabled, see Table 1.

The total $(2k-1)!!$ grows much faster than the number of planar diagrams, and at $k = 9$, $N = 1$ planar diagrams are under 0.02%. The large-N limit is the limit that keeps only the few diagrams.

Detour: Catalan numbers and the semicircle

A subsection you may skip.

The $g = 0$ column of Table 1 is the Catalan numbers $\frac{1}{k+1}\binom{2k}{k}$. As $N \to \infty$ the distribution of the eigenvalues of M approaches the semicircle $\rho(x) = \frac{1}{2\pi}\sqrt{4 - x^2}$ ($|x| \le 2$) (Wigner's semicircle law) Known . Integrating numerically the $2k$-th moment $\int x^{2k}\rho(x)\,dx$ of this density agrees with the Catalan numbers to 10 digits for $k \le 6$ Computed . Counting planar diagrams is the same as finding the large-N limit, and Brézin–Itzykson–Parisi–Zuber (1978) used this as a way of counting Known .

05

The sum of diagrams becomes a sum over random surfaces

We read the ribbon diagrams inside out. Replace four-point vertices by squares, three-point vertices by triangles, and propagators by the gluing of two edges, and each diagram becomes one closed surface glued from polygons. (5) is

$$\ln Z = \sum_{\Sigma} N^{\chi(\Sigma)}\, w_{\lambda}(\Sigma) \qquad (10)$$

a sum over **random surfaces**, where N counts the topology and λ the number of polygons (w_λ is a weight that includes the momentum integrals). As a metaphor it is a sum over shapes of a membrane; more precisely, (10) is a rearrangement of a series, and convergence as a sum over surfaces has to be shown separately.

Physics reads this membrane as the world sheet of the string swept out by the tube of confinement (17-03). N^{2-2h} has the form of the perturbative expansion of a string with string coupling $1/N$; in the problem statement, 'may be equivalent to a string theory with 1/N as the string coupling constant' Physics . In two-dimensional QCD, the coefficients of the $1/N$ expansion are fixed by sums over maps of surfaces, and it can be read as a theory of closed strings (Gross–Taylor 1993) Known .

06

Fluctuations disappear: factorisation and the master field

One more thing follows from the counting (4): the covariance of quantities normalised with $\frac{1}{N}\operatorname{tr}$ is smaller by $1/N^2$ Known .

$$\left\langle \tfrac{1}{N}\operatorname{tr} A \cdot \tfrac{1}{N}\operatorname{tr} B \right\rangle = \left\langle \tfrac{1}{N}\operatorname{tr} A \right\rangle \left\langle \tfrac{1}{N}\operatorname{tr} B \right\rangle + O\left(N^{-2}\right) \quad (11)$$

This is **factorisation**, and for $A = B$ the standard deviation shrinks as $1/N$.

Table 2 $\frac{1}{N}\operatorname{tr} M^4$ from random samples of the matrix (7) Computed

N	Mean	$2 + 1/N^2$	N × standard deviation
2	2.2541(80)	2.2500	7.19
8	2.0152(38)	2.0156	6.09
128	2.0003(15)	2.0001	6.06

At $N = \infty$, quantities made from traces (for the matrix model (7), quantities unchanged under conjugation $M \to UMU^\dagger$; for gauge theory, gauge-invariant quantities) stop being random, and the measure behaves as though concentrated on a single configuration. By custom, where it concentrates is called the **master field**. As a metaphor it is the picture of a heaving sea freezing; more precisely, what freezes is the distribution of quantities made from traces, while the components of the matrix keep fluctuating.

The representative case in gauge theory where factorisation works is the **Wilson loop**. It is the $\frac{1}{N}\operatorname{tr}$ of the matrix (an $N \times N$ unitary matrix) obtained by parallel transport with the gauge field along a closed path C, defined in 17-01 § 04. The **Schwinger–Dyson equations** are relations among expectation values that come from the integral not changing when the integration variable is shifted. Putting factorisation into these equations for Wilson loops gives equations closed in the expectations of loops alone (Makeenko–Migdal 1979) Physics . In the plane $\mathbb{R}^2$, Lévy constructed the large-N limit of the Yang–Mills measure (7-02) and showed that each Wilson loop converges in probability to a deterministic limit, and that these equations hold both at finite N and in the limit Known .

07

Large N seen on the two-dimensional lattice: the Gross–Witten transition

We write as much lattice vocabulary as needed (for details see 5-01). Replace spacetime by a mesh of lattice points, and on each edge ℓ joining two neighbouring points place a group element $U_\ell \in U(N)$. Write U_p for the product of the elements on the four sides of the smallest square of the lattice (a **plaquette**) p, taken in order around it; the action is made from $\operatorname{tr} U_p$. The $\frac{1}{N}\operatorname{tr}$ of the product of edge elements along a closed path is the lattice **Wilson loop**. Each U_ℓ follows the **Haar measure** (the uniform probability measure on the group, 1-08 § 03) multiplied by the weight of the action.

On the two-dimensional lattice, the **tree gauge**, which sets some of the edge elements to 1, makes the plaquettes independent, and a Wilson loop enclosing n faces is exactly w^n (7-02 § 02, § 03). Large N can be studied with the integral over a single plaquette. Weighting the Haar measure of U(N), set

$$d\mu(U) \;\propto\; \exp\Big(\frac{N}{\lambda}\operatorname{tr}(U+U^\dagger)\Big)\,dU, \qquad w_N(\lambda) = \Big\langle \tfrac{1}{N}\operatorname{Re}\operatorname{tr} U \Big\rangle \tag{12}$$

(in terms of the β of 7-02, $\beta = 2N/\lambda$). At finite N, Weyl's integration formula (which turns the Haar integral of a function depending only on traces into an integral over the eigenvalues $e^{i\theta 1},\dots,e^{i\theta N}$) and Heine's identity (which turns that eigenvalue integral into a determinant) give a determinant whose entries are the modified Bessel functions

$$I_n(x) = \frac{1}{2\pi}\int_0^{2\pi} e^{x\cos\theta}\cos n\theta\, d\theta$$

Known .

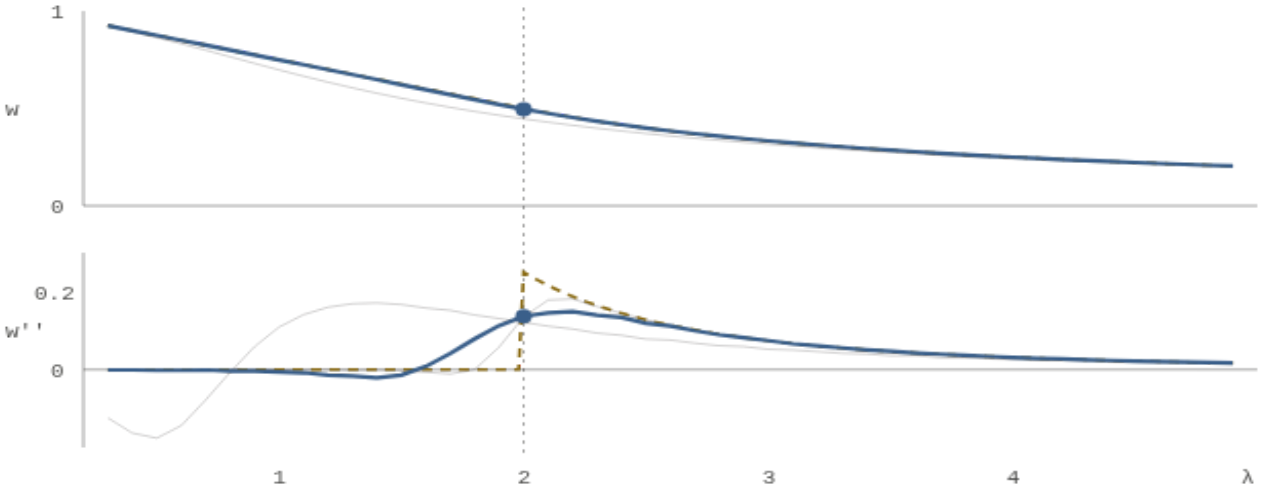
$$Z_N = \det\big[I_{j-k}(2N/\lambda)\big]_{j,k=1}^{N}, \qquad w_N = \frac{1}{2N}\frac{d\ln Z_N}{d(N/\lambda)} \tag{13}$$

Gross–Witten (1980) evaluated $N\to\infty$ by the **saddle-point method** (a method that approximates $\int e^{Nf(x)}dx$, containing a large parameter N, around the point where f is maximal) and found a third-order phase transition at $\lambda = 2$. The same transition appears in Wadia's 1979 preprint Known . In this normalisation the limit has the following form, and the finite-N values approach it as in Table 3 Computed .

$$w_\infty(\lambda) = \begin{cases} 1-\lambda/4 & (\lambda \le 2) \\ 1/\lambda & (\lambda \ge 2) \end{cases} \tag{14}$$

At $\lambda = 2$ both the value 1/2 and the slope −1/4 join up, and the second derivative jumps from 0 to 1/4 (the third derivative, for the free energy). The string tension is $\sigma = -\ln w$.

Figure 2 Choosing N draws $w_N(\lambda)$ in the upper panel and w''_N in the lower panel as thick lines (dashed is N=∞). Raising N brings the lower panel closer to the step at λ=2



N = 10 λ = 2.0 w = 0.49634 (N=∞ : 0.50000) σ = 0.700 w" ≈ 0.14 (N=∞ : step 0→0.25)

Table 3 $w_N(\lambda)$. The last column is formula (14)

λ	N=1	N=10	N=40	N=∞
1	0.69777	0.74937	0.74996	0.75000
2	0.44639	0.49634	0.49939	0.50000
4	0.24250	0.25000	0.25000	0.25000

Computed The numbers in this figure and table were computed on this machine ((13) evaluated to 300 digits; N=1,2 agree with direct integration to 10 digits; the second derivative in the figure is a difference with step 0.1). If JavaScript is disabled, see Table 3.

Three things can be read off Computed . (i) At finite N, Z_N is an entire function of N/λ and positive, so w_N is smooth, and the kink appears only at $N = \infty$. (ii) The manner of approach differs between the phases. The deviation is nearly $-1/(16N^2)$ at $\lambda = 1$; at $\lambda = 4$ it is exponentially small, 2×10^{-12} at $N = 20$; at $\lambda = 2$ doubling N only reduces it by about 1/2.5. (iii) Extending the strong-coupling formula $1/\lambda$ to $\lambda = 1$ gives 1, away from the correct 0.75. This is an example of the strong-coupling expansion not reaching beyond the transition (5-02).

Wadia's preprint (reprinted on arXiv) writes of this transition that it 'seems to be a pathology of the infinite N limit', and holds that in this model it is unrelated to the physical continuum limit (the $\lambda \to 0$ side). Still, $N \to \infty$ can create singularities that finite N does not have, so exchanging $N \to \infty$ and $a \to 0$ needs a reason.

> 08

The four-dimensional lattice: ratios move slowly with N

In four dimensions we rely on lattice numbers. Lucini–Teper (2001) fitted the continuum limits of $m_{0^{++}}/\sqrt{\sigma}$ (the lightest glueball and the string tension, 17-01) for SU(2) to SU(5) with a linear function of $1/N^2$, and summarised: 'all SU(N) gauge theories are close to SU(∞), at least as far as the low-lying mass spectrum is concerned' Physics .

$$\left.\frac{m_{0^{++}}}{\sqrt{\sigma}}\right|_N = a + \frac{b}{N^2} \quad (15)$$

Refitting (15) to the four points of the paper with error weights gives $a = 3.366 \pm 0.106$, $b = 1.93 \pm 0.54$, $\chi^2 = 0.26$ (2 degrees of freedom), with central values the same as the paper's $3.37(15) + 1.93(85)/N^2$. A constant alone gives $\chi^2 = 12.9$ (3 degrees of freedom) and does not fit Computed .

Table 4 Continuum-limit values of $m_{0^{++}}/\sqrt{\sigma}$ and the line (15) (in the Athenodorou–Teper column, SU(3) is from the 2020 paper, SU(2) and SU(∞) from the 2021 paper)

N	Lucini–Teper	Line (15)	Athenodorou–Teper
2	3.844(61)	3.848	3.781(23)
3	3.607(87)	3.580	3.405(21)
4	3.49(14)	3.487	
5	3.38(16)	3.443	
∞	3.37(15)	3.366	3.072(14)

Physics Extrapolated values from primary sources. Computed "Line (15)" is the fit on this machine (correlations and systematic errors ignored).

From SU(2) to SU(∞) the ratio changes by about 12.5% on the line (15), and by about 19% in the Athenodorou–Teper values Computed . The SU(∞) values of the two primary sources differ by about 2.0σ Computed . All are extrapolations presupposing that the continuum limit exists, and σ is extracted through the string picture. "N=3 is close to N=∞" is an observation within that premise.

09

What can be said as mathematics, and what cannot

Large-N results that have become theorems lie on the following shores Known .

- **Theories of planar diagrams**: 't Hooft (1983) treated a theory of planar diagrams alone in four-dimensional Euclidean space, and showed that 'if all particles involved have non-vanishing masses' and the couplings are small enough, an absolutely convergent procedure gives Green functions (expectations of products of fields; the

propagator is one of them) that uniquely solve the Dyson–Schwinger equations (the same as the Schwinger–Dyson equations of § 06) (abstract: 'an absolutely convergent procedure exists to obtain Green functions that uniquely solve the Dyson-Schwinger equations'). The gap is an assumption.

- **The strong-coupling lattice**: Chatterjee (2015) computed the expectations of large-N Wilson loops for SO(N) at strong coupling as absolutely convergent sums over trajectories of strings on the lattice, and Chatterjee–Jafarov, for the same strong-coupling SO(N), proved the $1/N$ expansion of Wilson loops as an asymptotic expansion (the first term is genus 0, higher orders surfaces of higher genus).
- **All N and β**: Cao–Park–Sheffield expressed the expectations of Wilson loops, for any matrix size $N \geq 1$, $\beta > 0$ and dimension $d \geq 2$, as sums over **embedded maps** (graphs drawn on a surface, mapped onto the lattice). For U(N) the weight of one map is proportional to the product of N^{χ} and the weights of the lattice edges (Weingarten functions, which themselves depend on N).
- **Two dimensions**: the planar master field (§ 06) and the area law at all couplings (7-02).

Seam

(1) The sum over surfaces holds for all β, but there is no estimate producing the area law or the gap on the continuum-limit side (large β). (2) Exchanging $N \to \infty$ and $a \to 0$ is not justified. (3) Even if something can be shown at $N = \infty$, the prize problem requires finite N, and the $1/N$ expansion has no guarantee of convergence. The problem statement too, about four dimensions, only goes as far as writing 'there might some day be an asymptotic solution in a large N limit'.

Large N is used again for strings (17-05) and for holography, which also takes $\lambda \to \infty$ (17-08). The distance to finite N is in 17-10.

What this chapter can say

What is said	What is not said
Known (4)(5) (as rearrangements of series), the difference between U(N) and SU(N).	Convergence of f_h. That the $1/N$ expansion approximates the theory in four dimensions.
Computed (6), Tables 1–3, Figures 1 and 2, the moments of the semicircle density, the check of (13) at N=1,2, the fit of Table 4.	Table 4 is a refit of extrapolated values from primary sources.

What is said	What is not said
Known Large N and the $1/N$ expansion for strong-coupling SO(N), sums over surfaces, the planar master field, the theory of planar diagrams assuming positive masses.	The area law and the gap on the weak-coupling side. Exchanging $N \to \infty$ and $a \to 0$. The return to finite N.
Physics The dominance of planar diagrams, the string coupling $1/N$, the loop equations, the smallness of the N-dependence of the ratios.	Not theorems of the four-dimensional continuum theory. This book does not itself show either the mass gap or the construction of the continuum limit (the published construction of the continuum limit: Volume 2).

Sources and reproduction

Item	Kind	Source / tools
(1), planar diagrams	Known	G. 't Hooft, Nucl. Phys. B72 (1974) 461. Abstract checked.
Quotations from the problem statement	Known	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory" (Clay) § 6.1, § 6.6. Text checked.
Counting with matrix integrals	Known	E. Brézin, C. Itzykson, G. Parisi, J.-B. Zuber, CMP 59 (1978) 35. Abstract checked.
(9) / the semicircle law	Known	J. Harer, D. Zagier, Invent. Math. 85 (1986) 457 / E. P. Wigner, Ann. of Math. 62 (1955) 548. Bibliographic data only.
The factorisation (11), the difference between U(N) and SU(N)	Known	Treated as standard facts of large N (within the framework of 't Hooft 1974 and Brézin–Itzykson–Parisi–Zuber 1978). For these two points, the texts of both papers: bibliographic data only. (11) was also checked with the numbers of Table 2.
The determinant (13)	Known, Computed	The standard rewriting by Weyl's integration formula and Heine's identity (original sources: bibliographic data only, not checked). Checked on this machine against direct integration for N=1,2.
Two-dimensional QCD and strings	Known	D. J. Gross, W. Taylor, NPB 400 (1993) 181, hep-th/9301068. Abstract checked.
Loop equations	Physics	Yu. M. Makeenko, A. A. Migdal, PLB 88 (1979) 135. Abstract checked.
The planar master field	Known	T. Lévy, arXiv:1112.2452. Abstract checked.
The transition, the quotation from Wadia	Known	D. J. Gross, E. Witten, PRD 21 (1980) 446. Abstract checked. S. R. Wadia, arXiv:1212.2906 (a reprint of the 1979 preprint). Abstract checked. (14) not checked in the texts of either paper; confirmed as the limit of the finite-N values.
Table 4, quotations	Physics	B. Lucini, M. Teper, JHEP 0106 (2001) 050, hep-lat/0103027 (text checked). A. Athenodorou, M. Teper, arXiv:2007.06422 Table 17 (SU(3)) / arXiv:2106.00364 Table 23 (SU(2)), Table 31 (SU(∞)). Tables in the text checked.

Item	Kind	Source / tools
The theorems of § 09	Known	G. 't Hooft, "Rigorous construction of planar diagram field theories in four-dimensional Euclidean space", CMP 88 (1983) 1. S. Chatterjee, arXiv:1502.07719. S. Chatterjee, J. Jafarov, arXiv:1604.04777. S. Cao, M. Park, S. Sheffield, arXiv:2307.06790. Abstracts checked.
Tables, figures, (6), (13)	Computed	Python 3 on this machine (numpy, decimal at 300 digits).

Next chapter to read: 17-05 String theory and extra dimensions — what is needed to read the sum over surfaces as a theory of strings.

String theory and extra dimensions — 26, 10 and 11 are numbers that come from conditions of consistency

> 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 quantising a string as a fundamental object takes away the free choice of the dimension of spacetime. How 26 for the bosonic string is derived in two ways (the first excited state / the central charge), and 10 for the superstring by the same balance. That 11 is a number that comes from supergravity.

Prerequisite chapters — 17-04 Large N. We use the zero-point energy and $-1/12$ of 17-03. Group representations (1-02), Lorentz transformations (2-04) and the harmonic oscillator (2-06) are written on the spot as far as needed.

Said first — The derivations are at the level of physics. The two-dimensional computations on the world sheet are exact, but string theory has no non-perturbative definition, and the existence of M-theory is a hypothesis. The story of extra dimensions shows nothing about the construction of four-dimensional Yang–Mills or about the mass gap. This book does not itself show either the mass gap or the construction of the continuum limit (the published construction of the continuum limit is read in Volume 2).

The order of this chapter

1. Making the string a fundamental object — how it differs from the string of the tube
2. Counting the transverse oscillations — the mass formula
3. Why 26 (I) — the first excited state is massless
4. Counting the levels — the tachyon and the growth (Detour: partition numbers and Hardy–Ramanujan, Detour: the growth of states and the Hagedorn temperature)
5. Why 26 (II) — the balance of the central charge

6. Why 10 — world-sheet supersymmetry (Detour: 26 − 10 = 16 and the heterotic string)
7. Why 11 — the number of supergravity
8. The "critical dimension" is a critical central charge — it need not be geometry
9. Back to Yang–Mills — the string of the tube, and holography

---

01

Making the string a fundamental object

17-03 treated the tube of confinement as a single string, and in 17-04 § 05 $1/N$ played the role of the string coupling. String theory too is said to have begun from models of mesons (**dual resonance models**). The scattering amplitude of Veneziano (1968) is often cited as the starting point; for example, the abstract of Neveu–Schwarz (1971) opens with 'A new dual-resonance model of mesons is constructed ...' Known . For mesons, spin J and the square of the mass line up on a straight line (Regge trajectories). If a string of tension T and length $2R$ rotates with angular velocity $1/R$, its ends move at the speed of light, and integrating the energy and the angular momentum (2-04) gives $E = \pi T R$ and $J = \pi T R^2/2$, that is

$$J = \alpha' E^2, \qquad \alpha' = \frac{1}{2\pi T}. \ (1)$$

For the quantised open string, the maximum spin at level N is $J_{\max} = \alpha' M^2 + 1$ (Tong § 3.1.3). Putting the $\sqrt{\sigma} = 485$ MeV of 17-01 into T gives $\alpha' = 0.677\ \text{GeV}^{-2}$, $\sqrt{\alpha'} \approx 0.16$ fm Computed . The **fundamental string** of this chapter is not the effective theory of a tube; it is required to be quantisable without contradiction at however short a distance. Then the dimension D of spacetime stops being a free input. In this chapter, following the string literature, we write D for the dimension of spacetime. It is the same as the d of 17-03 and of the table of notation in the appendix, not the dimension of space alone.

02

Counting the transverse oscillations: the mass formula

A string sweeps out a surface (the **world sheet**) in spacetime $\mathbb{R}^{1,D-1}$. In **light-cone gauge** we use the light-cone coordinates $X^{\pm} = (X^0 \pm X^{D-1})/\sqrt{2}$, recombining time X^0 and one longitudinal direction X^{D-1}, and make X^+ proportional to the world-sheet time τ ($X^+ = x^+ + \alpha' p^+ \tau$; x^+ and p^+ are the $+$ components of the position and

momentum of the centre of mass of the string; Tong (2.13)). The oscillations of X^- do not vanish, but they are fixed by the transverse oscillations through the constraint equations. Only the $D-2$ transverse directions oscillate independently (Goddard–Goldstone–Rebbi–Thorn 1973) Known . This is the same counting as the transverse displacement $h \in \mathbb{R}^{d-2}$ of 17-03 ($d = D$). The n-th oscillation in direction i is a harmonic oscillator (2-06), and each application of the creation operator α^i_{-n} raises the **level** N by n. The mass of the open string is

$$\alpha' M^2 = N - a \tag{2}$$

where $-a$ is the zero-point energy (the sum of $\frac{1}{2}n$ over all oscillators). Its finite part is the $-1/12$ of 17-03 (17-03 § 04), so

$$a = -\frac{D-2}{2}\sum_{n\geq 1} n \;\longrightarrow\; -\frac{D-2}{2}\cdot\Big(-\frac{1}{12}\Big) = \frac{D-2}{24}. \tag{3}$$

The Lüscher term $-\pi(D-2)/(24R)$ of 17-03 and the $(D-2)/24$ of the Nambu–Goto levels (17-03 § 06) are this same number Known . Tong admits that this treatment of the sum is 'famously unconvincing', and derives it again by the central-charge method, the same as § 05 of this chapter (Tong § 4.4.1). The closed string has right-moving oscillations α and left-moving oscillations $\tilde{\alpha}$, and its levels satisfy $N = \tilde{N}$ (Tong (2.26)).

03

Why 26 (I): the first excited state must be massless

The internal states of a particle form a representation of the rotations that leave its momentum unchanged (the **little group**) (1-02; Wigner 1939). For mass $M > 0$ we can go to the rest frame and the little group is $SO(D-1)$; for zero mass it is the $SO(D-2)$ that preserves the momentum $(p, 0, \ldots, 0, p)$ (precisely, it is the group $ISO(D-2)$ of rotations and translations in $D-2$ dimensions, but what acts on finitely many states is the rotation part $SO(D-2)$). A photon in four dimensions has 2 polarisations, a massive spin-1 particle 3 Known . The level-1 states $\alpha^i_{-1}|0\rangle$ of the open string are a transverse vector with $D-2$ states, one short of the $D-1$ components a massive vector needs. To keep Lorentz symmetry it must be massless, and

$$\alpha' M^2 = 1 - a = 0 \implies a = 1 \overset{(3)}{\implies} \frac{D-2}{24} = 1 \implies D = 26. \tag{4}$$

Tong makes the same argument for the first excited state of the closed string ($(D-2)^2$ states) and writes 'these states will have to be massless. And this is only the case if the dimension of spacetime is D = 26' Physics .

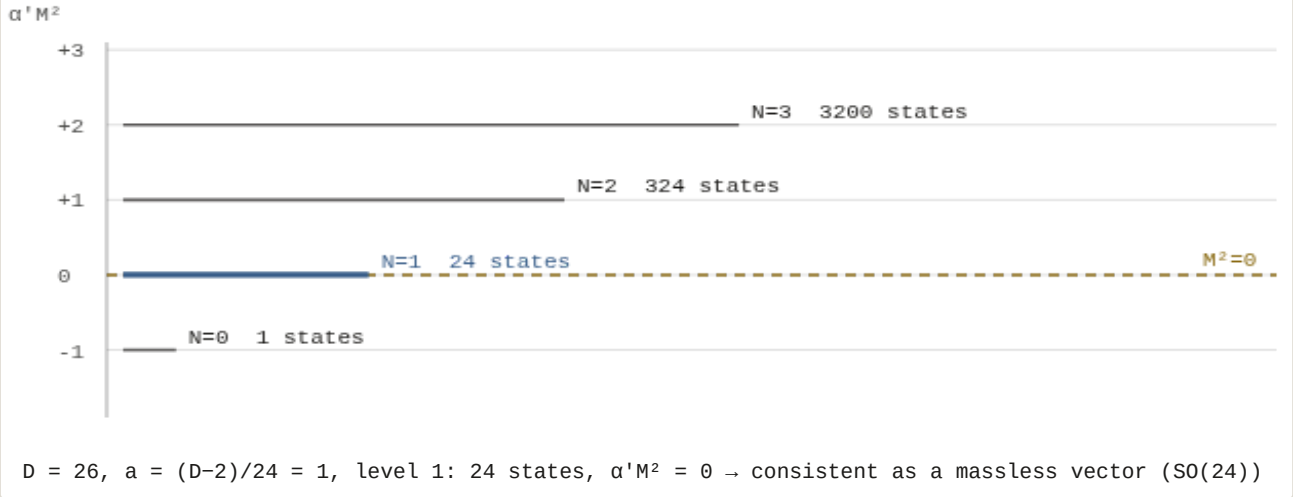
Checking with the Lorentz algebra

More carefully, we check whether the conserved quantities $M^{\mu\nu}$ that generate Lorentz transformations (rotations and boosts) satisfy the relations of the Lorentz algebra in light-cone gauge too. $M^{\mu\nu}$ includes the angular momentum carried by the oscillations of the string. The crux is $[M^{i-}, M^{j-}] = 0$, where i, j are transverse directions and $-$ is the light-cone direction X^- (Tong § 2.4). Since the oscillations of X^- are fixed by the transverse oscillations, this commutator can be written with the transverse oscillators.

$$[M^{i-}, M^{j-}] \propto \sum_{n>0} \left[\left(\frac{D-2}{24} - 1\right)n + \frac{1}{n}\left(a - \frac{D-2}{24}\right)\right](\alpha^i_{-n}\alpha^j_n - \alpha^j_{-n}\alpha^i_n) + (\alpha \leftrightarrow \tilde{\alpha}) \quad (5)$$

The proportionality factor $2/(p^+)^2$ is omitted. This is the formula for the closed string, and $(\alpha \leftrightarrow \tilde{\alpha})$ is a term of the same form for the left-moving oscillations. It vanishes for every n only when $D = 26$ and $a = 1$ Known . The original source is Goddard–Goldstone–Rebbi–Thorn (1973), whose abstract also states that the angular momentum operators generate the Lorentz group $O(D-1,1)$ 'only when the first excited state is a photon, i.e., a spin-one massless state, and when D = 26'. The first excited state of the closed string has $24^2 = 576 = 299 + 276 + 1$ states Computed , and Tong assigns these to the symmetric traceless, antisymmetric and trace fields $G_{\mu\nu}, B_{\mu\nu}, \Phi$ of $SO(24)$, reading the first as a massless spin 2, that is, the graviton Physics .

Figure 1 Varying D moves the levels $\alpha' M^2 = N - a$ up and down (the length of a line is the logarithm of the number of states). The thick level 1 (1/2 for the superstring) sits on $M^2 = 0$ only for D=26 for the bosonic string and D=10 for the superstring



Computed The numbers in this figure were computed on this machine. If JavaScript is disabled, see (4) and Table 2.

```
04
```

Counting the levels: the tachyon and the growth of the number of states

$N = 0$ of the open string at $D = 26$ has $\alpha' M^2 = -1$: a **tachyon**, with negative squared mass (Tong § 3.1.1). It is a field whose mass term has the opposite sign (2-02), meaning that we are expanding around a maximum of the potential Physics . The number of states d_N at level N is the number of ways of splitting N into a sum of "positive integers in 24 colours".

$$\sum_{N \geq 0} d_N \, q^N = \prod_{n \geq 1} (1 - q^n)^{-24} \tag{6}$$

The 324 at level 2 is the dimension of the symmetric traceless rank-2 tensor of $SO(25)$, and the 3200 at level 3 is $2900 + 300$ (symmetric traceless rank 3 and antisymmetric rank 2) Computed . According to Tong all excitations fit into representations of $SO(D-1)$, and the condition is on level 1 alone. That the numbers at levels 2 and 3 match sums of dimensions of representations of $SO(D-1)$ we checked for $D = 4$–39 Computed . Both are polynomials in D of degree at most three, so this makes them identities valid for any D.

Table 2 Levels of the open bosonic string ($D = 26$). The ratio is to the Bessel form of the detour ($n = N - 1$)

N	$\alpha' M^2$	d_N	Ratio
0	−1	1	
1	0	24	
2	1	324	1.000009
3	2	3200	0.999998
10	9	6.392×10^{8}	1.00000000
400	399	1.446×10^{91}	1.00000000

Computed The numbers in this table were computed on this machine.

Detour: partition numbers and Hardy–Ramanujan

A subsection you may skip.

With one colour the coefficients of (6) are the partition numbers $p(N)$ ($p(4) = 5$ from $4 = 3+1 = 2+2 = 2+1+1 = 1+1+1+1$), and Hardy and Ramanujan (1918) gave $p(N) \sim \frac{1}{4\sqrt{3}\,N} e^{\pi\sqrt{2N/3}}$ Known . For (6) with 24 colours, setting $n = N - 1$ gives $d_N \approx \frac{1}{\sqrt{2}} n^{-27/4} e^{4\pi\sqrt{n}}$, but the approach is slow, and the ratio is still 0.71 at $N = 400$. What fits well is the form $2\pi n^{-13/2} I_{13}(4\pi\sqrt{n})$ using the modified Bessel function I_{13}, which already agrees to 5 digits at $N = 2$ (Table 2) Computed . $I_\nu(x)$ is a function that behaves like $e^x/\sqrt{2\pi x}$ when x is large. Putting this

in gives $2\pi n^{-13/2}\, e^{4\pi\sqrt{n}}/\sqrt{8\pi^2\sqrt{n}} = \frac{1}{\sqrt{2}} n^{-27/4} e^{4\pi\sqrt{n}}$, and the two formulas agree at large n. The ratio 0.71 is the error from replacing I_{13} by this approximation.

Detour: the growth of states and the Hagedorn temperature

A subsection you may skip.

An open string at level N has $\alpha' M^2 = N - 1$, so setting $n = N - 1$ as in the detour above gives $\alpha' M^2 = n$, that is $M = \sqrt{n/\alpha'}$. From the asymptotic form of the detour above, the number of states grows at the rate $e^{4\pi\sqrt{n}} = e^{4\pi\sqrt{\alpha'}M}$ (the power of n in front changes slowly compared with the exponential). Write Θ for the temperature (to distinguish it from the tension T; in units with Boltzmann's constant 1). The weight of a state of mass M is the Boltzmann factor $e^{-M/\Theta}$ (2-08), so the product of number and weight has the form $e^{(4\pi\sqrt{\alpha'} - 1/\Theta)M}$, and for $\Theta > \Theta_H = 1/(4\pi\sqrt{\alpha'})$ it grows with M and the sum diverges. This Θ_H is the **Hagedorn temperature** Physics . Putting $d = D = 26$ and $\sigma = 1/(2\pi\alpha')$ into the Hagedorn length L_H of 17-03 § 06 ($\sigma L_H^2 = \pi(d-2)/3$, where σ is the tension, the T of this chapter) gives $L_H = 4\pi\sqrt{\alpha'}$, and its reciprocal $1/L_H$ coincides with Θ_H Computed . It is the same reading as in 17-03, where $1/L$ became the temperature when the periodic direction was regarded as time.

05

Why 26 (II): the balance of the central charge

The second derivation looks at the D fields X^μ on the world sheet as two-dimensional free scalar fields (2-02). A two-dimensional theory whose form does not change under a change of scale (a **conformal field theory**) has a **central charge** c, and one free scalar has $c = 1$. More precisely, a conformal field theory is a theory invariant under angle-preserving maps (**conformal transformations**), of which changes of scale are a part. In two dimensions there are infinitely many such maps, and this constrains the theory strongly. c appears in the ground-state energy on a cylinder of circumference L (when the left and right c are equal).

$$E_0(L) = -\frac{\pi c}{6L} \quad (7)$$

(Blöte–Cardy–Nightingale 1986, Affleck 1986; Tong § 4.4.1) Known . The $-\pi(D-2)/(6L)$ of the closed tube of 17-03 is the case $c = D - 2$, and about the $-\pi c/(24R)$ of a tube with ends Tong writes 'c counts the number of degrees of freedom of the QCD flux tube'. The Lüscher coefficient measures the central charge of the tube.

In Polyakov's formulation the metric of the world sheet is also a field to be integrated, and changes of coordinates and changes of scale of the metric (Weyl transformations) are gauge symmetries. Fixing the gauge brings in auxiliary fields (ghosts; the same

mechanism as Faddeev–Popov in 4-08), with central charge -26. A quantum breaking of a gauge symmetry (an anomaly) is not allowed, so the total must be 0 (Tong § 5.3) Physics .

$$c_{\text{matter}} + c_{\text{ghost}} = D \cdot 1 + (-26) = 0 \implies D = 26 \quad (8)$$

The $D - 2$ fields and $-1/12$ in light cone, and the D fields of the covariant form with the ghosts' -26, select the same 26. More precisely, (8) is a condition only as long as the scale of the metric (Weyl) is kept as a gauge symmetry. For $D \neq 26$ the scale degree of freedom does not disappear but remains, and behaves as a new field called the Liouville field (the non-critical string of Tong § 5.3.2). The abstract of Polyakov (1981) too writes 'At D = 26 the string dynamics is that of harmonic oscillators ..., otherwise it is described by the nonlinear integrable theory', treating the $D \neq 26$ side as a different theory.

06

Why 10: world-sheet supersymmetry

The bosonic string has no fermions (relatives of the electron and the quarks), and has a tachyon. Add to the world sheet fermions ψ^μ as partners of X^μ ($c = 1/2$ each), make the **world-sheet supersymmetry** that exchanges the two a gauge symmetry, and the ghosts add central charge $+11$.

$$D\left(1 + \frac{1}{2}\right) + (-26 + 11) = 0 \implies D = 10 \quad (9)$$

In Tong § 5.3.1, 'c = 26 − 11 = 15' and 'D(1 + 1/2) = 15' Physics . We check it in light cone too. Depending on the boundary condition, the fermion oscillations take half-integer labels $r = \frac{1}{2}, \frac{3}{2}, \ldots$ (Neveu–Schwarz, NS) or integer ones (Ramond, R), and their zero-point energy has the opposite sign, $-\frac{1}{2}\sum r$. The finite parts are in Table 1.

Table 1 The sum minus $1/\varepsilon^2$ (cutoff $e^{-\varepsilon r}$)

ε	Integer $\sum n e^{-\varepsilon n}$	Half-integer $\sum r e^{-\varepsilon r}$
0.1	-0.0832917	0.0416302
0.001	-0.0833333	0.0416667
→ 0	-1/12	+1/24

Computed The numbers in this table were computed on this machine (closed forms checked against direct sums).

The zero-point energy per transverse direction is $-\frac{1}{24} - \frac{1}{48} = -\frac{1}{16}$ for NS and $-\frac{1}{24} + \frac{1}{24} = 0$ for R Computed . The NS levels are $\alpha' M^2 = N - \frac{D-2}{16}$ (N also takes half-integers), and the first excited state $\psi^i_{-1/2}|0\rangle$ is again a vector with $D - 2$ components. The

massless condition $\frac{1}{2} - \frac{D-2}{16} = 0$ gives $D = 10$, agreeing with (9) ("Superstring" in Figure 1). The tachyon of the NS ground state, $\alpha' M^2 = -\frac{1}{2}$, is said to be removed by a rule that discards part of the states (attributed to Gliozzi–Scherk–Olive 1977) Physics .

Detour: 26 − 10 = 16 and the heterotic string

A subsection you may skip.

Green–Schwarz (1984) named $SO(32)$ and $E_8 \times E_8$ as the groups for which the gauge and gravitational anomalies in 10 dimensions cancel, and Gross–Harvey–Martinec–Rohm (1985) built the heterotic string as 'a chiral combination of the closed D=26 bosonic and D=10 fermionic strings' Known . The difference between 26 and 10 is 16, and $SO(32)$ and $E_8 \times E_8$ both have rank 16 and dimension 496 (E_8 has 240 roots and rank 8, giving 248; 1-04) Computed . How this 16 becomes a gauge group is in 17-07.

07

Why 11: a number of supergravity, not of strings

11 is not a critical dimension of strings. The abstract of Cremmer–Julia–Scherk (1978) reads 'We present the action and transformation laws of supergravity in 11 dimensions' Known . **Supergravity** is a field theory containing the graviton and its supersymmetric partner (the gravitino), and that 11 is its upper limit is said to follow from the classification of Nahm (1978).

The skeleton is this Physics . **Supercharges** are the generators of the symmetry that exchanges bosons and fermions (supersymmetry), and Q is the number of their real components. With Q supercharges, the range of helicities (the amount of rotation about the direction of motion) of a massless multiplet is $Q/8$, and for spin at most 2 (range at most 4) $Q \leq 32$. 11 dimensions is said to be the last dimension that fits within this.

The fields of 11-dimensional supergravity are the graviton, a **3-form** and the gravitino. A 3-form is an antisymmetric field with three indices, the electromagnetic potential A_μ (2-03) with three indices instead of one. Counting the degrees of freedom with the little group $SO(9)$ (the $SO(D-2)$ of $D = 11$), the 128 bosons, graviton $45 - 1 = 44$ and 3-form $\binom{9}{3} = 84$, balance the gravitino's $9 \cdot 16 - 16 = 128$ Computed . In counting the gravitino we used, only as a number, that a **spinor** of $SO(9)$ (the quantity carrying the internal degrees of freedom of a fermion) has 16 components.

Witten (1995) argued in the abstract 'It is argued ... that eleven-dimensional supergravity arises as a low energy limit of the ten-dimensional Type IIA superstring'.

The **IIA string** is a kind of closed superstring in 10 dimensions in which the supersymmetries carried by the right- and left-moving oscillations have opposite handedness (chirality). With λ the coupling of the IIA string, the radius of the 11th circle in 11-dimensional units is

$$r(\lambda) = \lambda^{2/3} \quad (10)$$

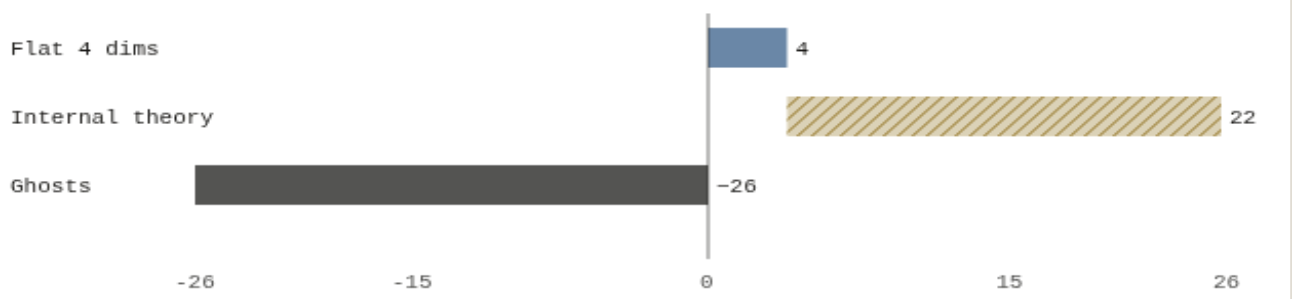
(formula (2.13) of that paper) Physics . At weak coupling the circle shrinks, and states with momentum n around the circle have mass $\sim |n|/\lambda$ in the string metric (formula (2.14)): heavy and not seen. As a metaphor, the 10 dimensions of the string are how a thin 11-dimensional tube looks from afar. More precisely, this is a conjecture from low-energy arguments, and the IIA string at strong coupling has no independent definition. Hořava–Witten (manuscript 1995, published 1996) **proposed** a link between the heterotic string and 11 dimensions in the form 'We propose'. The following paper (1996) argued that the gauge group is fixed to one E_8 for each boundary, wrote that this gives new material testing 'the hypothesis that there is a consistent quantum M-theory', and treated the existence of M-theory as a hypothesis. This 11-dimensional theory (M-theory) has no complete definition.

08

The "critical dimension" is a critical central charge: extra dimensions need not be geometry

What (8)(9) require is not a dimension but a central charge. Tong writes 'the "critical dimension" of string theory is something of a misnomer: it is really a "critical central charge"' Known . For flat four-dimensional spacetime, for the bosonic string it suffices to add **any conformal field theory with** $c = 22$ to the $c = 4$ of $X^0, \dots, X^3$, and Tong continues 'It is what we really mean when we talk about the "extra hidden dimensions of string theory"'. For the superstring flat four dimensions give $c = 6$ and the rest is $c = 9$ Computed . As a metaphor, only the total of the ledger is fixed and the items are free. More precisely, the internal theory must also satisfy other conditions (world-sheet supersymmetry and so on), and it can be read as geometry, as a manifold, only in special cases. How the extra dimensions then become invisible is in 17-06, and how gauge fields come out in 17-07.

Figure 2 Varying the number of flat dimensions d_{flat} changes the breakdown of the central-charge ledger. The internal theory (hatched) takes up what is removed, and the total stays 0



Bosonic string, flat 4 dims c = 4, internal c = 22, ghosts c = −26, total 0

Computed The arithmetic of this figure was checked on this machine. If JavaScript is disabled, read it as $d_{\text{flat}} + c_{\text{int}} = 26$ (for the superstring $\frac{3}{2}d_{\text{flat}} + c_{\text{int}} = 15$).

09

Back to Yang–Mills: the four-dimensional tube, and the dimensions of holography

Then is the string of the four-dimensional tube of 17-03, not being 26-dimensional, inconsistent? The abstract of Polchinski–Strominger (1991) states that the long-distance effective theory of four-dimensional strings such as 'QCD strings' is 'an interacting Poincaré-invariant conformal field theory with four bosons and c=26', and showed the compatibility to second order in perturbation Physics . The 4 bosons, which would have $c = 4$ if free, carry $c = 26$ through their interaction. What the critical dimension constrains is the fundamental string; the effective theory of a long tube is consistent as it is, in four dimensions. But the effective theory breaks down at short distances, and there is no mathematical derivation either (17-03 § 06).

The extra dimensions of the fundamental string return to Yang–Mills through holography (17-08), and the breakdown differs from model to model. Maldacena's $AdS_5 \times S^5$ (AdS_5 is a five-dimensional spacetime of constant negative curvature, with the four-dimensional directions and a radial direction; S^5 is the five-sphere) is 10 = 4 + radial + S^5; Witten 1998 § 4 is 11 = 4 + two circles + radial + S^4 (one circle fewer, 10, on reducing to IIA); the improved holographic QCD of Gürsoy–Kiritsis is 5 = 4 + radial. The dimensions that make the gap also differ by model: in Witten's type, a two-dimensional cigar formed by the radial direction and a shrinking circle; in the bottom-up type, a single radial direction and an infrared singularity (17-09) Physics . These extra dimensions are tools of a dual description, not claimed to be real dimensions, and are a separate matter from the experimental lower bounds (17-06).

Seam

(1) String theory has no non-perturbative definition, and the very existence of M-theory is a hypothesis. (2) Holographic computations work on the side of $N \to \infty$ and 't Hooft coupling $\to \infty$, the opposite direction to the limit that returns to four-dimensional pure Yang–Mills (17-10). (3) The problem statement (Jaffe–Witten,

2000) writes 'at present we do not know' of a non-trivial relativistic field theory in four dimensions satisfying axioms such as Wightman's (4-09). The four-dimensional theory was constructed by the construction published in 2026, as a limit of Wilson's lattice (the committee's assessment; Volume 2). The road of extra dimensions does not construct a four-dimensional theory in this sense.

What this chapter can say

What is said	What is not said
Known The counting of (3)(5) (as a two-dimensional world-sheet computation in light cone).	A non-perturbative definition of string theory. The existence of M-theory.
Physics The reading of (4) (reading the first excited state as massless) and the balance of (8)(9) (the central charge of the ghosts and the anomaly argument).	The treatment of the non-critical string (the Liouville field) for $D \neq 26$.
Computed Tables 1 and 2, Figures 1 and 2, 576, 496, 128, the Hagedorn temperature (the $1/L_H$ of 17-03), α'.	Why the Bessel form of Table 2 fits so well.
Physics The readings of the graviton and the tachyon, the relation between 11 dimensions and IIA, Hořava–Witten (a proposal).	Checking "11 is the upper limit" in the original source.
Physics That the effective theory of the four-dimensional tube is consistent with $c = 26$ (to second order in perturbation). The breakdown of the dimensions of the models.	Consistency of the effective string at all orders. That a tube exists in four-dimensional Yang–Mills. Constructing a four-dimensional theory from extra dimensions. This book does not itself show either the mass gap or the construction of the continuum limit (the published construction of the continuum limit: Volume 2).

Sources and reproduction

Item	Kind	Source / tools
$J_{\max}$, (2)–(9), each quotation	Known, Physics	D. Tong, "String Theory", arXiv:0908.0333 § 2.2–2.4, § 3.1, § 4.4.1, § 5.3 (including the non-critical string of § 5.3.2). Text checked.
Light-cone quantisation and $D = 26$ (the original source of (4)(5)) / 'At D = 26 …'	Known, Physics	P. Goddard, J. Goldstone, C. Rebbi, C. B. Thorn, Nucl. Phys. B56 (1973) 109 / A. M. Polyakov, Phys. Lett. B103 (1981) 207. Abstracts checked.

Item	Kind	Source / tools
Meson models, NS and R / the rule removing the tachyon	Known, Physics	G. Veneziano, Nuovo Cim. A57 (1968) 190 (bibliographic data only; its standing as the starting point of dual resonance models not checked) / A. Neveu, J. H. Schwarz, Nucl. Phys. B31 (1971) 86 / P. Ramond, Phys. Rev. D3 (1971) 2415 (abstracts checked) / F. Gliozzi, J. Scherk, D. Olive, Nucl. Phys. B122 (1977) 253 (abstract only; the attribution of the projection not checked).
The little group / formula (7)	Known	E. P. Wigner, Ann. of Math. 40 (1939) 149 / H. W. J. Blöte, J. L. Cardy, M. P. Nightingale, PRL 56 (1986) 742 / I. Affleck, PRL 56 (1986) 746. Bibliographic data only (content checked in Tong).
Anomaly cancellation / the heterotic string	Known	M. B. Green, J. H. Schwarz, Phys. Lett. B149 (1984) 117 / D. J. Gross, J. A. Harvey, E. J. Martinec, R. Rohm, PRL 54 (1985) 502. Abstracts checked.
11-dimensional supergravity / the upper limit	Known	E. Cremmer, B. Julia, J. Scherk, Phys. Lett. B76 (1978) 409 (abstract checked) / W. Nahm, Nucl. Phys. B135 (1978) 149 (abstract only; "11 is the upper limit" not checked).
(10), KK masses, Hořava–Witten	Physics	E. Witten, Nucl. Phys. B443 (1995) 85, hep-th/9503124 (text checked; formulas (2.13)(2.14)) / P. Hořava, E. Witten, Nucl. Phys. B460 (1996) 506, hep-th/9510209 (manuscript 1995-10) and Nucl. Phys. B475 (1996) 94, hep-th/9603142 (abstracts checked).
The four-dimensional tube and $c = 26$	Physics	J. Polchinski, A. Strominger, PRL 67 (1991) 1681. Abstract checked.
Asymptotic form of the partition numbers	Known	G. H. Hardy, S. Ramanujan, Proc. London Math. Soc. (2) 17 (1918) 75. Bibliographic data only.
Breakdown of the models / the problem statement	Physics, Known	J. Maldacena, Adv. Theor. Math. Phys. 2 (1998) 231, hep-th/9711200 / E. Witten, Adv. Theor. Math. Phys. 2 (1998) 505, hep-th/9803131 § 4 / U. Gürsoy, E. Kiritsis, JHEP 02 (2008) 032, arXiv:0707.1324 (abstract) / A. Jaffe, E. Witten, "Quantum Yang–Mills Theory" (the Clay problem statement, 2000). Witten 1998 § 4 and the problem statement: text checked.
Tables, figures, arithmetic	Computed	Python 3 on this machine. zero_point.py (Table 1), levels.py (Table 2, $SO(D-1)$, Hagedorn, α'), counts.py (576, 128, 496, the balance).

Next chapter to read: 17-06 How extra dimensions become invisible — why extra dimensions that can be read as geometry do not show up in four-dimensional experiments.

How extra dimensions become invisible — waves on a small circle look like a tower of heavy 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 extra dimensions being invisible means that fluctuations in those directions become a sequence of heavy particles (the KK tower), which cannot be produced at low energy. That the lowest rung of the tower is the first eigenvalue of the Laplacian, bounded by the diameter when the curvature is non-negative (Zhong–Yang). Two ways of hiding, Calabi–Yau and branes, and the experimental lower bounds.

Prerequisite chapters — 17-05 String theory and extra dimensions. We use 2-02 § 06, 2-05 § 08 and 1-10 § 07. Fourier series and the Rayleigh quotient are written on the spot, and the metric, the Laplacian, Ricci curvature and the Einstein equations are defined on the spot as far as they are used.

Said first — Whether the real world has extra dimensions is not known, and the experimental numbers are lower bounds. The gap of the KK tower comes from the internal space being small, and is a different thing from the mass gap of Yang–Mills (which arises from interaction on $\mathbb{R}^4$). This book does not itself show either the mass gap or the construction of the continuum limit (the published construction of the continuum limit is read in Volume 2).

The order of this chapter

1. What "seeing" means — the resolution is ħc/E
2. A field on a circle — one field becomes a tower
3. The lowest rung of the tower — the first eigenvalue and the diameter (the minimum of metric, Laplacian and Ricci curvature)
4. Kaluza and Klein — charge is momentum
5. The six dimensions of strings — Calabi–Yau
6. Confining to a brane — ADD and RS (Detour: extra dimensions without a gap)
7. Experimental lower bounds — Newton's law and colliders

8. Back to Yang–Mills — the KK gap is not the mass gap

01

What "seeing" means — the resolution is ħc/E

To see, in physics, is to collide particles and examine how they bounce off. To resolve a structure of size ℓ needs a probe with wavelength shorter than ℓ, that is, high energy (2-05).

$$\ell \approx \frac{\hbar c}{E}, \qquad \hbar c = 197.327\ \text{MeV}\cdot\text{fm} \quad (1)$$

1 TeV gives 2.0×10^{-19} m, the LHC collision energy of 13.6 TeV gives 1.5×10^{-20} m, and the Planck mass 1.22×10^{19} GeV gives 1.6×10^{-35} m Computed . From here on we set $\hbar = c = 1$ (2-04 § 08). As a metaphor, a garden hose seen from afar is a line, but up close the direction going around it becomes visible. More precisely, what this chapter says is that "fluctuations along a small circle need high energy".

02

A field on a circle — one field becomes a tower

Consider the five-dimensional spacetime $\mathbb{R}^{1,3} \times S^1$ with a circle of radius R attached to each point of space (the circle coordinate y identified with $y + 2\pi R$). A massless scalar field $\Phi(x, y)$ obeys the wave equation (2-02), and since it is periodic in y it can be expanded in a Fourier series (1-11).

$$(\partial_t^2 - \nabla^2 - \partial_y^2)\Phi = 0, \qquad \Phi(x,y) = \frac{1}{\sqrt{2\pi R}} \sum_{n\in\mathbb{Z}} \phi_n(x)\, e^{iny/R} \quad (2)$$

Substituting, ∂_y^2 is replaced by $-n^2/R^2$, and the equation splits coefficient by coefficient.

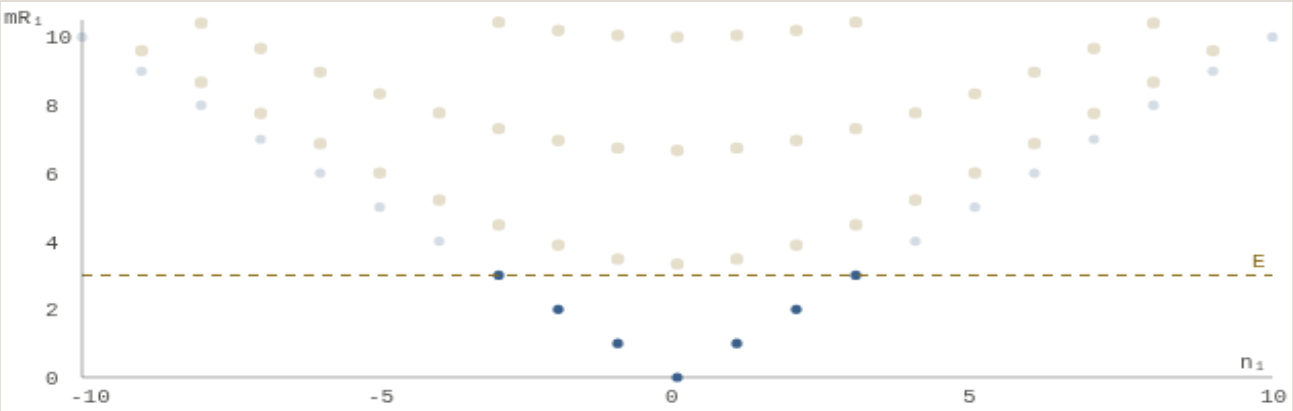
$$\left(\partial_t^2 - \nabla^2 + \frac{n^2}{R^2}\right)\phi_n = 0, \qquad m_n = \frac{|n|}{R} \quad (3)$$

This is the Klein–Gordon equation of mass m_n (2-02 § 06). One five-dimensional field looks, in four dimensions, like a massless **zero mode** ϕ_0 and an infinite sequence of fields of mass $1/R, 2/R, \ldots$. This is the **Kaluza–Klein tower** (KK tower) Known . The PDG writes 'At energies smaller than 1/R, the KK modes can be neglected, leaving the zero-mode action'. The couplings of the zero mode are diluted; for a quartic interaction

$\lambda_4 = \lambda_5/(2\pi R M_5)$ (PDG formula 84.6), weaker the larger the circle. M_5 is the fundamental mass scale of the five-dimensional theory, put in to make λ_5 dimensionless.

For a product of δ circles (a torus T^δ with radii R_i), each tuple of integers n_i gives a field with $m^2 = \sum_i n_i^2/R_i^2$. Figure 1 is the case of two circles.

Figure 1 The KK tower of two circles ($R_1 \geq R_2$). Each point is a four-dimensional field (horizontal n_1, vertical the mass). Varying E (dashed) and the ratio of radii changes the fields below E



E·R₁ = 3.00, R₂/R₁ = 0.30, fields below E: 7 (Weyl estimate 8.5), the thin direction is not yet visible

Table 1 The number of four-dimensional fields of mass at most E (including the zero mode). In brackets the Weyl estimate $\pi R_1 R_2 E^2$ (the area $4\pi^2 R_1 R_2$ times $E^2/4\pi$; the detour of 1-10)

R_2/R_1	$ER_1 = 2$	**5**	**10**	**30**
1	13 (12.6)	81 (78.5)	317 (314.2)	2821 (2827.4)
0.1	5 (1.3)	11 (7.9)	23 (31.4)	267 (282.7)
One circle	5	11	21	61

Computed The numbers of Table 1 and Figure 1 were computed on this machine (the figure is drawn on the spot from formula (3)). If JavaScript is disabled, see Table 1.

The row $R_2/R_1 = 0.1$ is the same as one circle for $ER_1 \leq 5$, and the thin direction first appears at $E \gtrsim 1/R_2$. The count grows as E to the power of (the number of visible extra dimensions).

03

The lowest rung of the tower — the first eigenvalue and the diameter

The lowest rung $1/R$ can be restated as an inequality. For a periodic function with mean 0, $f = \sum_{n\neq 0} c_n e^{iny/R}$, Parseval's identity gives (Wirtinger's inequality) Known

$$\int_0^{2\pi R} |f'|^2\, dy = 2\pi R \sum_{n\neq 0} \frac{n^2}{R^2}|c_n|^2 \;\geq\; \frac{1}{R^2}\int_0^{2\pi R} |f|^2\, dy \qquad (4)$$

This is the simplest example of the Poincaré inequality of 2-05 § 08 and 1-10 § 07, and the best constant R^2 is the reciprocal of the gap $1/R^2$ (the square of the lowest rung of the tower).

Measuring lengths — metric, diameter, Laplacian

To replace the circle by a general compact manifold K (1-05) we need a way of measuring lengths. An inner product given on each tangent space (1-05 § 06), varying smoothly from point to point, is called a **metric**. It is a way of measuring lengths and angles. With a metric we can measure the lengths of curves, and the length of the shortest path joining two points is their distance. The maximum of that distance over all pairs of points is the **diameter** d (for a circle of circumference $2\pi R$, $d = \pi R$; the d of this chapter is the diameter, not the spacetime dimension d of other chapters). The size of the gradient of a function, $|\nabla f|$, is also fixed by the metric, and the operator $-\Delta_K$ with $\int_K |\nabla f|^2 = \int_K \bar{f}\,(-\Delta_K f)$ is called the **Laplacian** (the Laplace–Beltrami operator). On a flat torus, in coordinates y_i measured by length, $-\Delta_K = -\sum_i \partial^2/\partial y_i^2$.

If spacetime is the unwarped product $\mathbb{R}^{1,3} \times K$ and the field is a scalar, expanding in the eigenfunctions of $-\Delta_K$ (eigenvalues $0 = \lambda_0 < \lambda_1 \le \cdots$) gives $m_j^2 = \lambda_j$. The first KK mass is the minimum of the Rayleigh quotient (1-10 § 06).

$$m_1^2 = \lambda_1(K) = \inf\left\{ \frac{\int_K |\nabla f|^2}{\int_K |f|^2} \;:\; \int_K f = 0,\ f \not= 0 \right\} \tag{5}$$

For form or spinor fields, or in warped backgrounds, it becomes the eigenvalue of another operator (the Hodge Laplacian, the Dirac operator and so on). In this chapter we restate "extra dimensions are invisible" as the Poincaré inequality on the part without the zero mode, $\int |\nabla f|^2 \ge \lambda_1 \int |f|^2$.

When the diameter suffices — Ricci curvature

Multiplying the metric by s^2 (multiplying all lengths by s) multiplies λ_1 by s^{-2}, so shrinking raises the tower. That "small" may be measured by the diameter d alone is a theorem when the Ricci curvature is non-negative.

Ricci curvature, as a metaphor, is the average of how much a bundle of geodesics (locally shortest paths) leaving a point around one direction gathers together as it goes. On a sphere it is positive (meridians gather at the poles), on a flat torus 0, on a saddle-shaped surface negative (they spread). More precisely, for a direction v of length 1, it is the sum $\mathrm{Ric}(v,v) = \sum_i K(v, e_i)$ of the bending (sectional curvature) of the $n-1$ sections spanned by v and the directions e_i orthogonal to v, which is $n-1$ times the average. In two dimensions there is a single section, so it is the same as the Gauss curvature. The bending itself is the curvature in the sense of 1-07 § 05 (the

discrepancy after going round a small loop) when the bundle of tangent spaces is given the connection fixed by the metric. $\mathrm{Ric}_K \geq 0$ is the condition that this sum is at least 0 at every point in every direction Known .

$$\mathrm{Ric}_K \geq 0 \implies \lambda_1(K) \geq \frac{\pi^2}{d^2} \qquad \text{(Zhong–Yang; equality only for the circle: Hang–Wang)} \,($$

Table 2 The ratio $\lambda_1 d^2/\pi^2$. (6) says it is at least 1 Computed

K	Ratio	K	Ratio
Circle	1	Square torus	2
Rectangular torus 1×0.2	1.04	Hexagonal torus	1.7778
Rectangular torus 1×0.05	1.0025	Spheres S^2, S^3, S^6	2, 3, 6

The tori were computed numerically and agree to 5 digits with the closed forms (rectangle $1+\rho^2$, hexagon $16/9$). The spheres are known values.

For thin tori the ratio approaches 1 but does not reach it. The curvature condition cannot be dropped. Join two spheres by a thin neck and put into (5) a function that is $+1$ on one and -1 on the other: the numerator comes only from the neck, so making the neck thinner leaves the diameter nearly unchanged while the Rayleigh quotient becomes as small as we like. At the base of the neck the curvature becomes strongly negative, and $\mathrm{Ric} \geq 0$ fails. This is a well-known example Known .

04

Kaluza and Klein — charge is momentum around the circle

The idea of extra dimensions goes back to Kaluza and Klein in the 1920s (the introduction of the PDG). Five-dimensional spacetime also has a metric in the sense of § 03 (in spacetime, what corresponds to the inner product at each point has, like the $\eta = \mathrm{diag}(-1,1,1,1)$ of 2-04, exactly one negative sign, and it may vary from point to point). Its 15 components split into 10 for the four-dimensional metric $g_{\mu\nu}$, 4 for a vector A_μ and 1 for a scalar ϕ. Rewriting formulas 5 and 8 of the review by Overduin–Wesson with signature $(-,+,+,+,+)$ gives[Note 1]

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu + \phi^2\left(dy + \kappa A_\mu dx^\mu\right)^2, \qquad \kappa = \sqrt{16\pi G} \quad (7)$$

The **Einstein equations** are the equations of gravity that tie the bending (curvature) of the metric to the energy of matter. Here we do not go into their content and use only the counting of components. If no component depends on y, the five-dimensional Einstein equations split into 10+4+1=15 four-dimensional equations (ibid. formula 6). Setting ϕ constant, among the 15 the equations for the metric and vector components become the four-dimensional Einstein equations and Maxwell's equations (ibid.

formula 7) Known . But the remaining scalar equation is compatible with $\phi =$ constant only when $F_{\mu\nu}F^{\mu\nu} = 0$ (pointed out by Jordan and Thiry, ibid. § 3.4). The change of coordinates $y \mapsto y + \kappa\lambda(x)$ induces $A_\mu \mapsto A_\mu - \partial_\mu\lambda$, and the gauge transformation (2-03) appears as a re-choice of the coordinate in the direction of the circle.

Klein (1926) connected Kaluza's theory with wave mechanics (abstract) and made the circle small. The momentum n/R of the n-th component in (2) becomes, under (7), a charge (ibid. formula 25, $\phi = 1$).

$$q_n = \frac{n\sqrt{16\pi G}}{R}, \qquad n \in \mathbb{Z} \quad (8)$$

Charge is restricted to integer multiples of q_1 (the title of Klein's short communication of the same year is 'The Atomicity of Electricity as a Quantum Theory Law'). As a metaphor, charge is momentum around an invisible circle; more precisely, it is the momentum n/R times a constant κ fixed by convention.

Setting $q_1 = e$ ($e^2/4\pi = \alpha \approx 1/137.036$) gives $R = 2\ell_P/\sqrt{\alpha} \approx 3.8 \times 10^{-34}$ m (about 23 times the Planck length), and the first rung of the tower is $1/R \approx 5.2 \times 10^{17}$ GeV, about 10^{21} times the electron mass Computed . Overduin–Wesson, taking the radius as the Planck length, estimate the discrepancy at about 22 orders of magnitude, and write that it 'played a large role in the abandonment of five-dimensional Kaluza-Klein theory'. If the internal space has a non-abelian symmetry, non-abelian gauge fields can come out (17-07) Physics .

05

The six dimensions of strings — curling up into Calabi–Yau

Spacetime in superstring theory is ten-dimensional (17-05 § 06), so to make it look four-dimensional a six-dimensional internal space K is needed. Candelas–Horowitz–Strominger–Witten (1985) studied vacua leaving one supersymmetry in four dimensions, and wrote in the abstract 'This condition permits only a few possibilities'. The K satisfying the condition is said to be a Kähler manifold with Ricci curvature 0 (a complex manifold whose metric and complex structure mesh) whose holonomy (1-07) lies in SU(3): a **Calabi–Yau manifold** Physics .

That such metrics exist is a mathematical theorem. Yau announced in 1977 'We announce a proof of Calabi's conjectures on the Ricci curvature of a compact Kähler manifold' (the paper is from 1978). As a consequence, a compact Kähler manifold with vanishing first Chern class has a Ricci-flat Kähler metric Known . If it is Ricci-flat, (6) applies, so the scalar KK tower on a Calabi–Yau of diameter d starts above π/d.

On the size, the introduction of the PDG writes 'The extra dimensions were supposed to be compactified at a scale close to the Planck scale, and thus not testable experimentally in the near future'. The shape affects the particle content (the model of CHSW has 'an E6 gauge group with four standard generations of fermions').

06

Confining to a brane — ADD and RS

In 1998, Arkani-Hamed–Dimopoulos–Dvali (ADD) proposed a way of hiding in which the fields of the Standard Model are confined to a three-dimensional membrane (a 3-brane) and only gravity spreads into $4+\delta$ dimensions. Since they can be probed only by gravity, the extra dimensions may be large. The abstract reads 'The observed weakness of gravity on distances $\gtrsim 1$ mm is due to the existence of n ≥ 2 new compact spatial dimensions large compared to the weak scale' Physics . The weakness of gravity is the same dilution as the λ_4 of § 02, and the fundamental scale M_D and the reduced Planck mass $M_P \approx 2.4 \times 10^{18}$ GeV are related by (PDG formula 84.10)

$$M_P^2 = R^{\delta} M_D^{2+\delta} \quad (9)$$

Table 3 R fixed by (9) (reproducing the "10^9 km, 0.5 mm, ..., 0.1 MeV^{-1}" of PDG formula 84.11) Computed

δ	$M_D = 1$ TeV	$M_D = 4$ TeV
1	1.2×10^{12} m	1.8×10^{10} m
2	4.8×10^{-4} m	3.0×10^{-5} m
6	2.6×10^{-14} m	4.2×10^{-15} m

$\delta = 1$ is the size of the solar system, 'clearly ruled out' (PDG), and $\delta = 2$ gives 0.5 mm. The next year, 1999, Randall–Sundrum (RS) used a slice of five-dimensional anti-de Sitter space (PDG formula 84.16).

$$ds^2 = e^{-2ky}\eta_{\mu\nu}dx^{\mu}dx^{\nu} + dy^2, \qquad 0 \leq y \leq \pi R \quad (10)$$

Anti-de Sitter space is a spacetime of constant negative curvature (17-08). The metric (10) is the slice $0 \leq y \leq \pi R$ of that spacetime with curvature radius $1/k$. Scales at the end $y = \pi R$ are redshifted by a factor $e^{-k\pi R}$. The PDG writes that $k \sim M_P$ and $R \simeq 11/k$ give the TeV; at $kR = 11$, $ke^{-k\pi R} \approx 2.4$ TeV, and changing kR from 10 to 12 moves it from 55 TeV to 0.1 TeV Computed . Making a large ratio with an exponential resembles dimensional transmutation (17-01). More precisely, in RS the exponent comes from geometry (kR), in dimensional transmutation from the running of the coupling ($1/g^2$). The view that connects warped extra dimensions with strongly coupled four-dimensional theories is AdS/CFT (17-08), at the level of a conjecture Physics .

Detour: extra dimensions without a gap

A subsection you may skip.

The paper of RS from the same year argued that four-dimensional gravity can be reproduced even if the extra dimension is infinitely long: 'even without a gap in the Kaluza-Klein spectrum, four-dimensional Newtonian and general relativistic gravity is reproduced to more than adequate precision' Physics . A gap in the tower is not essential for invisibility.

07

Experimental lower bounds — Newton's law and colliders

Deviations from Newton's law are by custom expressed in Yukawa form (PDG formula 84.12).

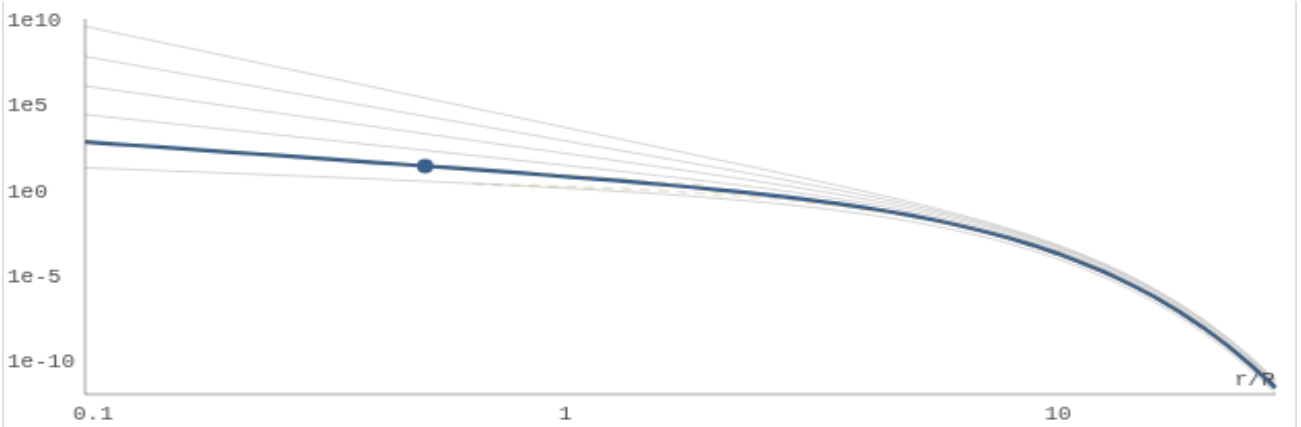
$$V(r) = -G_N \frac{m_1 m_2}{r}\left(1 + \alpha\, e^{-r/\lambda}\right) \quad (11)$$

The exchange of a field of mass m gives a Yukawa potential e^{-mr}/r, and one term is added for each KK mode. In a scalar (Newtonian) model, summing the whole tower of T^δ (radius R) gives

$$\frac{V(r)}{V_{\mathrm{N}}(r)} = \sum_{k \in \mathbb{Z}^\delta} e^{-|k|\, r/R} \xrightarrow{\;r \gg R\;} 1 + 2\delta\, e^{-r/R} \quad (12)$$

2δ is the number of fields on the first rung; for $\delta = 1$ the sum is $\coth(r/2R)$. For $r \ll R$, $V \propto r^{-1-\delta}$: the law of a space of $3+\delta$ dimensions. For the graviton the coefficient changes. The KK modes of the graviton are massive spin-2 fields, and the contribution of their polarisations is said to multiply the coefficient of the first rung by 4/3; the PDG takes, for a two-dimensional torus, $\alpha = 16/3$ (4/3 of the 4 in (12)) and $\lambda = R$ Physics . (12) is not the formula for gravity itself but a model for seeing how summing the tower produces a Yukawa-type deviation.

Figure 2 $V/V_{\mathrm{N}} - 1$ of (12) (log scale). Varying δ and r/R changes the deviation and the dimension of space that can be read off. The dashed lines are two approximations




```
δ = 2  r/R = 0.501  V/V_N = 25.13, readable dimension of space ≈ 5.0 (3 far away, 5)
```

Table 4 The ratio V/V_{N} of (12)

δ	$r/R = 0.1$	**1**	**10**
1	20.02	2.164	1.000091
2	628.3	6.507	1.000184
6	3.72×10^9	3721	1.000593

Computed The numbers of Figure 2 and Table 4 were computed on this machine (sums over $|k| \le 420$; $\delta = 1$ agrees with $\coth$ to 12 digits, and $r \to 0$ with the integral approximation to 5 digits). If JavaScript is disabled, see Table 4.

That the deviation vanishes as $e^{-r/R}$ for $r \gg R$ is the quantitative content of "invisible".

Table 5 Lower bounds from PDG 2025 (lengths converted with (1) Computed)

What was searched for	Lower bound
Inverse-square law (torsion balance)	For $\delta = 2$, $R < 30$ μm, $M_D > 4.0$ TeV
Supernova SN1987A / neutron stars	For $\delta = 2$, $M_D > 27$ / 1700 TeV
Graviton emission at the LHC (ATLAS / CMS)	$M_D > 5.9$–11.2 / 5.5–10.7 TeV ($\delta = 2$–6)
The lightest KK graviton of RS	ATLAS 4.5 TeV ($k/M_P = 0.1$), CMS 2.2–5.6 TeV
Gauge fields propagating in flat extra dimensions	$1/R > 4.16$ TeV (a reinterpretation by ATLAS; 4.7×10^{-20} m), $\gtrsim 6$ TeV from LEP2
All fields propagate (UED)	$1/R > 1.4$–1.5 TeV. The minimal model is excluded together with dark matter

One of the inverse-square-law experiments (Tan et al. 2016) confirms $|\alpha| \le 1$ down to $\lambda = 59$ μm (abstract). The observed gauge fields are four-dimensional down to about 5×10^{-20} m, and gravity follows the inverse-square law down to some tens of μm. Extra dimensions have not been found, but neither has it been shown that there are none.

```
08
```

Back to Yang–Mills — the KK gap is not the mass gap

The KK gap is not the mass gap. The lowest rung of the tower is a gap that comes only from the internal space being compact; free fields have it too, and the zero mode has no gap. It is the same computation as the gap of a free field in a box (2-02 § 06, 2-05 § 06). What the prize problem asks for is a gap arising from interaction on $\mathbb{R}^4$, which has no compact directions (4-11), and Jaffe–Witten write 'no present ideas point the direction to establish the existence of a mass gap that is uniform in the volume'.

Do not mix the two kinds of extra dimensions. The lower bounds of § 07 concern real extra dimensions. The extra dimensions of holographic models (17-08, 17-09) (the radial direction of anti-de Sitter space, S^5, the auxiliary circle of Witten's model) are tools of a dual description or of regularisation, and are not claimed to be real dimensions.

Within the models the tower is visible. In models that read the gap from geometry, the KK tower of the extra dimensions comes at the same height as the glueballs (in the Maldacena–Núñez model of $\mathcal{N} = 1$ supersymmetric Yang–Mills, if the gravity approximation is valid, 'comparable to the masses of Kaluza Klein states'; the authors also write that quantising the string on this background might allow the KK states to be separated) Physics . Removing the tower requires shrinking the extra dimensions, and then the coupling becomes weak (in four dimensions, because of asymptotic freedom), which is the opposite side from the region where the computation works (Csáki et al., of Witten's three-dimensional model: 'This is the opposite of the limit that is required for the supergravity description to be valid'). The order of limits is treated in 17-10.

Seam

What can be said as a theorem about "invisible" is only about linear fluctuations on a fixed internal space ((4)–(6)). Within the range we searched we found no result justifying that the tower separates in an interacting theory, in particular in the limit that returns to four dimensions.

Note 1 Overduin–Wesson take the four-dimensional signature $(+,-,-,-)$ and set $\kappa \equiv 4\sqrt{\pi G}$ (formula 8). (7) is aligned with this book's signature $(-,+,+,+)$, and $\kappa = \sqrt{16\pi G}$ is the same value. Back to text

What this chapter can say

What is said	What is not said
Known (2)–(8), Yau's theorem.	General lower bounds for towers in warped backgrounds or of forms and spinors. The consistency of keeping ϕ constant (the scalar equation requires $F_{\mu\nu}F^{\mu\nu} = 0$).
Computed Tables 1–5, Figures 1 and 2, Klein's radius, reproducing the numbers of the PDG with (9).	(12) is a scalar model. The 16/3 for the graviton is quoted from the PDG; the origin of the factor 4/3 is not derived in this chapter.
Physics The Calabi–Yau condition, ADD and RS, non-abelian gauge fields, the height of the towers in the models.	Whether real extra dimensions exist. That the towers of the models separate in the limit returning to four dimensions. This book does not itself show either the mass gap or the construction of the continuum limit (the published construction of the continuum limit: Volume 2).

Sources and reproduction

Item	Kind	Source / tools
(2)(3), (9)–(11), 16/3, Table 5, quotations	Known, Physics	Z. Demiragli, A. Pomarol, "84. Extra Dimensions" (PDG, PRD 110 (2024) 030001, revised August 2025). Text checked.
(7)(8), Kaluza, Klein	Known	J. M. Overduin, P. S. Wesson, Phys. Rep. 283 (1997) 303, gr-qc/9805018 § 3, § 4 (text checked; formulas 5–8 and 25, the condition for constant ϕ in § 3.4). Th. Kaluza, Sitzungsber. Preuss. Akad. Wiss. (1921) 966 (bibliographic data only). O. Klein, Z. Phys. 37 (1926) 895 (abstract checked); Nature 118 (1926) 516 (bibliographic data only).
(6)	Known	J. Q. Zhong, H. C. Yang, Sci. Sinica Ser. A 27 (1984) 1265 (bibliographic data only; the claim follows the survey by Ling–Lu, whose bibliographic data is not checked). F. Hang, X. Wang, IMRN 2007, rnm064, doi:10.1093/imrn/rnm064 (bibliographic data only).
The Calabi–Yau condition	Physics	P. Candelas, G. T. Horowitz, A. Strominger, E. Witten, NPB 258 (1985) 46. Abstract checked (the term Calabi–Yau and SU(3) holonomy not checked in the text).
Yau's theorem	Known	S.-T. Yau, PNAS 74 (1977) 1798 (abstract checked); CPAM 31 (1978) 339 (bibliographic data only).
ADD, RS, the inverse-square law	Physics	N. Arkani-Hamed, S. Dimopoulos, G. Dvali, PLB 429 (1998) 263. L. Randall, R. Sundrum, PRL 83 (1999) 3370; ibid. 4690. W.-H. Tan et al., PRL 116 (2016) 131101. Abstracts checked.
Quotations in § 08	Known, Physics	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory" (the Clay problem description; the PDF published by Clay; text checked). J. Maldacena, C. Núñez, PRL 86 (2001) 588, arXiv:hep-th/0008001 (abstract checked). C. Csáki, H. Ooguri, Y. Oz, J. Terning, JHEP 9901 (1999) 017, arXiv:hep-th/9806021 (text checked).

Item	Kind	Source / tools
Tables, figures, numbers	Computed	Python 3 (numpy) on this machine: kk_count.py, zhong_yang.py, units_klein.py, add_rs.py, newton_kk.py (the values embedded in Figure 2 are the same computation output to 4 decimal places).

Next chapter to read: 17-07 Where gauge fields come from — extracting gauge groups from branes and geometry.

Where gauge fields come from — the group is read off from the number of branes, a lattice, the shape of a singularity

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 — Four mechanisms by which, on the side of extra dimensions, the gauge group becomes an output (symmetries of the internal space, stacked branes, an internal lattice of the string, collapsing spheres). That the positions of branes become eigenvalues of matrices, and that ADE comes out of the linear algebra of "negative definite". The road down to pure Yang–Mills, and the expectations it relies on.

Prerequisite chapters — 17-06 How extra dimensions become invisible. We use roots (1-04), holonomy (1-07) and the central charge (17-05). Intersection numbers and lattices are defined on the spot.

Said first — All the mechanisms are at the level of physics, and neither string theory nor M-theory has a mathematical definition. That a group comes out of geometry and that the four-dimensional quantum Yang–Mills theory exists and has a mass gap are different questions. This book does not itself show either the mass gap or the construction of the continuum limit (the published construction of the continuum limit is read in Volume 2).

The order of this chapter

1. Is the group an input or an output? — four sources
2. Symmetries of the internal space — non-abelian Kaluza–Klein
3. Strings on branes — N branes give U(N)
4. Positions become matrices — eigenvalues are brane positions
5. The heterotic string — a 16-dimensional lattice (Detour: the same tower, different groups)
6. Compactifying on Calabi–Yau — the commuting part

7. Collapsing spheres and ADE — negative definiteness
8. Folding and geometric engineering
9. From eleven dimensions to pure Yang–Mills — two circles and an expectation
10. Back to Yang–Mills — the group is not where the difficulty lies

01

Is the group an input or an output? — four sources

In four-dimensional Yang–Mills theory the gauge group G is an input chosen by hand. The text of the prize problem (Jaffe–Witten) too says 'for any compact simple gauge group G', allowing any of the list in 1-04 (0-03). In theories with extra dimensions the group becomes an output, read off from something else Physics .

Symmetry	Number of branes	Lattice	Singularity
Isometries of the internal space § 02	N D-branes give U(N) § 03–04	The 16 left-over dimensions of the heterotic string § 05–06	The arrangement of collapsing spheres gives ADE § 07–08

What any of these roads fixes is the set of fields and the group seen at low energy, not the existence of a four-dimensional quantum theory. Of the four, three (branes, lattice, singularity) are stories within supersymmetric settings (theories with a symmetry exchanging bosons and fermions), and what that means for pure Yang–Mills we look back on in § 10.

02

Symmetries of the internal space — non-abelian Kaluza–Klein

In 17-06 § 04 the rotation of the circle became U(1). If the group $\mathrm{Isom}(K)$ of transformations of the internal space K that do not change distances (isometries) is large, each of its infinitesimal generators (Killing vector fields) gives one gauge field, and the commutators of the generators become the structure constants (1-03) Physics . For the generators of rotations of a sphere, $L_a = \varepsilon_{abc} x_b \partial_c$,

$$[L_a, L_b] = -\varepsilon_{abc}\, L_c, \qquad \dim SO(n+1) = \frac{n(n+1)}{2} \quad (1)$$

(the left formula was checked numerically Computed); S^n gives $n(n+1)/2$ gauge fields, and $\mathbb{CP}^2$ (the set of all complex lines through the origin of $\mathbb{C}^3$; real dimension 4) gives the 8 gauge fields of SU(3) Known . An example of a seven-dimensional space whose isometry group is locally (as a Lie algebra) SU(3)×SU(2)×U(1) (3-07) is $\mathbb{CP}^2 \times S^2 \times S^1$ ($4+2+1=7$ dimensions), and together with four dimensions the whole is $4+7=11$-dimensional.

Witten (1981) studied the road of compactifying eleven-dimensional supergravity on seven-dimensional spaces, and wrote in the abstract 'It is possible to obtain an SU(3) × SU(2) × U(1) gauge group, but the proper fermion quantum numbers are difficult to achieve'. The seven-dimensional spaces Witten studied are not necessarily the same as the example above.

03

Strings on branes — stacking N gives U(N)

Impose on open strings a boundary condition attaching their ends to a subspace, and that subspace is a **D-brane** (Polchinski 1995: 'extended objects defined by mixed Dirichlet-Neumann boundary conditions in string theory') Physics . One with p spatial dimensions is called a Dp-brane. Labelling the two ends of a string with brane numbers $1,\dots,N$, the states have $N\times N$ matrices as coefficients, and the massless vectors form the adjoint representation of U(N) (1-04) (Polchinski's lectures of 1996: 'this global symmetry of the world-sheet theory is elevated to a gauge symmetry in spacetime'). There are N^2 kinds of strings, and $\dim U(N) = N^2$.

Separating the branes makes the strings between them heavy (Witten 1995b § 3; T is the string tension, a_i the positions of the branes).

$$m_{ij} = T\,|a_i - a_j| \quad (2)$$

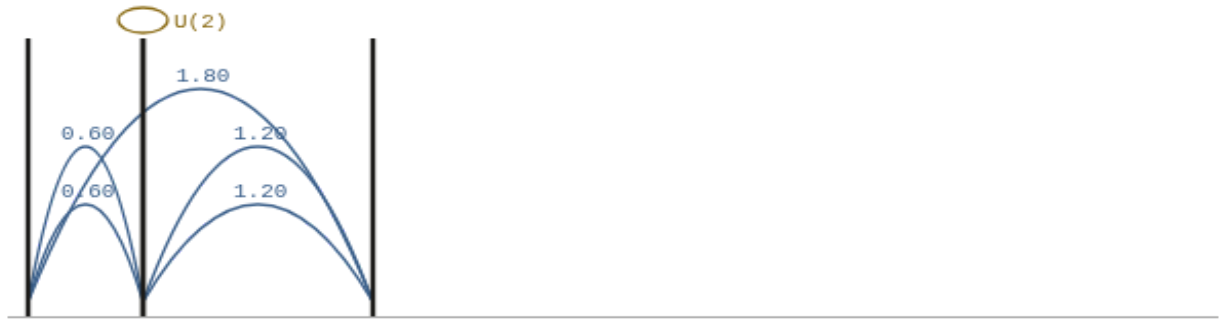
If they are stacked $n_1, n_2, \dots$ at a time, the remaining group is $U(n_1)\times U(n_2)\times\cdots$, and

$$\#\{\text{massless vectors}\} = \sum_k n_k^2, \qquad \#\{\text{heavy W}\} = N^2 - \sum_k n_k^2 \quad (3)$$

(the same count appears in Polchinski's lectures). Separating branes is the Higgs mechanism seen from the side of geometry. More precisely, when the adjoint scalars X^i of § 04 take diagonal values $\mathrm{diag}(a_1,\dots,a_N)$, the off-diagonal gauge fields acquire the masses (2). This mechanism, in which the values of scalars make gauge fields heavy, is called the Higgs mechanism (3-07) Physics .

Figure 1 Varying the spacings d_1, d_2, d_3 of four branes (vertical lines) changes the remaining group and the W masses. Arcs are heavy strings (numbers are mass/T); the loops at the top are strings between stacked branes. A

heavy string has two orientations ($i \to j$ and $j \to i$), so each arc, each mass, counts as 2




```
Group U(2)×U(1)², massless 6, heavy 10, W masses/T (2 each): 0.60, 0.60, 1.20, 1.20, 1.80
```

Table 1 How four branes are stacked, and the numbers in (3) (a heavy string has two orientations, so 2 for each W mass)

Stacking	4	3+1	2+2	2+1+1	1+1+1+1
Remaining group	U(4)	U(3)×U(1)	U(2)×U(2)	U(2)×U(1)²	U(1)⁴
Massless / heavy	16 / 0	10 / 6	8 / 8	6 / 10	4 / 12

Computed Table 1 was counted on this machine, and Figure 1 is drawn on the spot from (2)(3). If JavaScript is disabled, see Table 1.

The W masses of four separated branes are d_1, d_2, d_3 and their consecutive sums, 6 in all: the relation between the 3 simple roots and the 6 positive roots of SU(4) (A_3). In § 07 the same structure appears as the areas of spheres.

04

Positions become matrices — the theory on the branes

The theory on stacked branes is 'the dimensional reduction of ten-dimensional supersymmetric Yang-Mills theory with gauge group U(n) from ten to p+1 dimensions' (Witten 1995b § 3) Physics . Splitting the components of the ten-dimensional gauge field into directions along the branes and perpendicular to them gives

$$A_M = (A_0, \dots, A_p;\, X^{p+1}, \dots, X^9), \qquad V = -\frac{T^2}{2} \sum_{i,j} \operatorname{tr}\left([X^i, X^j]^2\right) \;\geq 0 \quad (4)$$

where X^i are the $9-p$ scalars in the adjoint representation ($N \times N$ Hermitian matrices), and V is their potential. Since X is Hermitian, $[X^i, X^j]$ is anti-Hermitian and $\operatorname{tr}[X^i, X^j]^2 \leq 0$. Therefore $V \geq 0$. If $V = 0$, the matrices can be diagonalised simultaneously, and Witten reads the diagonal entries as the positions of the branes. The off-diagonal entries are the strings between branes, and the eigenvalues of the map $\delta \mapsto \sum_k [X^k, [X^k, \delta]]$ are $|a_\lambda - a_\mu|^2$ (times T^2, the square of (2)). We checked agreement within 4×10^{-15} on an example of two directions conjugated by a unitary matrix Computed .

On D3-branes the theory is four-dimensional $\mathcal{N} = 4$ U(N) supersymmetric Yang–Mills with 6 scalars and fermions (the starting point of 17-08), not pure Yang–Mills. The matrix indices have the same structure as the double lines of 17-04.

05

The heterotic string — a 16-dimensional lattice becomes a group

The heterotic string is a closed string whose right-movers follow the ledger of the superstring (central charge 15) and whose left-movers follow that of the bosonic string (26) (17-05 § 06, Gross–Harvey–Martinec–Rohm 1985). The left-movers are 16 short of the 10 of flat ten dimensions; Polchinski's lectures (1994), on the supersymmetric heterotic string, write 'an additional (0, 16) is needed from another unitary CFT', and go on to say that the only solutions are $E_8 \times E_8$ or SO(32) Physics .

A 16-dimensional lattice

In the lattice picture this 16 is seen as a torus $\mathbb{R}^{16}/\Gamma$ on which 16 left-moving bosons are wound Physics . The momentum P lies on the lattice Γ, and the states with $P \cdot P = 2$ become massless vectors. On the heterotic string wound on a torus, Witten (1995 § 4.6) writes 'The gauge symmetry is extended precisely when $V^\perp$ contains some P of $d_P = 1$'. Witten's $V^\perp$ is the space of left-moving charges. Witten takes the sign of the inner product opposite to this chapter and writes $d_P = -(P,P)/2$, so $d_P = 1$ corresponds to $P \cdot P = 2$ in this chapter's sign. Γ must be **even** (every $P \cdot P$ even) and **self-dual** (the points having integer inner product with all points are exactly Γ). This condition is said to come from the requirement that the one-loop amplitude, whose world sheet is a torus, not depend on how the shape of the torus is chosen (modular invariance), and Polchinski's lectures also write 'The constraints of modular invariance are strong'.

$$\dim G = 16 + \#\{P \in \Gamma : \ P \cdot P = 2\} \quad (5)$$

There are only two even self-dual lattices in 16 dimensions Known . Both can be written as "$x \in \mathbb{Z}^n \cup (\mathbb{Z} + \frac{1}{2})^n$ with $\sum x_i$ even": two copies of $n = 8$ side by side give $E_8 \oplus E_8$, and $n = 16$ gives D_{16}^+. Both have 480 roots, and (5) is 496, but joining roots with non-zero inner product, the former splits into two sets of 240 ($E_8 \times E_8$) and the latter is connected ($\mathfrak{so}(32)$) Computed . The ten-dimensional group is written precisely as $\mathrm{Spin}(32)/\mathbb{Z}_2$. $\mathrm{Spin}(32)$ is the double cover of SO(32), the same relation as SU(2)→SO(3) in 1-02 (the $\mathbb{Z}_2$ divided out is a different element from the one that returns to SO(32)). The lattice spanned by the roots of D_{16}^+ is a sublattice of index 2 in D_{16}^+, and this $/\mathbb{Z}_2$ is said to correspond to that difference.

Seen from eleven dimensions, Hořava–Witten **proposed** that the two E_8 sit one on each of the walls at the two ends of the segment $S^1/\mathbb{Z}_2$ (the 1995 abstract says 'We propose'; the 1996 abstract writes that the group is fixed to 'a copy of E8 for each boundary component', explicitly presupposing 'the hypothesis that there is a consistent quantum M-theory') Physics .

Detour: the same tower, different groups

A subsection you may skip.

The eigenvalues of the Laplacian on a flat torus $\mathbb{R}^n/\Gamma$ are $4\pi^2\, v\cdot v$ for each point v of the dual lattice (the n-dimensional version of 17-06 § 02) Known ; if the lattice is self-dual, the number on each rung of the KK tower is the number of points of squared length $2k$. Counting, $E_8 \oplus E_8$ and D_{16}^+ both give 480, 61920 and 1050240 for $k = 1, 2, 3$, agree on every rung with $k \le 6$, and also agree with $480\,\sigma_7(k)$ (σ_7 is the sum of the 7th powers of the divisors) Computed .

It is known that they agree on every rung, and Milnor (1964) is said to have given these two tori as an example of being non-isometric although their eigenvalues agree including multiplicities Known . The tower is the same, but the groups carried by the heterotic string differ.

06

Compactifying on Calabi–Yau — holonomy and the commuting part

To bring $E_8 \times E_8$ down to four dimensions, compactify on a six-dimensional Calabi–Yau manifold K (17-06 § 05). The holonomy of K (1-07) lies in SU(3), and in the standard construction the holonomy of the gauge field is said to be aligned with an SU(3) inside E_8 too Physics . A gauge transformation visible in four dimensions is a g constant over all of K that is unchanged by parallel transport along every closed path, that is, it must commute with every element of the holonomy Known .

$$G_4 = \{\, g \in E_8 :\ gh = hg \ \text{ for all } h \in SU(3) \,\} \quad (6)$$

Of the 240 roots of E_8, 72 are orthogonal to the roots of SU(3), giving E_6 (dimension 78) Computed . The abstract of Candelas–Horowitz–Strominger–Witten (1985) writes 'a model that in four dimensions has an E6 gauge group with four standard generations of fermions'.

Table 3 The part of E_8 that commutes with SU(n) (248 = remaining group + SU(n) + remainder)

Holonomy	SU(2) (K3)	SU(3) (Calabi–Yau)	SU(4)	SU(5)
Remaining group	E_7	E_6	SO(10)	SU(5)
Breakdown	133+3+112	78+8+162	45+15+188	24+24+200

Computed Table 3 was counted by building the 240 roots of E_8 in coordinates; the types were fixed by rank, number of roots and the determinant of the Cartan matrix.

The remaining 162 are $\mathbf{2}\times\mathbf{27}\times\mathbf{3}$ (pairs of the 27-dimensional representation of E_6 and the 3-dimensional one of SU(3), and their conjugates), and the four-dimensional fermions are said to come from here.

07

Collapsing spheres and ADE — the linear algebra of negative definiteness

A **K3 surface** is a Calabi–Yau manifold of complex dimension two (real dimension four), with holonomy SU(2) (the left end of Table 3) Known . Witten (1995 § 4.6), on the correspondence between the heterotic string wound on a torus and the IIA string wound on a K3 surface, writes 'Whenever the heterotic string on a four-torus gets an enhanced gauge group G, the corresponding K3 gets an orbifold singularity of type G' Physics . Then 'a configuration of r two-spheres S_i (with an intersection matrix given by the Dynkin diagram of G) collapses to a point' (Dynkin diagrams are defined later in this section), and each sphere gives one U(1). The W bosons are two-dimensional membranes wrapped on the collapsing spheres (as described in Bershadsky–Sadov–Vafa 1996 § 1). Their mass is said to be the tension of the membrane times the area of the sphere, the same form as the $T|a|$ of § 03.

Intersection numbers and negative definiteness

Why only ADE? Look at the **intersection numbers** $M_{ij}=S_i\cdot S_j$ of the spheres. For two spheres this is the sum over intersection points counted with orientation signs ±1. The intersection number of a sphere with itself (the self-intersection number) counts in the same way the intersection points with a slightly shifted copy, and for spheres in K3 it is -2 Known . Defining the "inner product" of classes of spheres by this intersection number, on K3 the inner product is positive in three directions and negative in nineteen (signature $(3,19)$) Known . According to Witten, in the collapsing limit the classes of the spheres lie in the subspace $V^{\perp}$ on which this inner product is negative definite, so M is negative definite ($x^{\mathsf T}Mx<0$ for every $x\neq0$). The collapsing spheres S_i can be taken to be **holomorphic curves** (curves of complex dimension one, locally the zero sets of holomorphic functions of complex coordinates), and intersection points of distinct holomorphic curves all count as positive because of the complex structure, so $M_{ij}\geq0$ Known . With $x=e_i+e_j$, $-4+2M_{ij}<0$ gives $M_{ij}\leq1$. Hence $M=-(2I-A)$, where A is the adjacency matrix of the graph whose vertices are the spheres and whose edges are the intersections. This graph is called the **Dynkin diagram** (those that appear here are of "simply laced" type, all edges single). We need the following.

$$x^{\mathsf{T}}(2I - A)\,x = 2\sum_i x_i^2 - 2\sum_{\text{edges } ij} x_i x_j \;>\; 0 \qquad (x \;/= 0) \quad (7)$$

If the graph is not connected we consider each component (the group is a product over components). If a connected graph has a cycle, the vector equal to 1 on the cycle makes the left-hand side at most 0, so it must be a tree. Examining the 201 trees with at most 10 vertices, only types A (a straight line), D (forked at one end) and E_6, E_7, E_8 remained (Table 4) Computed . That these are the only ones for any number of vertices is known as part of the classification of simple Lie algebras (1-04) Known .

Figure 2 Choosing a type changes the arrangement of collapsing spheres (circles are spheres, lines are intersections), the smallest eigenvalue of (7) and the dimension of the group. "Fold" shows by colour the orbits of the permutations of § 08



E₈: 8 spheres, 240 roots, group dimension 248, smallest eigenvalue 0.010956 (positive definite)

Table 4 The number of trees with n vertices, and the types satisfying (7) (for $n \le 3$ only the straight line)

n	4	5	6	7	8	9	10
Number of trees	2	3	6	11	23	47	106
Types satisfying it	A D	A D	A D E	A D E	A D E	A D	A D

Computed The numbers of Table 4 and Figure 2 were computed on this machine. The smallest eigenvalue agrees to 6 digits with $2 - 2\cos(\pi/h)$ (h is the number of roots ÷ the rank), and the figure is drawn with this formula. One beyond E_8 there is an integer vector making (7) 0 (centre 6, arms 3 / 4, 2 / 5, 4, 3, 2, 1). If JavaScript is disabled, see Tables 4 and 5.

08

Folding and geometric engineering — A to G are all there

ADE are only the types whose roots have equal lengths. B, C, F and G come out when the singularities are further bundled in another direction. Bershadsky et al. (1996 § 3) write that going once around over another parameter can permute the collapsing spheres by an automorphism of the Dynkin diagram (a permutation of vertices preserving edges), and that then 'the actual gauge group will be smaller', listing the following (the last is the cyclic permutation of three legs) Physics .

$$A_{2n-1} \to C_n, \qquad D_n \to B_{n-1}, \qquad E_6 \to F_4, \qquad D_4 \to G_2 \quad (8)$$

With this every type from A to G comes out of geometry. The number of orbits into which the permutation bundles the positive roots agrees with the number of positive

roots of the folded algebra (Table 5) Computed .

Table 5 Folding: number of orbits of positive roots = number of positive roots of the algebra after

Before → after	$A_5 \to C_3$	$D_5 \to B_4$	$E_6 \to F_4$	$D_4 \to G_2$
Dimension	35 → 21	45 → 36	78 → 52	28 → 14
Positive roots (fixed) → orbits	15 (3) → 9	20 (12) → 16	36 (12) → 24	12 (3) → 6

Computed Table 5 was counted by generating roots from the Cartan matrices ($A_3 \to C_2$, $A_7 \to C_4$ and $D_6 \to B_5$ also agree).

But the same paper notes that 'a given A-D-E singularity may *not* correspond to an enhanced gauge symmetry'. Computing field theories by this mechanism is **geometric engineering**; Katz–Klemm–Vafa (1997) reduced the mathematically established results of four-dimensional $\mathcal{N} = 2$ (17-11) to the T-duality of strings (strings wound on a circle of radius R and on a circle of radius α'/R give the same theory; α' is from (1) of 17-05), and write 'This construction is not based on any duality conjecture' Physics . What it can handle is supersymmetric theories with at least 8 supercharges (17-05 § 07) (corresponding to four-dimensional $\mathcal{N} = 2$ or more); within the range we searched we found no example handling the mass gap of non-supersymmetric pure Yang–Mills by geometric engineering.

09

From eleven dimensions to pure Yang–Mills — two circles

Witten (1998 § 4) describes a road from eleven dimensions down to four-dimensional **pure** Yang–Mills Physics . The starting point is the six-dimensional (2,0) theory on M5-branes (membranes of M-theory extending in five spatial dimensions) ((2,0) is the name of a type of six-dimensional supersymmetry; Witten calls it 'exotic'), which 'can apparently be interpreted' as M-theory on $AdS7 \times S^4$ ($AdS7$ is a seven-dimensional spacetime of constant negative curvature, treated in 17-08) (according to Maldacena's conjecture). No description by a Lagrangian is said to be known, and there is no mathematical definition either. This is wrapped on two circles.

$$(2,0) \xrightarrow{C_1(R_1)} \text{5-dim. } SU(N),\ g_5^2 = R_1 \xrightarrow{C_2(R_2)} \text{4-dim. } SU(N),\ g_4^2 = \frac{R_1}{R_2} \quad (9)$$

Making the fermions antiperiodic (changing sign once around) on C_2, the fermions acquire masses of order $1/R_2$, the scalars too 'very plausibly' become heavy at one loop, and 'If this is so, the low energy theory will be the pure SU(N) theory without supersymmetry'. In terms of the IIA string this is N D4-branes wrapped on $\mathbb{R}^4 \times S^1$, and the theory on the branes of § 03–04 comes down to four dimensions.

Witten writes that he would like to keep $\eta = g_4^2 N$ small, so that the coupling is weak at the scale of the circles and strong at much longer distances. A one-loop rough guide:

putting $b_0 = 11N/(48\pi^2)$ (formula (1) of 17-01), $\mu = 1/R_2$ and $g^2(\mu) = g_4^2 = \eta/N$ into (3) of 17-01 gives

$$\Lambda R_2 \;\sim\; \exp\!\Big(-\frac{24\pi^2}{11\,\eta}\Big) \quad (10)$$

which is 4×10^{-10} for $\eta = 1$ and 0.12 for $\eta = 10$ Computed . The coefficient changes with the convention for the coupling and the choice of $\mu = 1/R_2$, so this is a guide to the order of magnitude. But supergravity can compute only 'at least if η is large', and Witten writes of connecting to small η 'This looks like a tall order'. The order of limits is in 17-10, and how the gap appears as geometry in 17-09.

10

Back to Yang–Mills — the group is not where the difficulty lies

On every road the group comes with extra fields or supersymmetry (the $9-p$ scalars on the branes, the fields of mass $1/R_2$ from the two circles, and so on) Physics . And for the prize problem, the group is not where the difficulty lies. The problem statement allows any compact simple group; what is asked is that its quantum Yang–Mills theory exists on $\mathbb{R}^4$ satisfying the axioms (4-10) and has a mass gap (4-11). Geometry can produce all the names from A to G, but within supersymmetric settings, and the road to pure Yang–Mills (§ 09) relies on expectations and on the order of limits.

Seam

(1) Every construction stands on string theory or M-theory, neither of which has a non-perturbative mathematical definition. (2) Trying to define a four-dimensional theory from the side of extra dimensions makes the definition of the starting point ((2,0) theory and so on) the harder part. (3) Within the range we searched we found no result connecting the group coming out of geometry with the existence of a four-dimensional quantum theory with a gap.

What this chapter can say

What is said	What is not said
Known (1), the commutation of (6), that the diagrams satisfying (7) are limited to ADE, the signature and positivity of the intersection numbers of K3, the classification of 16-dimensional lattices and Milnor's example (the last two: bibliographic data only).	Proofs of the classification and of the properties of intersection numbers are not written.
Computed Tables 1–5, Figures 1 and 2, the eigenvalues around (4), the numbers of (10).	Lattice point counts up to $k \leq 6$. (10) is a one-loop rough guide, and its coefficient may change with convention.
Physics The U(N) of D-branes, the groups of the heterotic string, E_6, the proposal of Hořava–Witten, ADE and folding, geometric engineering, the pure SU(N) of (9).	None is a mathematical theorem. The existence of string theory, M-theory and the (2,0) theory. That the scalars become heavy at one loop. This book does not itself show either the mass gap or the construction of the continuum limit (the published construction of the continuum limit: Volume 2).

Sources and reproduction

Item	Kind	Source / tools (text = the text of the manuscript checked, abstract = the abstract che
§ 01, § 10	Known	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory", the official problem descripti the Clay Mathematics Institute (text).
§ 02	Physics	E. Witten, NPB 186 (1981) 412 (abstract; the choice of seven-dimensional spaces checked).
§ 03, § 04	Physics	J. Polchinski, PRL 75 (1995) 4724, hep-th/9510017 (abstract); TASI Lectures on D-Branes, hep-th/9611050 (text). E. Witten, NPB 460 (1996) 335, hep-th/9510135 § 3 (text).
§ 05	Physics	D. J. Gross and 3 others, PRL 54 (1985) 502; NPB 256 (1985) 253 (abstracts). J. Polchinski, "What is String Theory?" (lectures of 1994), hep-th/9411028 § 2.6 (tex Hořava, E. Witten, NPB 460 (1996) 506; NPB 475 (1996) 94 (abstracts).
§ 05, § 07	Physics	E. Witten, NPB 443 (1995) 85, hep-th/9503124 § 4.6 (text). M. Bershadsky, V. Sadc Vafa, NPB 463 (1996) 420, hep-th/9511222 § 1 (text; the formula mass = tension × area for the membrane not checked).
Intersection numbers in § 07	Known	E. Witten, hep-th/9503124 § 4.6 (text; the squared length of the class of a sphere -2, the collapsing spheres are holomorphic). The general theory of the signatur the intersection form of K3 and the positivity of intersections of holomorphic cu W. Barth and 3 others, *Compact Complex Surfaces*, Springer (bibliographic data only).
Lattices, detour	Known	J. H. Conway, N. J. A. Sloane, *Sphere Packings, Lattices and Groups*. J. Milnor, PN 51 (1964) 542. Both bibliographic data only.

Item	Kind	Source / tools (text = the text of the manuscript checked, abstract = the abstract che
§ 06	Physics	P. Candelas and 3 others, NPB 258 (1985) 46 (abstract; the construction aligning holonomies not checked).
§ 08	Physics	M. Bershadsky and 5 others, NPB 481 (1996) 215, hep-th/9605200 § 3 (text). S. Ka Klemm, C. Vafa, NPB 497 (1997) 173, hep-th/9609239 (abstract).
§ 09	Physics	E. Witten, ATMP 2 (1998) 505, hep-th/9803131 § 4 (text). (10) is the arithmetic of putting $g^2 = \eta/N$ into (1)(3) of 17-01 ($b_0 = 11N/(48\pi^2)$).
Tables, figures, numbers	Computed	Python 3 (numpy) on this machine: branes.py, ade.py, e8_commutant.py, lattice scales.py.

Next chapter to read: 17-08 Holography — the $\mathcal{N} = 4$ theory on branes, and the correspondence in which the radial direction becomes scale.

Holography — reading a four-dimensional field theory through the geometry of gravity in one more dimension

> 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 — AdS/CFT is the conjecture that a four-dimensional field theory and a theory of gravity in five dimensions (times a sphere) are the same. That depth from the boundary measures scale, the dictionary linking mass and dimension, Wilson loops computed with a hanging string, and where the line falls for the correspondences that can be written exactly.

Prerequisite chapters — 17-05. We also use the 't Hooft coupling λ of 17-04 and the Nambu–Goto action of 17-03. For D-branes see 17-07. Tools not in the prerequisite chapters, such as the metric, geodesics and sources, are defined on the spot, one line each.

Said first — AdS/CFT is a conjecture, not a theorem. The representative example, $\mathcal{N} = 4$ supersymmetric Yang–Mills theory, is conformal and has no mass gap; it is not pure Yang–Mills theory. There is no example in which holography has shown the mass gap as a mathematical theorem (models with a gap at the level of physics are in 17-09), and this book itself shows nothing about the mass gap or the construction of the continuum limit (the published construction of the continuum limit is read in Volume 2).

The order of this chapter

1. What is claimed (Detour: why "hologram")
2. The two faces of D3-branes
3. Anti-de Sitter space
4. Radial direction = scale — depth measures size (Figure 1)
5. The dictionary — mass and dimension
6. Hanging a string — Wilson loops (Figure 2)

7. Putting in a single scale — a model of the area law
8. Correspondences that can be written exactly — theorems and proposals
9. The distance to Yang–Mills

01

What is claimed — the boundary theory and gravity in the interior

The claim of holography is this. **A field theory without gravity in d dimensions describes, in other words, the same physics as a theory with gravity in $d+1$ dimensions.** The extra dimension is "depth" from the boundary; the former is called the boundary theory and the latter the interior (bulk) theory. The representative example is Maldacena (1997, ATMP 1998; the first year is when it appeared on arXiv, the second the year of the journal).

$$\mathcal{N}=4 \text{ SYM } \left(U(N),\, g_{\mathrm{YM}}\right) \quad \overset{\text{conjecture}}{\Longleftrightarrow} \quad \text{type IIB string theory on } AdS_5\times S^5 \ \left(N,\, g_s\right) \ (1)$$

On the left is four-dimensional $\mathcal{N}=4$ supersymmetric Yang–Mills theory (SYM below), on the right ten-dimensional string theory. **Type IIB** is one of the ten-dimensional superstring theories, differing from the type IIA of 17-05 in the combination of handedness (chirality) of its supersymmetries. The dimensions are 10 = 4 (boundary) + 1 (depth) + 5 (S^5) (17-05 § 09). The original paper writes 'We are led to the conjecture', and (1) is still a conjecture Physics . This chapter deals with the entrance; models in which the mass gap turns into geometry are left to 17-09.

Detour: why "hologram" — counting by area

A subsection you may skip.

The entropy of a black hole is proportional not to its volume but to the area A of its horizon (Bekenstein 1973, Hawking 1975) Known .

$$S/k_B = A/(4\ell_P^2), \qquad \ell_P^2 = G\hbar/c^3 \ (2)$$

For a solar mass the radius is 2953 m and $S/k_B = 1.05\times 10^{77}$, far less than the 2.6×10^{115} obtained by dividing the volume by the cube of the Planck length Computed . From here 't Hooft (1993) and Susskind (1994, J. Math. Phys. 1995) set up the idea of counting degrees of freedom by area, and Susskind wrote 'much like a holographic image' Physics . As a metaphor, it is a hologram recording a three-dimensional image on a flat plate. More precisely, (1) is a claim that two quantum theories are equivalent, and the "plate" is at the infinity of spacetime.

02

The two faces of D3-branes — where the conjecture came from

Open strings ending on N stacked D3-branes (membranes with three spatial dimensions on which the ends of open strings sit) become $U(N)$ gauge fields at low energy (17-07). Seen from gravity, the same branes curve spacetime. Maldacena read that in the low-energy limit the theory on the branes and the spacetime just near the branes describe the same thing Physics . The **metric** (line element) ds^2 is the formula giving the squared distance when the coordinates shift slightly, and the length of a curve is $\int ds$ (for flat space $ds^2 = dx_1^2 + dx_2^2 + \cdots$). The metric made by the branes is

$$ds^2 = f^{-1/2} dx_{\parallel}^2 + f^{1/2}\left(dr^2 + r^2 d\Omega_5^2\right), \qquad f = 1 + \frac{4\pi g_s N \alpha'^2}{r^4} \quad (3)$$

Here $x_{\parallel}$ are the four coordinates along the branes (in $dx_{\parallel}^2$ only the time direction has the opposite sign; 2-04), r is the distance from the branes, $\alpha' = \ell_s^2$ is the square of the string length, and g_s is the string coupling. As $r \to 0$, $f \approx R^4/r^4$ ($R^4 = 4\pi g_s N \alpha'^2$), and setting $z = R^2/r$ gives

$$ds^2 = \frac{R^2}{z^2}\left(dx_{\parallel}^2 + dz^2\right) + R^2 d\Omega_5^2 \quad (4)$$

the product of AdS_5 of radius R and S^5 Known . There are two ways of writing the correspondence of couplings. In the convention widely used today $g_{\mathrm{YM}}^2 = 4\pi g_s$ and $R^4/\alpha'^2 = 4\pi g_s N = \lambda$ ($\lambda = g_{\mathrm{YM}}^2 N$); in the convention of the original paper, $g_{\mathrm{YM}}^2 = 2\pi g_s$, $R^4/\alpha'^2 = 2\lambda$ Known . The difference is only in the normalisation of the action; today's convention is the normalisation in which the instanton action is $8\pi^2/g_{\mathrm{YM}}^2$ (the review of Aharony–Gubser–Maldacena–Ooguri–Oz (3.8), § 4.2). (2) of 17-04 ($\frac{1}{2g^2}\int \mathrm{tr}\, F^2$ with matrix traces, $\lambda = g^2 N$) has the same normalisation, so with the λ of 17-04, $R^4/\alpha'^2 = \lambda$. In this chapter we do not use λ directly and write $R^2/\alpha' = \sqrt{4\pi g_s N}$.

Supergravity (the field theory with gravity that remains at low energy in string theory) is reliable when $R \gg \ell_s$, that is $g_s N \gg 1$, and quantum effects in the interior correspond to the $1/N$ corrections of the gauge theory (original paper) Physics . Making this compatible with $g_s \ll 1$ requires $N \gg g_s N \gg 1$.

Table 1 The string coupling $g_s = (R/\ell_s)^4/(4\pi N)$ needed for radius R

R/ℓ_s	$N = 3$	$N = 100$
2	0.424	0.0127
3	2.15	0.0645
5	16.6	0.497

R/ℓ_s	$N=3$	$N=100$
10	265	7.96

Example: taking the curvature radius to be 5 times the string length, $N=3$ gives $g_s=16.6$, a strong string coupling, and only at $N=100$ does it come down to 0.5 Computed .

03

Anti-de Sitter space — the boundary is infinitely far away

Anti-de Sitter space AdS_{d+1} is, in a flat space whose metric (§ 02) has signature $(-,-,+,\dots,+)$, the hyperboloid

$$-X_{-1}^2-X_0^2+X_1^2+\cdots+X_d^2=-R^2 \quad (5)$$

and the group $SO(2,d)$ preserving (5) is its symmetry as it stands (appendix of Maldacena). Choosing coordinates again gives the form of the first half of (4) (Poincaré coordinates). Below we consider the Euclidean version, moved to imaginary time (2-04). Writing the boundary coordinates as $x=(x_1,\dots,x_d)$,

$$ds^2=R^2\,\frac{\sum_{i=1}^{d}dx_i^2+dz^2}{z^2}\qquad (z>0) \quad (6)$$

For $d=1$ it is the upper half-plane, and Figure 1 draws this case. Example: the length going up along $x=0$ from $z=1$ to $z=\varepsilon$ is $\int_\varepsilon^1 R\,dz/z=R\ln(1/\varepsilon)$, infinite as $\varepsilon\to 0$. The boundary $z=0$ is infinitely far away Known . The dimension of the symmetry is $6\cdot 5/2=15$ for $SO(2,4)$ (in the Euclidean version $SO(1,5)$, of the same dimension 15; 1-03), and the four-dimensional conformal transformations number 15 (translations 4, Lorentz 6, dilatation 1, special conformal 4); in the words of Witten (1998a), 'SO(2, d) ... acts on M_d as the conformal group' Known .

04

Radial direction = scale — depth measures size

In (6), under $(x,z)\mapsto(sx,sz)$ both numerator and denominator are multiplied by s^2, and ds^2 does not change. This **isometry** of the interior (a transformation preserving length) is, on the boundary, the dilatation $x\mapsto sx$: "looking at something s times larger on the boundary" and "looking s times deeper" become the same operation. In the words of Susskind–Witten (1998), 'scale size gets transmuted into a spatial dimension' Known .

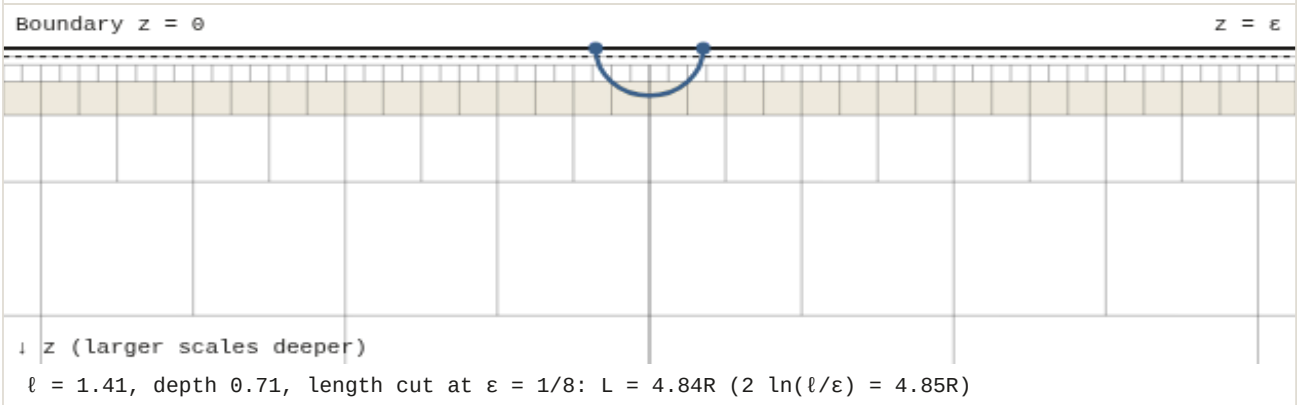
The **geodesics** of (6) (the shortest curves joining two points) are semicircles orthogonal to the boundary, and the semicircle hanging from the boundary interval

$[-\ell/2, \ell/2]$ goes down to depth $\ell/2$. The length cut off at depth ε is logarithmic in ℓ.

$$L(\ell) = 2R \ln \frac{\ell + \sqrt{\ell^2 - 4\varepsilon^2}}{2\varepsilon} \approx 2R \ln \frac{\ell}{\varepsilon}, \qquad e^{-mL} \approx \left(\frac{\varepsilon}{\ell}\right)^{2mR} \tag{7}$$

The amplitude for a particle of mass m to propagate through the interior is roughly e^{-mL}, so exponential decay turns into a power law on the boundary. Susskind–Witten (2.5) follow the same idea, writing it on a model of the sphere in a form without coefficients, $\delta^m/|X_1 - X_2|^m$. The exponent $2mR$ is an approximation for large mR; exactly it is (9) Physics .

Figure 1 The upper half-plane (6) and a dyadic tiling (every tile has area $R^2/2$). Changing the interval ℓ with the knob changes the depth of the semicircle and the L of (7). The coloured band is the level where the bottom lies



Computed The numbers in this figure were computed on this machine.

Table 2 L of (7) ($R = 1$; agrees with numerical integration along the semicircle). Each doubling of ℓ increases L by $2\ln 2 = 1.386$

ℓ/ε	L	$2\ln(\ell/\varepsilon)$
10	4.584863	4.605170
100	9.210140	9.210340
1000	13.815509	13.815511

Cutting near the boundary at $z = \varepsilon$ is an ultraviolet cutoff of the boundary theory. Susskind–Witten write 'the infrared regulator δ in the bulk theory also plays the role of an ultraviolet regulator in the boundary super Yang Mills theory' (δ corresponds to the ε of this chapter). They then divided the cut sphere into a lattice of size δ, assumed one degree of freedom per field per lattice cell, and counted the degrees of freedom as N^2/δ^3 (ibid. (3.1)) Physics . In Figure 1 the width of the tiles doubles at each level down, which recalls one step of the renormalisation group, grouping blocks in pairs (7-05). More precisely, it stands at the level where de Boer–Verlinde–Verlinde (1999, JHEP 2000) derived the form of the Callan–Symanzik equation (the formula describing how correlation functions change when the scale is changed; 4-04) from the Hamilton–Jacobi equation describing radial evolution in classical supergravity (an equation set

up by regarding the action as a function of the values at the end), and proposed 'We propose a direct correspondence' Physics .

05

The dictionary — fields in the interior and operators on the boundary

In the prescription of Gubser–Klebanov–Polyakov (1998) and Witten (1998a), the boundary value ϕ_0 of an interior field ϕ is read as the **source** of a boundary operator $\mathcal{O}$ (an external field added to the action in the form $\int \phi_0 \mathcal{O}\, d^d x$), and the following are set equal. The right end is the action I of the classical solution in the limit where supergravity can be used ($N, \lambda \to \infty$). The abstract of GKP begins 'We suggest'; it is a proposal Physics .

$$\left\langle \exp \int \phi_0\, \mathcal{O}\, d^d x \right\rangle_{\text{boundary}} = Z_{\text{interior}}\left[\phi \to \phi_0\right] \approx e^{-I[\phi_0]} \tag{8}$$

One consequence is the relation between mass and dimension. For solutions independent of x, the equation of a field of mass m is $z^{d+1}(z^{1-d}\phi')' = m^2 R^2 \phi$ (Witten 1998a (2.37)), and putting in $\phi = z^\alpha$ gives $\alpha(\alpha - d) = m^2 R^2$. We write the larger root as Δ (the letter customary in the literature, distinct from the Δ of the mass gap).

$$\Delta(\Delta - d) = m^2 R^2, \qquad \Delta = \frac{d}{2} + \sqrt{\frac{d^2}{4} + m^2 R^2} \tag{9}$$

A solution with $\phi \approx \phi_0 z^{d-\Delta}$ near the boundary extends smoothly into the interior. Under the dilatation $(x, z) \mapsto (sx, sz)$, ϕ_0 is multiplied by $s^{\Delta - d}$, so for $\int \phi_0 \mathcal{O}$ to be invariant the **scaling dimension** of $\mathcal{O}$ (the Δ such that the operator is multiplied by $s^{-\Delta}$ under the dilatation $x \mapsto sx$) must be Δ, and the two-point function is proportional to $|x|^{-2\Delta}$ Physics . The condition for real roots, $m^2 R^2 \geq -d^2/4$, is the Breitenlohner–Freedman bound, which allows negative squared masses down to this point Known .

Table 3 Δ of (9) ($d = 4$). The right end is the power read off by solving the equation numerically from $z = 1$ to e^{-25} (agrees with $d - \Delta$)

$m^2 R^2$	**−3**	**0**	**5**	**12**
Δ	3	4	5	6
Numerical power	1.000000	0.000000	−1.000000	−2.000000

Example: if massless, $\Delta = 4 = d$, and the ϕ_0 in $\int \phi_0 \mathcal{O}\, d^4 x$ is a dimensionless number, that is, it plays the role of a coupling. The dilaton (the scalar whose expectation value fixes the string coupling) corresponds to $\operatorname{tr} F^2$. For the $\mathcal{N} = 4$ theory, Witten (1998a § 2.6) states that the operator obtained by differentiating the Lagrangian density with respect to the coupling (which includes $\operatorname{tr} F^2$) corresponds to the dilaton with $m^2 = 0$, and Csáki–Ooguri–Oz–Terning, citing earlier literature, write 'the operator tr F^2 in four

dimensions corresponds to the dilaton field ...' Physics . $\mathrm{tr}\, F^2$ creates the 0^{++} glueball (17-01), so in 17-09 we read glueball masses from the dilaton equation.

06

Hanging a string — computing the Wilson loop

The energy $E(L)$ of a heavy quark pair at distance L is read from the Wilson loop (5-01). In the prescription put forward at about the same time by Maldacena (1998) and Rey–Yee (1998, EPJC 2001), the quarks are the endpoints of a string ending on the boundary, and the area (Nambu–Goto) of the string hanging into the interior gives E Physics . In (4) the tension $1/(2\pi\alpha')$ is multiplied by R^2/z^2.

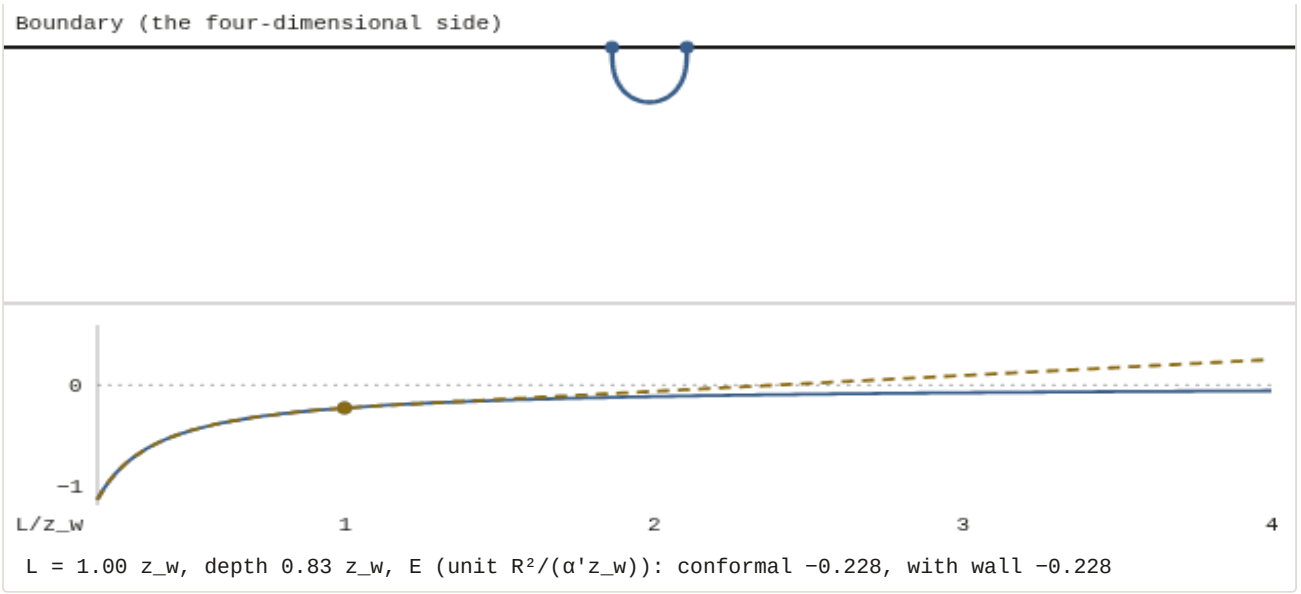
$$E = \frac{1}{2\pi\alpha'}\int_{-L/2}^{L/2} \frac{R^2}{z^2}\sqrt{1+z'^2}\,dx, \qquad L = 2z_0\int_0^1 \frac{y^2\,dy}{\sqrt{1-y^4}} = \frac{2\sqrt{2}\,\pi^{3/2}}{\Gamma(1/4)^2}z_0 = 1.19814\,z_0 \quad (10$$

The integrand does not contain x, so $z^2\sqrt{1+z'^2} = z_0^2$ (z_0 is the depth of the bottom) is conserved, giving the right-hand formula. The string goes down to a depth $z_0 = 0.8346L$ proportional to L. Subtracting the divergence at the boundary (the self-energy of point charges) by the amount for two straight strings gives

$$E(L) = -\frac{4\pi^2}{\Gamma(1/4)^4}\,\frac{R^2}{\alpha'}\,\frac{1}{L} = -0.22847\,\frac{\sqrt{4\pi g_s N}}{L} \quad (11)$$

where the constant agrees between numerical integration and the closed form to 8 digits, and is the same as (4.6)(4.8) of the original paper Computed . The original paper writes 'the energy goes as 1/L, a fact which is determined by conformal invariance. Notice that the energy goes as $(gN)^{1/2}$ as opposed to gN which is the perturbative result' Physics .

Figure 2 Top: a string with both ends on the boundary. Bottom: $E(L)$ (solid line (11); dashed line (12), with a wall). Changing L changes the depth and E, and with "Put a wall" the string lies down along the wall



Computed The numbers in this figure were computed on this machine (the shape of the string from a table of the integral (10), E from (11)(12)).

Table 4 E (unit $R^2/(\alpha' z_w)$) and the depth without a wall. $L_c = 1.19814\, z_w$

L/z_w	Conformal (11)	With wall (12)	$z0/z_w$
0.5	−0.45695	−0.45695	0.4173
1.19814	−0.19069	−0.19069	1.0000
2	−0.11424	−0.06307	1.6693
4	−0.05712	0.25524	3.3385
8	−0.02856	0.89186	6.6770

07

Putting in a single scale — cutting the geometry gives an area law

The $1/L$ of (11) is a consequence of conformality, so confinement ($E \approx \sigma L$, 17-01) needs a scale. The crudest model is to cut the interior at depth z_w. Polchinski–Strassler (2001, PRL 2002) stated that if there is a gap 'the gravitational redshift has a nonzero lower bound' and used the approximation of cutting the geometry Physics . Minimising (10) in this model, for $L \le L_c = 1.19814 z_w$ the conformal solution (bottom depth $0.8346 L \le z_w$) can be used as it is. For $L > L_c$ the bottom of the string reaches the wall. Where it touches the wall the string is tangent to it (slope 0), so the conserved quantity $z^2\sqrt{1+z'^2}$ becomes z_w^2, and the descending part is half of the conformal solution with $z_0 = z_w$. The halves on the two sides use up a width L_c, and the remaining $L - L_c$ is joined by a straight line along the wall.

$$E(L) = -0.19069\,\frac{R^2}{\alpha' z_w} + \sigma(L - L_c), \qquad \sigma = \frac{\sqrt{4\pi g_s N}}{2\pi z_w^2} \qquad (L > L_c) \quad (12)$$

A part proportional to L (an area law) appears, and at $L = L_c$ it joins (11) in both value and slope Computed . Example: at $L = 8z_w$ in Table 4 it is $+0.892$ against the conformal -0.029. As in (12) it has the form $\sigma(L - L_c)$ plus a constant, and for large L it lies on a straight line of slope σ. The string tension carries a factor $\sqrt{4\pi g_s N}$, while the glueball mass fixed by the box up to the wall ($\sim 1/z_w$) does not, so the larger the coupling, the further apart the two scales (the λ-dependence of the ratio and its sources are in 17-09 § 08). This is arithmetic within the model; the actual geometry is treated in 17-09 and 17-10.

08

Correspondences that can be written exactly — what is a theorem and what is a proposal

"Exact" has two meanings: being correct as mathematics closed on one side alone (hyperbolic geometry, Lie groups, ordinary differential equations), and the definitions of both sides agreeing within a model. Neither is a theorem of four-dimensional Yang–Mills theory, and (1) is not proved either.

Table 5 Correspondences and their level

Boundary side	Interior side	Level
Conformal group, dilatation	$SO(2,4)$, $(x, z) \mapsto (sx, sz)$	Known
$\ln(\ell/\varepsilon)$	Length of a geodesic (7)	The formula for the length Known (numbers Computed), its identification with the logarithm on the boundary Physics
Dimension Δ	The power of (9)	The power Known , the identification Physics
The form of Callan–Symanzik	Radial Hamilton–Jacobi	Physics (proposal)
$E(L)$	Minimal area of a string	The prescription Physics , (10)(11) Computed
—	Connectedness of the boundary	Known Mathematical theorem

The last row is a theorem of Witten–Yau (1999). We prepare three terms. An **Einstein manifold** is a manifold whose Ricci curvature (curvature averaged over directions) is a constant multiple of the metric; here the constant is negative (the space (6) is an example). A **conformal boundary** (in Penrose's sense) is a boundary at infinity, attached when multiplying the metric by the square of a function vanishing on the boundary makes it extend smoothly to the boundary (for (6), $z^2 ds^2$ extends to $z = 0$). $H_n(M; \mathbb{Z}) = 0$ means, roughly, that M has no n-dimensional "holes". If M is $n+1$-dimensional, N is its n-dimensional conformal boundary, M is complete (not cut off

partway in any direction), and the scalar curvature of N (curvature further condensed into a single number) is positive, then

$$H_n(M;\mathbb{Z}) = 0, \qquad \text{in particular } N \text{ is connected} \ (13)$$

(abstract: 'Let M be a complete Einstein manifold of negative curvature … N must be connected') Known . What this chapter uses is only the conclusion that "the boundary N is in one piece (connected)". AdS/CFT is the motivation, but it is not a theorem about the duality itself. Example: the Physics rows of Table 5 all use (1) as a premise.

09

The distance to Yang–Mills — conformality, limits, where things stand

The boundary theory of (1) is not pure Yang–Mills theory. It has six scalar fields and fermions, is conformal so its coupling does not run (in contrast with 4-06), and no scale Λ is born (4-07). The $N, \lambda \to \infty$ where the gravity computation works is the opposite side from what the problem statement (0-03) asks for: any compact simple group, and a theory that agrees with asymptotic freedom ($\lambda \to 0$) at short distances. A conformal theory has no gap: if there is a state of mass $m > 0$, dilatation gives states of mass sm (for any $s > 0$) too.

$$m^2 \in \operatorname{spec}(P^2) \ \Rightarrow \ s^2 m^2 \in \operatorname{spec}(P^2) \ (s > 0) \ \Rightarrow \ \inf\left(\operatorname{spec}(P^2) \setminus \{0\}\right) = 0 \ (14)$$

P^2 is the square of the four-momentum (2-04), and the $1/L$ of (11) is another face of this circumstance Known . Example: the Clay page on the problem still writes of the mass gap 'no proof of this property is known'. There is no example in which holography has shown the mass gap as a mathematical theorem (models with a gap at the level of physics are in 17-09); string theory has no non-perturbative definition, (1) is a conjecture, and the limits lie outside the problem statement, all at once (16-05). The road back to four dimensions is in 17-10, and an example where a gap is visible from a duality under supersymmetry in 17-11.

What this chapter can say

What is said	What is not said
Known The geometry of (4)–(6), $SO(2,4)$ and the conformal group, the power and the bound of (9), (13)(14).	A proof of (1). A mathematical definition of (8) (the quantum theory in the interior is undefined).

What is said	What is not said
Computed Tables 1–4, Figures 1 and 2, the constants of (10)(11), the numbers of (2).	The wall of (12) is a model, not the actual geometry of confinement.
Physics (1)(8), the prescription for Wilson loops, the radial direction and the renormalisation group, the wall model.	SYM has no gap. A bridge to pure Yang–Mills, finite N, asymptotic freedom. This book itself shows neither the mass gap nor the construction of the continuum limit (the published construction of the continuum limit is in Volume 2).

Sources and reproduction

Item	Kind	Source / tools
(1)(3)–(5), the quotations in § 02, the convention of the original paper	Physics, Known	J. M. Maldacena, Adv. Theor. Math. Phys. 2 (1998) 231, hep-th/9711200. Text checked.
Today's convention $g_{\mathrm{YM}}^2 = 4\pi g_s$ in § 02	Known	O. Aharony, S. S. Gubser, J. Maldacena, H. Ooguri, Y. Oz, Phys. Rept. 323 (2000) 183, hep-th/9905111. Text of (3.8) and § 4.2 checked.
The quotation in § 03, (2.37), the bound $m^2R^2 \geq -d^2/4$, the dilaton and the coupling operator in § 05	Known, Physics	E. Witten, Adv. Theor. Math. Phys. 2 (1998) 253, hep-th/9802150. Text checked for (2.35)–(2.37), the bound and the real-root condition (just after (2.35) and § 2.6), and the description of the dilaton (§ 2.6).
(8)	Physics	S. S. Gubser, I. R. Klebanov, A. M. Polyakov, Phys. Lett. B 428 (1998) 105, hep-th/9802109. Abstract checked.
The name of the bound	Known	P. Breitenlohner, D. Z. Freedman, Phys. Lett. B 115 (1982) 197; Ann. Phys. 144 (1982) 249. Bibliographic data only.
The idea of (7), the quotations in § 04	Physics	L. Susskind, E. Witten, hep-th/9805114. Text of (2.5), (3.1) and § 2– § 3 checked. J. de Boer, E. Verlinde, H. Verlinde, JHEP 0008 (2000) 003, hep-th/9912012. Text checked.
$\mathrm{tr}\, F^2$ and the dilaton	Physics	C. Csáki, H. Ooguri, Y. Oz, J. Terning, JHEP 9901 (1999) 017, hep-th/9806021. Text checked (the paper attributes this correspondence to earlier literature).
(10)(11), the quotations in § 06	Physics, Computed	J. M. Maldacena, Phys. Rev. Lett. 80 (1998) 4859, hep-th/9803002. Text checked.
The prescription of § 06 (the other one of the same period)	Physics	S.-J. Rey, J.-T. Yee, Eur. Phys. J. C 22 (2001) 379, hep-th/9803001. Abstract checked.
The quotation in § 07	Physics	J. Polchinski, M. J. Strassler, Phys. Rev. Lett. 88 (2002) 031601, hep-th/0109174. Text checked.
(13)	Known	E. Witten, S.-T. Yau, Adv. Theor. Math. Phys. 3 (1999) 1635, hep-th/9910245. Abstract and the opening of the text (M is $n+1$-dimensional, N is n-dimensional) checked.

Item	Kind	Source / tools
Detour	Known, Physics	J. D. Bekenstein, Phys. Rev. D 7 (1973) 2333; S. W. Hawking, Commun. Math. Phys. 43 (1975) 199 (bibliographic data only). G. 't Hooft, gr-qc/9310026; L. Susskind, J. Math. Phys. 36 (1995) 6377, hep-th/9409089 (abstracts checked).
The quotation in § 09	Known	Clay Mathematics Institute, “Yang–Mills & the Mass Gap” (web page). Text checked.
The requirements of the problem statement in § 09	Known	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory” (the official Clay problem description, PDF). For checking against the text, see the table of sources of 0-03.
Tables, figures, numbers	Computed	Python 3 (scipy) on this machine: params.py, geodesic.py, delta.py, wilson.py, bh_entropy.py.

Next chapter to read: 17-09 The mass gap becomes geometry — reading glueball masses as eigenvalues in the geometry of a shrinking circle.

The mass gap becomes geometry — when a circle shrinks and spacetime ends, a gap opens in the spectrum

> 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 argument by which, when spacetime ends in a "cigar" where a circle shrinks, an area law and positive, discrete glueball masses (radial eigenvalues) come out. The numbers retraced on this machine, and the record against the lattice. That the ratios move a lot depending on which scalar is taken as the reference.

Prerequisite chapters — 17-08 (AdS, the dictionary, the string computation). The variational principle is in 2-05 § 08 and 1-10, glueballs in 17-01 § 06.

Said first — The gap here is in the supergravity approximation, $N \to \infty$, $\lambda \to \infty$, and is a linear eigenvalue problem on a fixed background. AdS/CFT is a conjecture, and the models are not pure Yang–Mills theory. This chapter does not solve the mass gap problem, and this book itself shows nothing about it.

The order of this chapter

1. What to translate into geometry
2. Putting in a scale with temperature
3. The circle shrinks (Figure 1, Detour: the temperature of a black hole)
4. The string lies down at the bottom
5. Glueballs are eigenvalues (Figure 2)
6. Why positive, why discrete
7. Comparing with the lattice (Later analysis: which scalar is lightest)
8. Two scales
9. Another way of closing off

01

What to translate into geometry — the gap and the area law

The boundary theory of AdS is conformal and has no gap (17-08 (14)), and in 17-08 § 07 a scale was put in with a wall placed by hand. In this chapter we look at models in which **the geometry itself ends at a finite depth**. Witten (1998) read two things from this Physics. The area law, from the coefficient of the boundary directions dx_i^2 being at least a positive constant in the interior. The gap, from the wave equation in the interior having only positive, discrete eigenvalues $m_1^2 < m_2^2 < \cdots$. Said on the boundary, the latter means that two-point functions such as that of $\mathrm{tr}\, F^2$ become sums of free propagators.

$$\langle \mathcal{O}(x)\, \mathcal{O}(0) \rangle = \sum_{n \geq 1} |c_n|^2 \, \frac{e^{-m_n |x|}}{4\pi |x|} \qquad \text{(the boundary is three-dimensional)} \tag{1}$$

The fraction on the right is the propagator of mass m in three dimensions Known; at long distance the first term dominates, and the correlation length is $1/m_1$ (2-05 § 05, 4-11). Example: looking at the two-point function of $\mathrm{tr}\, F^2$ in the model of § 05 (the lowest dilaton mode), in the normalisation of § 03 where the temperature is $T = 1/\pi$, $m_1 = 3.404$ and the correlation length is 0.294 Computed. But among all the supergravity fields, another scalar coming from fluctuations of the metric is lighter, $m^2 = 5.457$ (the dilaton is 11.588) (Brower–Mathur–Tan 2000; § 07) Physics. (1) is the leading term in $1/N$, and Witten writes that interactions destroy the discreteness but the gap remains.

02

Putting in a scale with temperature — the circle of imaginary time and antiperiodicity

Thermal equilibrium at temperature T is the path integral with imaginary time wound on a circle of period $\beta = 1/T$ (the inverse temperature; the β of this chapter is not the lattice inverse coupling) (2-07 § 06). Bosons return to themselves after one turn, and fermions change sign (antiperiodic), so the momenta in the direction of the circle are

$$\omega_n = \frac{2\pi n}{\beta} \ \text{(bosons)}, \qquad \omega_n = \frac{(2n+1)\pi}{\beta} \ \text{(fermions)}, \qquad n \in \mathbb{Z} \tag{2}$$

Known. Fermions have no $\omega = 0$ and drop out at long distances. Witten put the $\mathcal{N} = 4$ theory on $\mathbb{R}^3 \times S^1$ to break supersymmetry, and since the scalars also become heavy at one loop, held that a three-dimensional gauge field remains at long distances Physics. Csáki–Ooguri–Oz–Terning (1998; COOT below) call this QCD_3. As a metaphor, the

circle is a mould. More precisely, the radius R of the circle is an ultraviolet cutoff of the three-dimensional theory (17-06 § 08). Example: the lowest fermion momentum is πT, and the lowest non-zero boson momentum $2\pi T$.

03

The circle shrinks — the geometry of the cigar

If there is a circle on the boundary, the interior becomes an AdS black hole. Witten scaled that solution to obtain the following metric (the overall R^2 omitted) Physics .

$$ds^2 = g\,d\tau^2 + \frac{d\rho^2}{g} + \rho^2 \sum_{i=1}^{3} dx_i^2, \qquad g = \frac{\rho^2}{b^2} - \frac{b^2}{\rho^2}, \qquad \rho \geq b \quad (3)$$

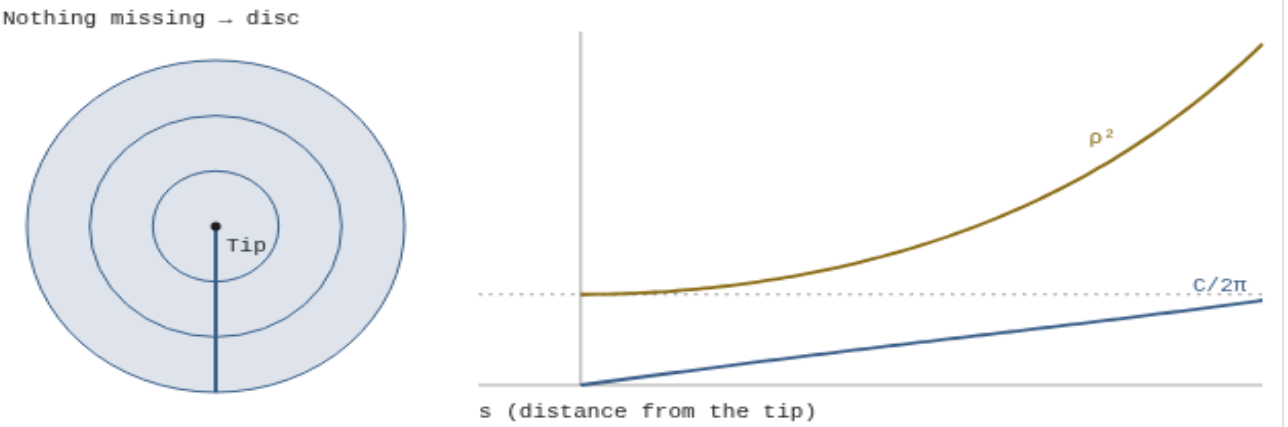
b is the curvature radius, and $\rho \to \infty$ is the boundary, where it returns to AdS (17-08 § 03). At $\rho = b$, $g = 0$ and the circle shrinks. Written with the distance r from the tip ($\rho \approx b + r^2/b$),

$$g\,d\tau^2 + \frac{d\rho^2}{g} \;\approx\; dr^2 + r^2\,d\theta^2, \qquad \theta = \frac{2\tau}{b} \quad (4)$$

these are polar coordinates in the plane. But if the period of τ deviates from πb, the tip is the vertex of a cone (a singularity). Smoothness fixes the period to πb. Near the boundary, dividing the metric by ρ^2 gives $d\tau^2/b^2 + \sum dx_i^2$, so the circumference measured on the boundary is $\pi b/b = \pi$, and the temperature is fixed to $T = 1/\pi$ independently of b Known . The numbers of this chapter are written in this normalisation (circumference π, $T = 1/\pi$). "$b = 1$" in the tables means that b was set to 1; the eigenvalues of (7) do not depend on b, since b disappears on setting $\rho = bu$.

(ρ, τ) is a cigar-shaped disc. As a metaphor, spacetime "ends there". More precisely, the tip is the centre of the disc. For $b = 1$ it can be written $\rho^2 = \cosh 2r$, agreeing with numerical integration to 8 digits Computed .

Figure 1 Left: near the tip (the edges of a sector of angle $2\pi\beta/\beta_1$ identified). Right: the radius of the circle $C/2\pi$ (blue) and the coefficient ρ^2 of dx^2 (yellow). Moving the period away from 1 makes the tip a cone



β/β₁ = 1.00, circumference/(2π×distance) → 1.00, smooth (centre of a disc), minimum of the dx² coefficient b² = 1

Values of the curves of Figure 1 ($\beta/\beta_1 = 1$, $b = 1$; s is the distance from the tip)

ρ	s	$C/(2\pi s)$	ρ^2
1.001	0.0316	0.9997	1.002
1.1	0.3186	0.9719	1.210
2	1.0317	0.9385	4.000
4	1.7324	1.1522	16.00

Computed The numbers in this figure were computed on this machine (the curves are $\rho^2 = \cosh 2s$; the table gives values from numerical integration, and closer to the tip $C/(2\pi s)$ comes closer to 1).

Detour: the temperature of a black hole is the same computation

A subsection you may skip.

If the imaginary-time Schwarzschild metric $f\,d\tau^2 + dr^2/f + \cdots$ ($f = 1 - r_s/r$) is smooth at $r = r_s$, the period is $4\pi r_s$. Reading this as $\hbar c/(k_B T)$ gives

$$T = \frac{\hbar c^3}{8\pi G M k_B} \quad (5)$$

and for a solar mass $T = 6.17 \times 10^{-8}$ K Computed . This derivation is attributed to the method of Gibbons–Hawking (1977) Known .

04

The string lies down at the bottom — the area law

The Wilson loop is fixed as $e^{-\alpha(D)}$ by the regularised area $\alpha(D)$ of a surface in the interior (17-08 § 06) Physics . In (3) the surface can only enter $\rho \geq b$, and the coefficient ρ^2 of dx_i^2 is at least b^2, so with A the enclosed area,

$$\alpha(D) \;\geq\; b^2 A - (\text{boundary and regularisation terms}), \qquad \sigma = b^2 \times (\text{string tension}) \quad (6)$$

and the area law (17-01 § 04) comes out (Witten). We computed the area per unit length $E(L)$ of a string joining two points at distance L by the same method as 17-08 (10).

Table 1 The string in the cigar (unit $R^2/(2\pi\alpha')$, $b = 1$). Conformal is 17-08 (11)

L	$E(L)$	**Conformal** $-1.4355/L$	$E - L$
0.1	−14.355	−14.355	–
1	−1.307	−1.436	−2.307
2	−0.032	−0.718	−2.032
6	4.000	−0.239	−2.000

At short distances it is the conformal $1/L$, at long distances $E \approx L - 2.000$, and the slope 1 is $\sigma = b^2$ Computed . The string goes down to the bottom $\rho = b$ and lies flat. No wall (17-08 § 07) is needed. Example: rescaling the boundary scale by s makes the temperature T/s and multiplies the tension by $1/s^2$, so $\sigma \propto T^2$ (Witten).

05

Glueballs are eigenvalues — the radial equation of the dilaton

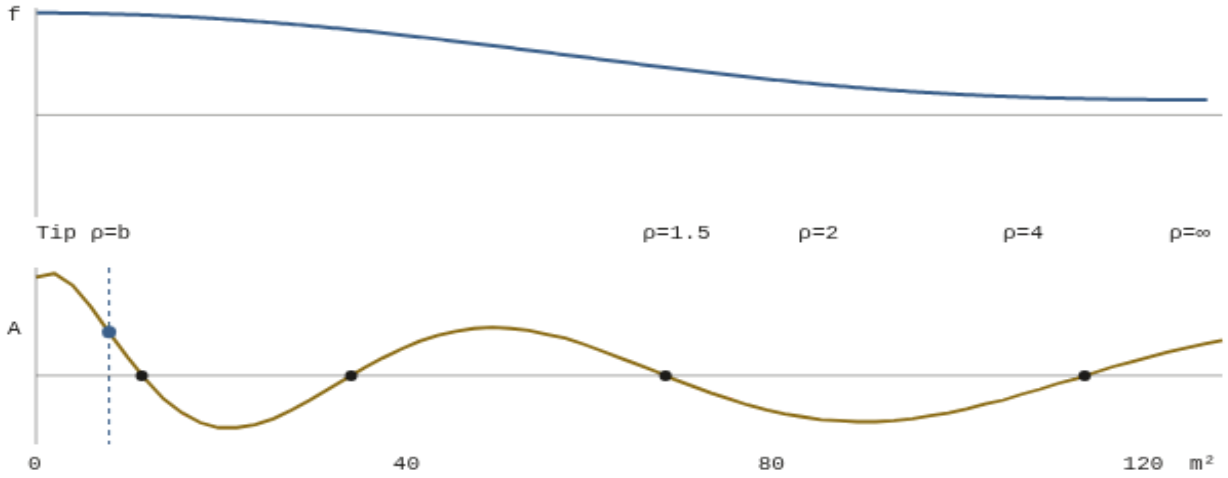
$\operatorname{tr} F^2$ corresponds to the dilaton (17-08 § 05). Witten studied waves $\phi = f(\rho)e^{ik\cdot x}$ on (3) that do not depend on τ (waves with momentum in the τ direction are heavier; a comparison within the same field; § 08). The equation is

$$-\frac{1}{\rho}\frac{d}{d\rho}\left(\rho^3 g \frac{df}{d\rho}\right) = m^2 f, \qquad m^2 = -k^2 \quad (7)$$

(for $b = 1$, $\rho^3 g = \rho^5 - \rho$) Physics , where m is the mass of a three-dimensional particle. There are two conditions. **Smooth at the tip**: one solution diverges as $\log(\rho - b)$, and the smooth solution is unique up to a constant multiple. **Normalisable at infinity**: of the solutions $f \sim 1$ and $f \sim \rho^{-4}$, the latter. The constant component is the source on the boundary (17-08 (8)).

Smoothness at the tip is the condition that the slope measured with the radial coordinate r of polar coordinates in the plane ((4)) is 0. Since $\rho - b \propto r^2$, the slope measured with ρ is not 0; from (7), $f'(1) = -m^2 f(1)/4$ ($b = 1$), which for the first eigenfunction is $-2.90 f(1)$ Computed . The abstract of de Mello Koch et al. (1998) also states this condition: there is no smooth solution whose slope in ρ vanishes at the horizon. The method is shooting: the zeros of the constant component $A(m^2)$ that the smooth solution from the tip leaves at infinity are the eigenvalues.

Figure 2 Top: the solution f shot from the tip (left) (horizontal axis θ, $\rho = \sec\theta$). Bottom: the constant component A left on the boundary (times $1 + m^2/4$). Varying m^2, at $A = 0$ the solution lands on the boundary at 0



m² = 8.0, constant component left on the boundary A = +0.148 (f=1 at the tip), nodes: 0

Computed The numbers in this figure were computed on this machine (the bottom is a table of a precise integration; the top is a Runge–Kutta integration of the same formula in the page, agreeing with the table to 4 digits).

Table 2 The eigenvalues m_n^2 of (7) ($b = 1$). There are $n - 1$ nodes. T_L is from (9)

n	This machine	COOT	$6n(n+1)$	$(\pi/T_L)^2(n+\frac{1}{2})^2$
1	11.5877	11.59	12	12.92
2	34.5270	34.53	36	35.89
3	68.9750	68.98	72	70.34
6	241.237	241.2	252	242.61

Shooting and the Chebyshev collocation method (moved to $u = 1/\rho^2$) agree to 7 digits and reproduce the 6 rows of Table 1 of COOT Computed . The ratio to the temperature is $m_1/T = 10.69$.

06

Why positive, why discrete — the variational principle and the length of the box

Positive. Multiplying (7) by ρf and integrating by parts, the boundary terms at the tip ($\rho^3 g = 0$) and at infinity vanish and

$$m^2 = \frac{\int_b^\infty \rho^3 g\,(f')^2\,d\rho}{\int_b^\infty \rho\, f^2\,d\rho} \tag{8}$$

If the numerator is 0 then f is constant, but a constant is not normalisable ($\int^\infty \rho\,d\rho = \infty$), so $m^2 > 0$ Known . The minimum of the right-hand side is m_1^2, and the gap is the

best constant of a Poincaré inequality (1-10, 2-05 § 08).

Discrete. Changing the variable to $t = \int_1^\rho d\rho'/\sqrt{\rho'^4 - 1}$ and the unknown function to $\psi = (\rho^6 - \rho^2)^{1/4} f$ (a Liouville transformation; $b = 1$), (7) becomes the Schrödinger equation on the interval $[0, T_L]$

$$-\frac{d^2\psi}{dt^2} + V(t)\,\psi = m^2 \psi, \qquad T_L = \int_1^\infty \frac{d\rho}{\sqrt{\rho^4 - 1}} = \int_0^1 \frac{ds}{\sqrt{1 - s^4}} = 1.31103 \quad (9)$$

Known . $V = (\rho^6 - \rho^2)^{-1/4} \frac{d^2}{dt^2} (\rho^6 - \rho^2)^{1/4}$ diverges at both ends: $V \approx -1/(4t^2)$ at the tip, $V \approx (15/4)/(T_L - t)^2$ at the end at infinity Computed . Even so, the length of the box T_L is finite, so the eigenvalues are discrete, just as for a particle in a box.

The type of each end also fixes the phase of the approximation. At an end where the potential is $(\nu^2 - \frac{1}{4})/u^2$ (u the distance from the end), the allowed solution is $\sqrt{u}\, J_\nu(mu)$ (J_ν the Bessel function of order ν), which away from the end oscillates like $\cos(mu - \nu\pi/2 - \pi/4)$. The tip has $\nu = 0$ and brings in a phase $\pi/4$, the end at infinity has $\nu = 2$ and brings in $5\pi/4$, so the n-th eigenvalue is $m_n T_L \approx \pi(n-1) + \pi/4 + 5\pi/4 = \pi(n + \frac{1}{2})$, that is $m_n \approx (\pi/T_L)(n + \frac{1}{2})$. At $n = 6$ the difference shrinks to 0.6% in m^2 (0.3% in m) Computed . The $6n(n+1)$ of Table 2 is an approximate formula from exact WKB analysis, which COOT cite as a private communication from Takei–Koike. Exact WKB analysis is a theory that handles mathematically divergent series expanded in a large parameter by summing them with Borel's method (4-12), and this formula assumes their summability (COOT note that there is no proof in this case).

In AdS the length $\int_0 d\rho/\rho^2$ down to the lower end $\rho = 0$ is infinite, solutions oscillate deep inside, and the spectrum is continuous from 0 (Witten). "Spacetime ends" here means "the Liouville length is finite". In the five-dimensional models of Gürsoy–Kiritsis–Nitti (2007) too, it is stated, by analysis within the model, that for the exponent α of the growth of the potential deep inside, confinement and a gap hold for $\alpha \geq 1$, and a purely discrete spectrum for $\alpha > 1$ (at $\alpha = 1$, a continuous spectrum above the gap). The same paper also has the sentence 'confinement and discrete spectra imply each other' Physics .

07

Comparing with the lattice — three and four dimensions

What supergravity produces are numbers in units of a single scale, so COOT matched the lowest dilaton state to the 0^{++} of the lattice (5-08) (in three dimensions, the lattice $N \to \infty$ value 4.065, printed as 4.07 in the table of COOT) and compared the rest. 0^{--} comes from the equation of a two-form field.

Table 3 Three dimensions (unit $\sqrt{\sigma}$). Lattice: Teper (values cited by COOT)

State	Lattice $N=3$	Lattice $N\to\infty$	COOT	This machine
0^{++*}	6.52(9)	6.18(13)	7.02	7.02
0^{++**}	8.23(17)	7.99(22)	9.92	9.92
0^{--}	6.48(9)	5.91(25)	6.10	6.10

Taking the dilaton as the reference, the ratio $0^{--}/0^{++} = 1.50$ agrees with the lattice 1.45 ± 0.03 to 3%, and the excited states come out 14% and 24% higher than the $N\to\infty$ lattice Computed . Which scalar to take as the reference is treated in "Later analysis" below.

Four dimensions is the D4-brane background of 17-07 §09 (radius R_2 of the antiperiodic circle). The equation of the dilaton-type scalar of COOT, with $y^2=\rho$ (the λ of the original paper) and in units placing the tip at $y=1$ (then $M_{\mathrm{KK}}=1/R_2=3$), is the following (the solutions at infinity are 1 and y^{-9}) Physics .

$$(y^7-y)\,f''+(10y^6-4)\,f'+m^2y^3f=0 \quad (10)$$

Table 4 Four dimensions (GeV, 0^{++} matched to 1.61). The right end is in units of $M_{\mathrm{KK}}=1/R_2$; 0^{++} and 0^{++*} are values from formula (10) of COOT, 0^{-+} the value from the three-form equation of COOT

State	Lattice (values cited by COOT)	COOT	This machine	m/M_{KK}
0^{++}	1.61(15)	Input	Input	2.03
0^{++*}	2.8 (no error)	2.38	2.38	3.01
0^{-+}	2.19(32)	1.83	1.93	2.44

$0^{-+}/0^{++}$ is 1.14 in COOT and 1.36 ± 0.32 on the lattice of the time. COOT obtain 0^{-+} from a component of the 11-dimensional three-form field. Solving that equation on this machine, the lowest value is $m^2=53.376$ (1.93 GeV if the 0^{++} of (10) is 1.61 GeV, ratio 1.20), agreeing with the value for the same equation (N4) in Table 2 of Brower–Mathur–Tan (2000; BMT below) Computed . The 1.83 GeV in the table of COOT cannot be reproduced from this equation. For the lattice 0^{++*}, a note says it may be a bound state of two glueballs.

Later analysis — which scalar is lightest

BMT turned all the supergravity fields (metric, three-form, size of the sphere) into radial eigenvalue problems. We list the four-dimensional scalar-type modes. "Assignment" is the quantum numbers J^{PC} read from the operator to which the mode couples.

Table 5 The lowest four-dimensional modes (m^2 in units placing the tip at 1; the right end in units of M_{KK})

Equation	Field of origin	Assignment	This machine	BMT Table 2	m/M_{KK}

Equation	Field of origin	Assignment	This machine	BMT Table 2	m/M_{KK}
S4	The $\tau\tau$ component of the metric	0^{++} (the lightest)	7.3083	7.308	0.901
T4	Metric, dilaton	2^{++}, 0^{++} and others (0^{++} couples to $\mathrm{tr}\, F^2$)	22.0966	22.097	1.567
V4	The τi components of the metric and others	1^{-+}, 0^{-+}	31.9853	31.985	1.885
(10) of COOT	Dilaton	0^{++} (COOT)	37.1699	–	2.032
N4	Three-form	1^{+-}, 0^{+-} (BMT), 0^{-+} (COOT)	53.3758	53.376	2.435

The T4 equation differs from (10) only in the first-order term.

$$(y^7 - y)\, f'' + (7y^6 - 1)\, f' + m^2 y^3 f = 0 \quad \text{(10')}$$

In the later analysis, the dilaton-type 0^{++} coupling to $\mathrm{tr}\, F^2$ is degenerate with 2^{++} at $1.567\, M_{\mathrm{KK}}$, and the lightest 0^{++} is another mode, S4, at $0.901\, M_{\mathrm{KK}}$ Physics . The 2.03 in Table 4 is the value from formula (10) of COOT. In three dimensions too there is a scalar (S3 of BMT, 5.457) lighter than the dilaton (11.588) Computed . BMT write 'the lowest mass scalar comes from the gravitational multiplet, not the dilaton', and call the agreement of ratios to a few per cent 'a misconception'. At the same time they write that the overall pattern of spins and masses corresponds well with the lattice results Physics .

So the ratios move depending on which modes are taken as numerator and denominator. The three-dimensional $0^{--}/0^{++}$ is 2.19 with S3 as the reference. In four dimensions, the three-form value of COOT divided by T4 is 1.55, and with the assignment of BMT (0^{-+} is V4) V4/T4 is 1.20 and V4/S4 is 2.09 Computed . For $SU(3)$ on today's lattice, $0^{-+}/0^{++} = 1.549(16)$ and $0^{++*}/0^{++} = 1.720(16)$ (the table of 17-01). The holographic ratios with the reference of COOT (1.14, 1.20, 1.48) come out 26%, 23% and 14% smaller respectively Computed .

08

Two scales — the tension is not predicted, and the scale of the circle does not go away

Glueball masses are fixed by an eigenvalue problem and do not depend on the coupling $\lambda = g_{\mathrm{YM}}^2 N$, but the string tension is proportional to λ. In the D4 model, from formula (2.7) of Kruczenski–Mateos–Myers–Winters (2003) Physics

$$\sigma = \frac{2}{27\pi}\,\lambda\, M_{\mathrm{KK}}^2, \qquad \frac{m0^{++}}{\sqrt{\sigma}} = \frac{m0^{++}}{M_{\mathrm{KK}}}\sqrt{\frac{27\pi}{2\lambda}} \quad (11)$$

(it may be off by a factor of 2 depending on the convention for λ). The value changes depending on which is put into $m0^{++}/M_{\mathrm{KK}}$ on the right: 2.03 from formula (10) of COOT, 1.567 from T4 of BMT (dilaton type), or 0.901 from S4 (the lightest).

Table 6 $m0^{++}/\sqrt{\sigma}$ from (11). The lattice $SU(\infty)$ is 3.072(14)

$m0^{++}/M_{\mathrm{KK}}$	$\lambda = 1$	10	30	100	1000	λ matching 3.072
2.03 (formula (10) of COOT)	13.2	4.19	2.42	1.32	0.42	19
1.567 (T4 of BMT)	10.2	3.23	1.86	1.02	0.32	11
0.901 (S4 of BMT)	5.87	1.86	1.07	0.59	0.19	3.7

At $\lambda \gg 1$, where supergravity is reliable, the ratio goes to 0 as $\lambda^{-1/2}$ Computed . The λ matching 3.072 is 3.7 to 19, moving by a factor of about 5 depending on which mode is taken, and in every case it is a fit, not a prediction. Klebanov–Strassler too acknowledge a 'large separation of scales' between the tension and the glueballs. On the lattice it is 3 to 4 (17-03).

In models cut off by a wall, the λ-dependence of the ratio is $\lambda^{-1/4}$. With a hard wall ((12) of 17-08) the tension is $\sigma = \sqrt{\lambda}/(2\pi z_w^2)$ and the glueball mass $\sim 1/z_w$ (Boschi-Filho–Braga 2003, imposing a Dirichlet condition at the wall). With a soft wall the tension is $\frac{R^2}{\alpha'}\frac{e}{4\pi}c$, from (19) of Andreev–Zakharov (2006) and $g = R^2/\alpha'$, and the squared mass is proportional to c ((15) of Karch–Katz–Son–Stephanov 2006; the case of mesons). In both the tension is proportional to $\sqrt{\lambda}$, so the ratio goes as $\lambda^{-1/4}$; the difference from the $\lambda^{-1/2}$ of the D4 model comes from the tension in (11) being proportional to λ Physics .

The other is the scale of the circle. In three dimensions, in the normalisation of § 03 the circumference on the boundary is π, so $R = 1/2$ and $m_1 = 3.404 = 1.70/R$ (COOT use another normalisation, placing the horizon at $b = 1/(2R)$). In four dimensions 0^{++} is $0.90/R_2$ to $2.03/R_2$, and whichever mode is taken, the glueballs are of the same order as the circle Computed . The KK modes on the sphere (17-06 § 02) are also 'of the same order' (COOT), and the states regarded as glueballs are chosen by hand. Removing them needs $\lambda \to 0$, which is the 'opposite' of the direction in which supergravity works. Witten himself called the computation at small $\eta = g_4^2 N$ 'a tall order' Physics . On the order of limits, see 17-10.

09

Another way of closing off — Klebanov–Strassler and the record

Klebanov–Strassler (2000) placed D3-branes (17-08 § 02) on a six-dimensional cone (the conifold) and built a background in which the singularity at the tip is deformed in the IR, leaving a finite sphere.

$$\sum_{i=1}^{4} z_i^2 = 0 \;\; \text{(conifold)} \;\longrightarrow\; \sum_{i=1}^{4} z_i^2 = \epsilon^2 \;\; \text{(deformed)}, \qquad z_i \in \mathbb{C} \tag{12}$$

Example: restricting z_i to real numbers, the right is an S^3 of radius ϵ, the sphere left at the tip Known . The flux of the five-form field through a cross-section of the cone (a quantity corresponding to the number of D3-branes) changes logarithmically with the radius r: $\mathcal{K}(r) = N + a\, g_s M^2 \ln(r/r_0)$. g_s is the string coupling (17-08 (3)), M the number of "fractional branes" wrapped on the sphere in the cone, and a a constant of order 1 (original paper). The authors read this decrease as a sequence (a cascade) in which the gauge group $SU(N+M) \times SU(N)$ shrinks step by step as $N \to N - M$ Physics . The deformation of the tip is held to be the breaking $\mathbb{Z}_{2M} \to \mathbb{Z}_2$ of the R-symmetry (the chiral symmetry of this theory) of the $\mathcal{N} = 1$ supersymmetric Yang–Mills remaining in the IR, made visible as geometry Physics . At the tip the coefficient of dx^2 is finite, so an area law comes out as in § 04. But the theory is $\mathcal{N} = 1$ supersymmetric, the UV cascade 'goes on forever', and the KK modes of S^3 and the glueballs 'both scale as $m/(g_s M)$'.

Table 7 The record (years are arXiv years)

Model	What fits	Input	What misses / is not predicted
Witten, QCD_3	$0^{--}/0^{++}$ 1.50 against 1.45(3) (with the dilaton as reference; 2.19 with S3 as reference)	0^{++}	Excited states 14–24%, tension, circle
Witten, QCD_4	$0^{-+}/0^{++}$ 1.14 against 1.36(32) (1998; with the dilaton of formula (10) of COOT as reference)	0^{++}	14–26% from today's lattice, ratios 1.14–2.09 depending on the choice of modes, $m/\sqrt{\sigma}$
Klebanov–Strassler (2000), Maldacena–Núñez (2000)	Area law, breaking of R-symmetry (chiral symmetry) (geometric)	—	$\mathcal{N} = 1$, same scale as KK
Gürsoy–Kiritsis–Nitti (2007)	$2^{++}/0^{++}$ 1.40 against 1.46(5)	0^{++*} $/0^{++}$	The potential is a fit, $N = \infty$

In every model the gap is shown as "a linear eigenvalue problem on a fixed background", and the part saying that the background describes pure Yang–Mills theory on $\mathbb{R}^4$ remains a conjecture. The Clay page too writes 'no proof of this property

is known'. The distance from the problem statement (0-03) is in 17-10, the seams in 16-05.

What this chapter can say

What is said	What is not said
Known (4)(8)(9)(12). Mathematics on a fixed background.	That the background describes pure Yang–Mills theory.
Computed Tables 1–6, Figures 1 and 2, (5), the behaviour of V of (9) at both ends. Reproducing the 0^{++} tables of COOT for QCD_3 and QCD_4. Reproducing the lowest values of S4, T4, V4, N4 in Table 2 of BMT and of S3 in three dimensions.	Reproducing the 0^{-+} (1.83 GeV) of QCD_4. The convention for λ in (11). Settling which mode is to be regarded as which state.
Physics The readings of (1)(3)(6)(7)(10)(10′)(11), glueball = eigenvalue, the assignment of quantum numbers, the comparisons of Tables 3, 4, 5 and 7.	Outside $N, \lambda \to \infty$. This book itself shows neither the mass gap nor the construction of the continuum limit (the published construction of the continuum limit is in Volume 2).

Sources and reproduction

Item	Kind	Source / tools
Witten's formulas and quotations	Physics, Known	E. Witten, Adv. Theor. Math. Phys. 2 (1998) 505, hep-th/9803131. Text checked.
(10), Tables 2–4, quotations from COOT, $6n(n+1)$	Physics	C. Csáki, H. Ooguri, Y. Oz, J. Terning, JHEP 9901 (1999) 017, hep-th/9806021. Text checked ($6n(n+1)$ is cited by that paper as a private communication from Y. Takei, T. Koike).
Table 5, (10′), the later analysis of § 07	Physics	R. C. Brower, S. D. Mathur, C.-I. Tan, Nucl. Phys. B 587 (2000) 249, hep-th/0003115. Text checked (Tables 1 and 2, the table for QCD_3, the quantum numbers of § 4, § 5.1).
Lattice	Physics	M. J. Teper, hep-lat/9804008, hep-lat/9711011; C. Morningstar, M. Peardon, PRD 56 (1997) 4043 (via COOT; bibliographic data only). A. Athenodorou, M. Teper, JHEP 11 (2020) 172; JHEP 12 (2021) 082 (from the checked survey records).
The condition at the tip, (11)	Physics	R. de Mello Koch et al., PRD 58 (1998) 105009, hep-th/9806125 (abstract, from the checked survey records); M. Kruczenski, D. Mateos, R. C. Myers, D. J. Winters, JHEP 0405 (2004) 041, hep-th/0311270, formula (2.7) (text checked: $\sigma = \frac{2}{27\pi} g_{\mathrm{YM}}^2 N_c M_{\mathrm{KK}}^2$).

Item	Kind	Source / tools
The $\lambda^{-1/4}$ of the wall models in § 08	Physics	H. Boschi-Filho, N. R. F. Braga, JHEP 0305 (2003) 009, hep-th/0212207 (abstract checked); O. Andreev, V. I. Zakharov, Phys. Rev. D 74 (2006) 025023, hep-ph/0604204, (19) and $g = R^2/\alpha'$ (text checked); A. Karch, E. Katz, D. T. Son, M. A. Stephanov, Phys. Rev. D 74 (2006) 015005, hep-ph/0602229, (15) (text checked). The tension with a hard wall is (12) of 17-08 (computed on this machine; that a linear term comes out with a wall was also checked in (12)–(16) of H. Boschi-Filho, N. R. F. Braga, C. N. Ferreira, Phys. Rev. D 73 (2006) 106006, hep-th/0512295). The exponent $-1/4$ is arithmetic on this machine combining these.
(12), § 09, Table 7	Physics, Known	I. R. Klebanov, M. J. Strassler, JHEP 0008 (2000) 052, hep-th/0007191 (text checked); J. M. Maldacena, C. Núñez, PRL 86 (2001) 588 (abstract).
The exponent α of § 06, Table 7	Physics	U. Gürsoy, E. Kiritsis, F. Nitti, JHEP 0802 (2008) 019, arXiv:0707.1349 (text checked).
Detour, Clay	Known	G. W. Gibbons, S. W. Hawking, 'Action integrals and partition functions in quantum gravity', Phys. Rev. D 15 (1977) 2752 (bibliographic data only). The Clay page on the problem (from the checked survey records).
Tables, figures, numbers	Computed	Python 3 (scipy) on this machine: glueball.py, cigar_wilson.py, fig_shoot.py, rev_bmt.py (Tables 5 and 6), rev_liouville.py (the V of (9)) and 5 others.

Next chapter to read: 17-10 Back to four dimensions — the order of limits and the distance from the problem statement.

Back to four dimensions — the direction that removes the extra dimensions is opposite to the direction in which geometry can compute

> This chapter organises established knowledge. § 03 is this book's own investigation.

What this chapter shows — That $R \to 0$, which removes the extra dimensions, is because of dimensional transmutation the same direction as $\lambda \to 0$. That geometry can compute in the opposite direction, $\lambda \to \infty$, where the KK tower remains at the same height as the glueballs. That the uniformity needed to exchange the limits is missing.

Prerequisite chapters — 17-09. We also use dimensional transmutation in 17-01, 't Hooft's counting in 17-04 and the Curie–Weiss model in 2-08.

Said first — (2)(3) and the arithmetic of the tables are physical estimates, not theorems. Within the range we searched we found no result justifying the limit from the models to the four-dimensional theory, and this book itself shows nothing about the mass gap or the construction of the continuum limit (the published construction of the continuum limit is read in Volume 2).

The order of this chapter

1. Three knobs — N, λ, R and the problem statement
2. Dimensional transmutation fixes the direction — $R \to 0$ is $\lambda \to 0$
3. Supergravity works in the opposite direction — the window and the floor (Figure 1)
4. Extra things remain — the KK tower and supersymmetry
5. $N \to \infty$ — when the group is fixed to one
6. The order of limits — exchanging them needs uniformity
7. Detour: seeing non-uniform convergence (Figure 2)
8. The distance from the problem statement — five requirements
9. What remains — the problem moves

01

Three knobs — N, λ, R and the problem statement

The question of this chapter is **whether the models of 17-09 can be brought back to the four-dimensional theory of the prize problem**. The knobs are three: the N of $SU(N)$, the 't Hooft coupling $\lambda = g^2 N$ at the scale M_{KK} (17-04), and the radius of the circle of the extra dimension $R = 1/M_{KK}$ (the spacing of the KK tower, 17-06). With Λ the glueball mass, the ratio ΛR measures how visible the extra dimensions are. The requirements of the problem statement (Jaffe–Witten § 4, 0-03) are

G fixed one at a time (e.g. $SU(3)$), space is $\mathbb{R}^4$ ($R = 0$), agrees with asymptotic free

and the third is that the correlation functions agree at short distances with the predictions of asymptotic freedom and perturbative renormalisation theory (§ 4; quotation in 0-03 § 07) Known . The models, conversely, can compute where $N = \infty$, $\lambda \gg 1$ and ΛR is of order 1 (§ 03, § 05).

Example: the Sakai–Sugimoto model fixes $M_{KK} \simeq 949$ MeV and $\kappa = \lambda N_c/(216\pi^3) \simeq 7.45 \times 10^{-3}$ from experimental values. For $N_c = 3$, $\lambda \approx 16.6$ Computed . This is the value in the coupling convention of Sakai–Sugimoto, and can differ from the λ of (2) in the next section by a factor of about 2 because of the difference in convention. The numbers put into (2) below are rough guides. The 776 MeV of the ρ meson is 0.82 times M_{KK}, at the same height as the tower.

02

Dimensional transmutation fixes the direction — $R \to 0$ is $\lambda \to 0$

Removing the extra dimensions means taking $R \to 0$ with Λ fixed. Λ is born from the running of the coupling (17-01, 4-07). Rewriting the one-loop formula $\mu\, dg/d\mu = -b_0 g^3$ ($b_0 = 11N/(48\pi^2)$) of (1) of 17-01 in terms of $\lambda = g^2 N$ gives

$$\mu \frac{d\lambda}{d\mu} = 2gN\,\mu \frac{dg}{d\mu} = -2b_0 N\, g^4 = -\frac{2b_0}{N}\,\lambda^2 = -\frac{11}{24\pi^2}\,\lambda^2$$

Dividing both sides by λ^2 gives $d(1/\lambda)/d\ln\mu = 11/(24\pi^2)$, so $1/\lambda$ is linear in $\ln\mu$. Below, λ without an argument is the value at $\mu = M_{KK}$.

$$\frac{1}{\lambda(\mu)} = \frac{1}{\lambda} + \frac{11}{24\pi^2}\ln\frac{\mu}{M_{KK}}$$

Lowering μ decreases the right-hand side, which becomes 0 at $\mu = M_{KK}\exp\big(-24\pi^2/(11\lambda)\big)$ (the one-loop coupling diverges here). This scale where the coupling becomes strong is a rough guide to Λ:

$$\Lambda R \;\sim\; \exp\!\Big(-\frac{24\pi^2}{11\,\lambda}\Big) \quad (2)$$

Physics . $\sim$ means "equal up to a constant multiple", and the constant changes with the definition of Λ (the Λ parameter, or the glueball mass) and with conventions. Solving the other way gives

$$\lambda(M_{KK}) \simeq \frac{24\pi^2/11}{\ln\big(1/(\Lambda R)\big)} \;\xrightarrow[R\to 0]{}\; 0 \quad (3)$$

(a constant multiple only adds a constant to the denominator, so it does not affect the direction of $R \to 0$), and **removing the extra dimensions is the same direction as $\lambda \to 0$**. Witten (1998 § 4) writes 'Asymptotic freedom should say that as $\eta \to 0$, the particle masses become exponentially small' (η is the λ of this chapter), and Csáki–Ooguri–Oz–Terning, putting the masses as $M^2 = f(\lambda)\Lambda_{UV}^2$, state 'To take the continuum limit $\Lambda_{\mathrm{UV}} \to \infty$, we have to simultaneously take $\lambda \to 0$' Physics ((2) is (10) of 17-07). As a metaphor, R plays the role of the lattice spacing; more precisely, it is the same logic as sending the bare coupling to 0 as the spacing $a \to 0$ (5-05).

Table 1 Estimates of ΛR (with the constant multiple in (2) set to 1) and the rough guide $\lambda^{-3/2}$ to the string correction

λ	One loop (2)	Two loops	$\lambda^{-3/2}$
1	4.45×10^{-10}	2.17×10^{-9}	1
5	0.0135	0.0334	0.0894
10	0.116	0.215	0.0316
16.6	0.274	0.409	0.0147
100	0.806	0.565	0.0010

Computed Two loops: from $\mu\, dg/d\mu = -b_0 g^3 - b_1 g^5$ ($b_1 = \frac{34N^2}{3(16\pi^2)^2}$), the formula of 17-01 with the next term added, the right-hand side of (2) multiplied by $(11\lambda/48\pi^2)^{-51/121}$ ($51/121 = b_1/(2b_0^2)$). The coefficients can change with convention, and the rows with large λ are rough guides outside the range where (2) can be used (that is also why, at $\lambda = 100$, the two-loop column is smaller than the one-loop one; § 03). Example: $\Lambda R = 10^{-3}$ needs $\lambda \le 3.12$ (two loops 2.66), and 10^{-6} needs $\lambda \le 1.56$ (1.41).

03

Supergravity works in the opposite direction — the window and the floor

Gravitational geometry can compute when $\lambda \gg 1$ (17-08). Csáki et al. evaluated the leading string correction to the f in the $M^2 = f(\lambda)\Lambda_{UV}^2$ of § 02, and for the 0^{++} glueball of **three-dimensional** pure Yang–Mills (QCD_3) obtained

$$f(\lambda) = c_0 + c_1 \lambda^{-3/2} + \cdots, \qquad c_1 < 0 \quad (4)$$

(the lowest 0^{++} state and the first five excited levels) Physics . $c_1 < 0$ means that with the ultraviolet cutoff Λ_{UV} fixed, decreasing λ lowers the mass, the direction consistent with the expectation of a continuum limit. The power in four dimensions is not in this paper. Below we borrow $p = 3/2$ as a rough guide for four dimensions.

The same paper writes, for the three-dimensional case, of taking $g_4^2 N \to 0$ as $R \to 0$ (to keep the QCD_3 coupling $g_3^2 N = g_4^2 N / R$ constant): 'This is the opposite of the limit that is required for the supergravity description to be valid'. The grounds for four dimensions are Witten's sentence in § 02 and the sentence 'To take the continuum limit ...' in the introduction of Csáki et al. Witten too writes of showing asymptotic freedom and computing masses at small η: 'This looks like a tall order, given our limited knowledge of worldsheet field theory with Ramond-Ramond fields in the Lagrangian' (world-sheet field theory is explained in § 09).

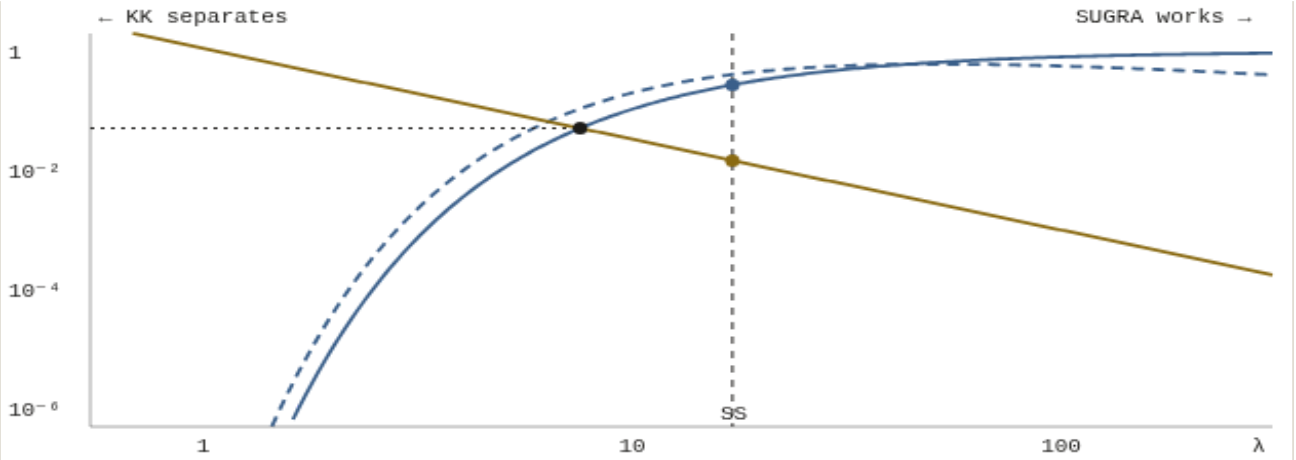
We look at the tug of war in numbers, with $e_{KK} = \Lambda R$ (the constant multiple in (2) set to 1) for how visible the tower is, and $e_{\text{str}} = \lambda^{-p}$ with coefficient 1 (borrowing (4), $p = 3/2$) as the rough guide to the string correction. We call the value at the crossing point λ^* of the two curves the **floor**.

$$F_p = e_{KK}(\lambda^*) = e_{\text{str}}(\lambda^*) \quad (5)$$

$\log e_{KK} + p \log \lambda$ is increasing in λ (at one loop and at two loops, if $p \geq 1$), so there is exactly one crossing point. The one-loop e_{KK} increases towards 1 and e_{str} decreases, so to the left of the crossing e_{str}, and to the right e_{KK}, exceeds F_p. That is, in the one-loop estimate $F_p = \min_\lambda \max(e_{KK}, e_{\text{str}})$, and **for any λ the larger of the two guides does not fall below F_p** Computed .

With the two-loop formula things differ. e_{KK} reaches its maximum 0.61 at $\lambda \approx 51$, and beyond that decreases towards 0 as $\lambda \to \infty$. So if the upper limit on λ is removed, the minimum of the max approaches 0 (for $p = 3/2$: 0.067 up to $\lambda \leq 10^4$, 0.038 for $\lambda \leq 10^5$, 0.014 for $\lambda \leq 10^6$) Computed . This is the behaviour of the perturbative formula for the running of the coupling extended to the large-λ side, outside the range where (2) can be used. The two-loop rows of Table 2 are crossing values, and also agree with the minimum of the max restricted to $\lambda \leq 316$ (the range of Figure 1).

Figure 1 Varying λ lets you read the values there of ΛR (blue; solid one loop, dashed two loops) and λ^{-p} (yellow). The horizontal dotted line is the one-loop floor F_p. SS is the λ of Sakai–Sugimoto (it can be off by a factor of about 2 by convention). The dashed line falling to the right of $\lambda \approx 51$ is outside the range where (2) can be used



λ = 16.63, ΛR: one loop 0.274, two loops 0.409, λ^−3/2 = 0.015, floor F = 0.051 (two-loop crossing 0.067)

Computed The numbers in this figure were computed on this machine (the curves are computed on the spot from the formulas; the floor agrees with Table 2).

Table 2 The floor F_p (the value at the crossing) and the crossing point λ^*

Formula for ΛR	p	λ^*	F_p
One loop	1	9.55	0.105
One loop	3/2	7.25	0.051
Two loops	1	7.77	0.129
Two loops	3/2	6.09	0.067

There is a **window** where both are around 5–13%, which does not contradict ratios agreeing with the lattice to a few per cent (when the three-dimensional $0^{--}/0^{++}$ is compared with the dilaton as reference; changing the reference mode or the states gives deviations of 14–26%: 17-09). But what a theorem needs is a **limit** sending the errors to 0, and in the one-loop estimate the floor blocks it. Both guides are crude, with coefficient 1; the point is not the value of the floor itself but that there is no direction making both small at once. Example: putting the $\lambda \approx 16.6$ of Sakai–Sugimoto straight into (2) gives $e_{KK} = 0.27$ (two loops 0.41) and $e_{\mathrm{str}} = 0.015$. Allowing for the difference of convention, halving λ (8.3) or doubling it (33) gives e_{KK} of 0.075–0.52 and e_{str} of 0.042–0.005, and their order does not swap Computed . This is a rough guide suggesting that the tower side remains large.

04

Extra things remain — the KK tower and supersymmetry

At $\lambda \gg 1$ the KK tower comes at the same height as the glueballs. The squared masses of fluctuations on a fixed background are eigenvalues of Laplace-type operators, and changing the metric as $g \mapsto s^2 g$ multiplies all eigenvalues by s^{-2}.

$$m_n^2 = \frac{c_n}{R^2} \qquad (c_n \text{ numbers fixed by the shape of the background alone}) \ (6)$$

If the background has a single length R, glueballs and KK modes alike are some multiple of $1/R$ (the counting of powers is Known ; the part saying that the background describes Yang–Mills is Physics). But the string tension involves λ, and Klebanov–Strassler write $T_s \sim g_s M\, m_{\text{glueball}}^2$ (the same circumstance as 17-08 § 07). That the tower does not separate is acknowledged by the authors in Table 3 (for Witten's background, by Csáki et al. and Klebanov–Strassler) Physics .

Table 3 What remains, and the authors' words

Model	What remains	The authors' words
Csáki et al. (Witten's background)	$SO(6)$ non-singlets	'their masses are of the same order as the SO(6) singlet states ... Presumably λ^{-1} corrections make them heavy'. Some fields removed by hand
Maldacena–Núñez	KK on the fivebranes	'the masses of glueballs are comparable to the masses of Kaluza Klein states on the fivebrane'
Klebanov–Strassler	KK on S^3	'in the SUGRA limit both scale as m/(g_sM)'. States explicitly that Witten's background also has extra modes
Polchinski–Strassler	$\mathcal{N} = 4$ matter	'the N=4 matter cannot be decoupled within the supergravity regime'
Sakai–Sugimoto	A tower of mass M_{KK}, $SO(5)$	'the same energy scale of the mass of the vector mesons'. A footnote says 'believed somewhat optimistically' (the latter in a footnote of hep-th/0507073)

How supersymmetry remains

In Witten's type (1998 § 4; fermions made antiperiodic on the circle), the fermions acquire masses of order $1/R$ at tree level, the scalars 'very plausibly get masses ... at one-loop level', and it continues 'If this is so, the low energy theory will be the pure SU(N) theory without supersymmetry' (§ 2.2, on the three-dimensional case, has a sentence to the same effect: at weak coupling the long-distance behaviour is 'very plausibly' only a three-dimensional gauge field). Klebanov–Strassler and Maldacena–Núñez preserve $\mathcal{N} = 1$ supersymmetry, so a further limit making the gluino (the fermion that is the supersymmetric partner of the gluon) heavy is needed. The five-dimensional model of Gürsoy–Kiritsis is of the kind (bottom-up) that puts a five-dimensional gravity model directly rather than deriving it from a string-theory background, and takes the β function as input ('using input ... as well as QCD') Physics . Example: within the range we examined, among the models built from string-theory backgrounds (top-down), only Witten's type yields non-supersymmetric pure Yang–Mills, and that too at the same scale as the tower.

05

$N \to \infty$ — when the group is fixed to one

Gravitational geometry becomes classical as $N \to \infty$. The quantities of the gauge theory are arranged by 't Hooft's counting (17-04) into a sum over genus h.

$$\log Z = \sum_{h \geq 0} N^{2-2h} F_h(\lambda) \quad (7)$$

What the gravity computation gives is the leading part of F_0 on the large-λ side, and for $SU(3)$ ($1/N^2 = 1/9$) all the remaining terms matter. That the ratios move slowly with N on the lattice (17-04) Physics does not give the existence of the theory at fixed N either. There is a gap in the kinds of group too: D-branes give $U(N)$ and $SU(N)$, and orientifolds (backgrounds divided by an operation that reverses the orientation of strings) give SO and Sp (17-07). Within the range we searched we found no holographic dual of four-dimensional pure Yang–Mills for the exceptional groups $G_2, F_4, E_{6,7,8}$, and geometric engineering can make types ADE, and types B, C, F, G by folding with outer automorphisms (automorphisms of Dynkin diagrams), but within supersymmetric theories (17-07). Example: in the problem statement, $SU(2)$ and G_2 are each a single theory.

06

The order of limits — exchanging them needs uniformity

The gravity computation takes the supergravity limit $\lambda \to \infty$ after the $N \to \infty$ that makes the string classical, and uses R as the unit.

$$N \to \infty \ \longrightarrow \ \lambda \to \infty, \qquad R \text{ fixed} \quad (8)$$

The problem statement's side fixes N and takes $R \to 0$, that is, by (3), $\lambda \to 0$. A standard theorem of analysis gives the condition under which two limits may be exchanged Known . A double sequence $a_{m,n}$ converges to b_m as $n \to \infty$ **uniformly in** m when

$$\sup_m \left| a_{m,n} - b_m \right| \xrightarrow[n \to \infty]{} 0$$

that is, when a single n makes the errors for every m small at once. Then the following holds.

$$a_{m,n} \xrightarrow[n \to \infty]{} b_m \ (m\text{-uniformly}), \quad a_{m,n} \xrightarrow[m \to \infty]{} c_n \implies \lim_{m \to \infty} b_m = \lim_{n \to \infty} c_n \quad (9)$$

Without uniformity the exchange is in general not allowed. Example: for $a_{m,n} = m/(m+n)$, $\lim_m \lim_n a_{m,n} = 0$ while $\lim_n \lim_m a_{m,n} = 1$ (since $\sup_m m/(m+n) = 1$, the

convergence to 0 as $n \to \infty$ is not uniform).

What is needed is that the $1/N$ corrections be uniformly small in λ, that the string corrections be uniformly small in N, and an unbroken road down to $\lambda \to 0$. (4) is the leading term of the correction in three dimensions, neither uniformity nor a limit. **Within the range we searched we found no result justifying this exchange.** As a metaphor, the supergravity region ('fixed ultraviolet cutoff … in the strong ultraviolet coupling regime') is in the position of a coarse lattice at strong coupling. More precisely, on the strong-coupling lattice the existence of the infinite-volume limit is proved (Osterwalder–Seiler 1978, 5-04), but this correspondence stays a metaphor.

07

Detour: seeing non-uniform convergence

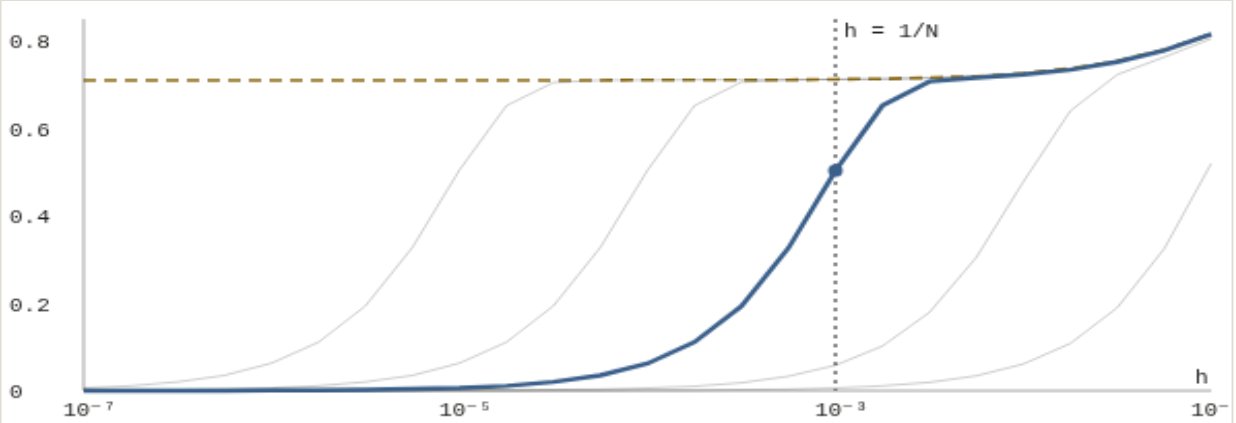
A section you may skip.

Spontaneous magnetisation changes value with the order of $h \to 0$ and $N \to \infty$ ((10) of 2-08). In the Curie–Weiss model ($J = 1$, inverse temperature $\beta = 1.25$) we count where it is non-uniform. The magnetisation per spin is written $\bar{m}$, as in 2-08, to distinguish it from mass. At finite N it can be written exactly as a sum of binomial coefficients.

$$\langle \bar{m} \rangle_{N,h} = \frac{\sum_k \binom{N}{k} \bar{m}_k \, e^{\beta N (\bar{m}_k^2/2 + h\bar{m}_k)}}{\sum_k \binom{N}{k} \, e^{\beta N (\bar{m}_k^2/2 + h\bar{m}_k)}}, \qquad \bar{m}_k = \frac{2k - N}{N} \tag{10}$$

By the symmetry of flipping all spins, $\langle \bar{m} \rangle_{N,0} = 0$, so $\lim_N \lim_{h \to 0} = 0$, while $\lim_{h \downarrow 0} \lim_N = \bar{m}^* = 0.71041$.

Figure 2 As N grows, the range over which $\langle \bar{m} \rangle_{N,h}$ (blue) clings to the $N = \infty$ curve (yellow dashed) extends to the left. The vertical dotted line is $h = 1/N$



N = 10³, ⟨m̄⟩ at h = 1/N = 0.503 ((11) at x = 1: 0.505)

Computed The numbers in this figure were computed on this machine ((10) evaluated exactly and embedded in steps of 0.25 in $\log_{10} h$).

Table 4 $\langle \bar{m} \rangle_{N,h}$ ($\beta = 1.25$). Bold is the $h = 1/N$ column

N	$h = 10^{-2}$	10^{-3}	10^{-4}	10^{-5}	$h = 0$
10^2	**0.4780**	0.0579	0.0058	0.0006	0
10^3	0.7231	**0.5028**	0.0625	0.0063	0
10^4	0.7255	0.7117	**0.5045**	0.0629	0
10^5	0.7257	0.7120	0.7105	**0.5047**	0
∞	0.7257	0.7120	0.7106	0.7104	$\bar{m}^*$ $(h \downarrow 0)$

Following a row to the right leads to 0, following a column down leads near $\bar{m}^*$, and the diagonal is about 0.5. For large N the distribution splits into two peaks at $\pm\bar{m}^*$ with weights in the ratio $e^{2\beta h N \bar{m}^*}$, so

$$\langle \bar{m} \rangle_{N,\, x/N} \longrightarrow \bar{m}^* \tanh(\beta \bar{m}^* x) \qquad (N \to \infty) \quad (11)$$

Against the value 0.50468 at $x = 1$, the exact value at $N = 10^5$ was 0.50467 Computed . **Non-uniformity sits in the band $h \sim 1/N$, which escapes to the left as N grows.** Here (10) can be written exactly, so the position of the band is known, but for the two knobs of (8) there is as yet no way of checking.

08

The distance from the problem statement — five requirements

The problem statement is '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 $\Delta > 0$', and existence includes satisfying axiomatic properties (Jaffe–Witten § 4) Known .

Table 5 The requirements, and where the extra-dimensional models stand (every row at the level of Physics)

Requirement	Where things stand
Fix the group G	$N \to \infty$ for $SU(N)$. No duals for exceptional groups found (§ 05)
On $\mathbb{R}^4$	ΛR of order 1 on $\mathbb{R}^4 \times$ (extra dimensions) (§ 03)
Asymptotically free at short distances	The ultraviolet is a five-dimensional theory or the (2,0) theory (17-07). The cascade of Klebanov–Strassler 'goes on forever as the scale is increased'
Exists, satisfying the axioms (4-10)	AdS/CFT is a conjecture (17-08 § 08). The string-theory side has no non-perturbative definition
Mass gap	A supergravity eigenvalue problem (semiclassical). Witten: 'A mass gap likewise can be demonstrated ... by using the large λ behavior of the metric' (this λ is a radial coordinate of the metric)

None of the rows of Table 5 is satisfied yet. Jaffe–Witten § 6 wrote 'at present we do not know any non-trivial relativistic field theory that satisfies the Wightman (or any other reasonable) axioms in four dimensions' (a sentence from 2000; the existence of the four-dimensional theory is given by a construction published in 2026 as the limit of

Wilson's lattice, which the committee has assessed; Volume 2). On the mass gap, the problem page of the Clay Mathematics Institute still writes 'no proof of this property is known' Known (the definition of the gap is in 4-11). Example: the ratios the models hit to a few per cent (17-09; only the three-dimensional ratio with the dilaton as reference) are also numbers obtained while satisfying none of the rows of Table 5.

09

What remains — the problem moves

The problem does not vanish; it moves. Witten, noting that a singularity appears in the geometry as $\eta \to 0$, writes 'Apparently, in this approach, the mysteries of four-dimensional quantum gauge theory are encoded in the behavior of string theory near this singularity', and Csáki et al. write that under the conjecture the problem of QCD at large N reduces to 'that of controlling the α′ corrections to the two-dimensional sigma-model with the Ramond-Ramond background' Physics . We give the meanings of the words briefly. α' is the square of the string length (the string tension is $1/(2\pi\alpha')$, (1) of 17-05). The two-dimensional sigma model is the field theory on the two-dimensional surface swept out by a string (the **world sheet**), whose fields are maps from the world sheet to spacetime. Ramond–Ramond fields are fields sourced by D-branes. The α' corrections are corrections coming from the curvature radius of spacetime being finite compared with the string length, and (4) evaluates the leading term of this expansion. The 'worldsheet field theory with Ramond-Ramond fields' that Witten named in § 03 (a field theory on the world sheet whose Lagrangian contains Ramond–Ramond fields) is also this two-dimensional theory.

$$\text{4-dim. pure YM} \xleftarrow{\text{limit (3)}} \text{extra-dim. models} \xleftarrow{\text{conjecture}} \text{strings, M-theory} \quad (12)$$

Example: neither arrow in (12) is a mathematical theorem (the left is unjustified, the right a conjecture), and the object at the right end has no non-perturbative definition either (17-13, 16-05).

An idea from this book's study (before verification)

Restricted to linear fluctuations of the supergravity fields, we see the "single length" of (6) as the reason the tower does not separate, and also as the reason that makes "a single scale" (17-12, 23-01) trivial on a fixed background. But the string tension involves λ (the formula of Klebanov–Strassler in § 04), and once strings are included the lengths are not exhausted by one. Going back to four dimensions is a change of vehicle from R to Λ, and the question remains after that. This is neither a theorem nor a claim of the literature (23-17).

What this chapter can say

What is said	What is not said
Known The powers in (6), (9), the quotations from the problem statement.	That the background describes four-dimensional pure Yang–Mills theory (a conjecture).
Computed Tables 1, 2 and 4, Figures 1 and 2, the floor of (5) (the value at the crossing) and that for the one-loop formula it is the minimum of the max, the maximum of the two-loop formula ($\lambda \approx 51$), $\lambda \approx 16.6$ and the range of a factor of 2 around it, the check of (11).	Coefficients independent of convention. e_{str} is a rough guide, not the actual string correction. The power in (4) is for the three-dimensional 0^{++}; the power in four dimensions is not given. The behaviour of the two-loop formula for $\lambda \gtrsim 51$ is outside the range of (2).
Physics That $R \to 0$ and $\lambda \to 0$ are the same direction, that the tower does not separate (the authors' words), the order in (8).	A justification of exchanging the limits (none found within the range searched). This book itself shows neither the mass gap nor the construction of the continuum limit (the published construction of the continuum limit is in Volume 2).

Sources and reproduction

Item	Kind	Source / tools
(1), the problem statement in § 08 and the sentence of § 6	Known	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory” § 4, § 6 (the Clay problem description). Text of the PDF checked.
The sentence on the Clay page in § 08	Known	Clay Mathematics Institute, “Yang-Mills & the Mass Gap” (the Millennium Prize Problems page). https://www.claymath.org/millennium/yang-mills-the-maths-gap/. Checked by reading the sentence.
The b_0 of (2) / the two-loop b 1	Known	The sources of 4-06 / W. E. Caswell, Phys. Rev. Lett. 33 (1974) 244 (bibliographic data only).
Witten's sentences, Witten's type	Physics	E. Witten, Adv. Theor. Math. Phys. 2 (1998) 505, hep-th/9803131 § 2.2, § 4. arXiv manuscript checked.
(4), the sentences of Csáki et al.	Physics	C. Csáki, H. Ooguri, Y. Oz, J. Terning, JHEP 9901 (1999) 017, hep-th/9806021 (the abstract, (1.1)(1.2) of the introduction, the QCD_3 sentence of § 2, the concluding sentences). arXiv manuscript checked.
Klebanov–Strassler in Tables 3 and 5	Physics	I. R. Klebanov, M. J. Strassler, JHEP 0008 (2000) 052, hep-th/0007191. TeX manuscript checked.

Item	Kind	Source / tools
The other models in Table 3 and § 04	Physics	J. M. Maldacena, C. Núñez, Phys. Rev. Lett. 86 (2001) 588, hep-th/0008001; J. Polchinski, M. J. Strassler, hep-th/0003136; U. Gürsoy, E. Kiritsis, JHEP 0802 (2008) 032, arXiv:0707.1324. Abstracts checked for all.
Sakai–Sugimoto in § 01 and Table 3	Physics	T. Sakai, S. Sugimoto, Prog. Theor. Phys. 113 (2005) 843, hep-th/0412141; ibid. Prog. Theor. Phys. 114 (2005) 1083, hep-th/0507073. TeX manuscripts checked.
The strong-coupling lattice in § 06	Known	K. Osterwalder, E. Seiler, “Gauge Field Theories on the Lattice”, Ann. Phys. 110 (1978) 440. Abstract checked.
(9)	Known	Uniform convergence and the exchange of limits (for example W. Rudin, Principles of Mathematical Analysis, Chapter 7). Bibliographic data only.
Tables, figures, numbers	Computed	Python 3 (numpy, scipy) on this machine: scale_window.py (Tables 1 and 2), floor_crossing.py (uniqueness of the crossing in (5), the maximum of the two-loop formula, the minimum of the max for $\lambda \le 316$, the range of a factor of 2 around $\lambda \approx 16.6$), floor_twoloop_check.py (the minimum of the max with the two-loop formula when the upper limit on λ is removed), curie_weiss_order.py (Table 4, Figure 2, (11)).

Next chapter to read: 17-11 Supersymmetry and Seiberg–Witten — an example where, under supersymmetry, a gap is visible from a duality.

2026-10-01 · chapter Yang–Mills and the mass gap Part 17 From the side of physics supersymmetry and electric–magnetic duality

Supersymmetry and Seiberg–Witten — switching to the dual, the strongly coupled vacuum becomes a weakly coupled theory of monopoles

> 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 low energies of the N=2 supersymmetric `SU(2)` theory are fixed by a single holomorphic function and its singularities and monodromies.
That at a strongly coupled point the mass of the monopole becomes 0, and that on breaking to N=1, confinement and a gap appear through monopole condensation as a weak-coupling calculation.
How far this is an argument of physics, and where theorems of mathematics begin.

Prerequisite chapters — 17-08. The tools are dual superconductivity in 17-02, dimensional transmutation in 17-01 and SU(2) in 1-04. Holomorphic functions, harmonic functions and the periods of a torus are defined on the spot.

Said first — The theory of this chapter is not pure Yang–Mills (it contains scalar fields and fermions). The central results are at the level of physics, and the N=2 theory itself has not been constructed mathematically. Nothing is shown about the mass gap of the problem statement.

The order of this chapter

1. What can be seen in a neighbouring theory — the one-loop coefficients
2. The minimum of supersymmetry — the set of fields, the flat directions, the coordinate u
3. A single holomorphic function — one loop and instantons
4. The contradiction with positivity — a harmonic function has no minimum in the interior
5. Exchanging electric and magnetic — $\tau \to -1/\tau$ and the BPS masses

6. Singularities and monodromy — the points where the monopole becomes light (Figure 1)
7. Writing with the periods of a curve — the periods of a torus and $\operatorname{Im}\tau > 0$ (Figure 2; Detour: at $u = 3$ we get $\tau = i$)
8. Reading the instanton coefficients from the exact solution — a series with rational coefficients, checked against one loop
9. The perturbation to N=1 — monopole condensation and the mass gap
10. The distance from pure Yang–Mills — the border between physics and mathematics

---

01

What can be seen in a neighbouring theory — three theories and their one-loop coefficients

Pure Yang–Mills is hard because the coupling becomes strong at the low energies where the gap is decided, and perturbation theory does not work (17-01). Gauge theories with two sets of supersymmetry (N=2) are the typical **neighbouring theory** that can be computed into the strongly coupled side, and Seiberg and Witten (1994, SW below) proposed to fix the low-energy effective theory exactly in the case of `SU(2)`. The abstract says 'The "strongly coupled" vacuum turns out to be a weakly coupled theory of monopoles, and with a suitable perturbation confinement is described by monopole condensation' Physics .

The three theories differ in their set of fields and in the coefficient b of the one-loop running (in the convention where pure Yang–Mills has $11N/3$). One Weyl fermion in the adjoint representation (a two-component fermion; the Dirac fermion of the electron has four components) adds $-2N/3$, and one complex scalar adds $-N/3$ Known .

$$b_{\mathrm{YM}} = \tfrac{11}{3}N, \qquad b_{\mathcal{N}=1} = \tfrac{11}{3}N - \tfrac{2}{3}N = 3N, \qquad b_{\mathcal{N}=2} = \tfrac{11}{3}N - 2\cdot\tfrac{2}{3}N - \tfrac{1}{3}N = 2N \quad (1)$$

Example: for `SU(2)` these are, in order, $22/3,\ 6,\ 4$. The N=2 theory has no gap, since a massless photon remains; a gap appears only when it is broken to N=1 in § 09. Section 6.6 of the problem statement also cites Seiberg–Witten as an example of a route to understanding the gap in theories with matter ('for example, see [41]') Known .

02

The minimum of supersymmetry — the set of fields, the flat directions, the coordinate u

Supersymmetry is a symmetry that exchanges bosons and fermions, and N counts how many sets of its generators (supercharges) there are. A set of fields carried into one another by supersymmetry is called a **multiplet**. The smallest set for an N=2 gauge theory is the gauge field A_μ, two Weyl fermions λ, ψ and a complex scalar ϕ, all in the adjoint representation (1-04). The potential of the scalar is

$$V(\phi) = \frac{1}{g^2}\,\mathrm{tr}\,[\phi, \phi^\dagger]^2 \quad (2)$$

and it is 0 if ϕ and $\phi^\dagger$ commute. For `SU(2)` a gauge transformation brings it to $\phi = \frac{1}{2} a \sigma^3$ ($a \in \mathbb{C}$), and since a and $-a$ are carried into each other, the gauge-invariant coordinate is $u = \mathrm{tr}\,\phi^2 = \frac{1}{2}a^2$ Known . The vacua with $V = 0$ form a family numbered by the complex number u (a moduli space), and SW argue that this family is not lifted in the quantum theory either Physics .

For $a \neq 0$ the only elements commuting with ϕ are the diagonal `U(1)`, and the remaining two gauge fields (the W bosons) acquire mass by the **Higgs mechanism** (the mechanism by which a scalar field that takes a value gives mass to gauge fields; 3-07). The `U(1)` photon stays massless, so there is no gap. Example: if $\phi = \sigma^+ = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, then $[\phi, \phi^\dagger] = \sigma^3$ and $V = 2/g^2 > 0$; if $\phi = \frac{1}{2} a \sigma^3$, then $V = 0$.

03

A single holomorphic function — the prepotential, one loop, instantons

A complex function is **holomorphic** if it can be differentiated as a function of a complex number. Under N=2 supersymmetry the low-energy effective theory can be written with a single holomorphic function $\mathcal{F}(a)$ (the **prepotential**) (Seiberg 1988, SW § 2) Physics . The complex number combining the coupling g and the θ angle,

$$\tau(a) = \frac{\theta}{2\pi} + \frac{4\pi i}{g^2} = \mathcal{F}''(a), \qquad ds^2 = \mathrm{Im}\,\tau(a)\,|da|^2 \quad (3)$$

gives the coupling of the photon and the distance (metric) on the family of vacua. Where $|a|$ is large, the theory is asymptotically free and weakly coupled (4-06). The original writes

$$\mathcal{F}(a) = \frac{i}{2\pi}\,a^2 \ln \frac{a^2}{\Lambda^2} + \sum_{k \geq 1} \mathcal{F}k \left(\frac{\Lambda}{a}\right)^{4k} a^2 \quad (4)$$

and continues 'Higher order perturbative corrections are absent. Instantons lead to new terms.' The first term is one loop, and the k-th term is the contribution of k instantons (3-05) Physics .

The power $4k$ comes from the $b = 4$ of (1). In the convention of (3) in 17-01, with $b_0 = b/16\pi^2 = 1/4\pi^2$,

$$\Lambda^4 = \mu^4 \exp\Big(- \frac{8\pi^2}{g^2(\mu)} \Big) \quad (5)$$

and the exponential is the weight of a single instanton. $(\Lambda/a)^4$ appears at no order of perturbation theory (4-12) Known . Example: in the units that put the singularities at $u = \pm 1$, one loop gives $\operatorname{Im}\tau \approx (2/\pi)\ln(2|a|)$ (from the numbers in § 07). The coefficient $2/\pi = 8\pi b_0$ agrees with the running of (5) Computed .

04

The contradiction with positivity — a harmonic function has no minimum in the interior

The metric of (3) must be positive. But the imaginary part of a holomorphic function is a **harmonic function**.

$$\Big(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \Big) \operatorname{Im}\tau = 0 \qquad (a = x + iy) \quad (6)$$

The value of a harmonic function at the centre of a circle equals its average over the circle, and unless it is constant it cannot have a minimum in the interior of a region (the minimum principle) Known .

If τ is a single function on the whole a-plane, then $\operatorname{Im}\tau$ goes to $+\infty$ as $|a| \to \infty$, so it has a minimum somewhere, contradicting the principle. The original writes 'if Im τ(a) is globally defined it cannot be positive definite as the harmonic function Im τ cannot have a minimum', and concludes that the description by a is only locally correct Physics . Example: the one-loop $(2/\pi)\ln(2|a|)$ is negative for $|a| < 1/2$, and at $|a| = 1/4$ it is -0.44; that is, "$4\pi/g^2$ is negative" Computed . On the strongly coupled side another variable is needed.

05

Exchanging electric and magnetic — $\tau \to -1/\tau$ and the BPS masses

Maxwell's equations with magnetic charges added were symmetric under exchanging electric and magnetic ((6) in 17-02). Including the θ angle, the exchange is $\tau \to -1/\tau$,

and shifting θ by 2π gives $\tau \to \tau + 1$; the two generate the group $SL(2,\mathbb{Z})$ (integer entries, determinant 1) Known . SW set $a_D = \partial\mathcal{F}/\partial a$ and wrote the metric symmetrically in a and a_D.

$$ds^2 = \operatorname{Im}\left(da_D\, d\bar{a}\right), \qquad \tau_D = -\frac{1}{\tau} \quad (7)$$

With a_D as the coordinate the form is the same, and the coupling changes to $-1/\tau$. The original stresses that this is another description of the same theory Physics . The mass of a particle with electric charge n_e and magnetic charge n_m has a lower bound, and the mass m_{BPS} of particles equal to the bound (BPS states) is

$$m_{\mathrm{BPS}} = \sqrt{2}\,|n_e\, a + n_m\, a_D| \quad (8)$$

(Witten–Olive 1978, SW § 4) Physics . The W boson is $(n_m, n_e) = (0,1)$, the monopole $(1,0)$, and the dyons $(1,n)$. Example: classically $a_D = \tau_{\mathrm{cl}} a$, and the monopole is $|\tau_{\mathrm{cl}}| \approx 4\pi/g^2$ times heavier than the W, about 12.6 times for $g = 1$. On the strongly coupled side the ratio is reversed (Figure 2).

06

Singularities and monodromy — the points where the monopole becomes light

Differentiating the one-loop term of (4) gives $a_D \approx \frac{i}{\pi} a\left(\ln \frac{a^2}{\Lambda^2} + 1\right)$. For large u, $a \approx \sqrt{2u}$, so when u goes once round a large circle, $a \to -a$ and $\ln a^2 \to \ln a^2 + 2\pi i$, and $a_D \to -a_D + 2a$. Written as a matrix acting on the column vector $(a_D, a)^{\mathsf{T}}$ (the **monodromy**),

$$\mathsf{M}_\infty = \begin{pmatrix} -1 & 2 \\ 0 & -1 \end{pmatrix} \quad (9)$$

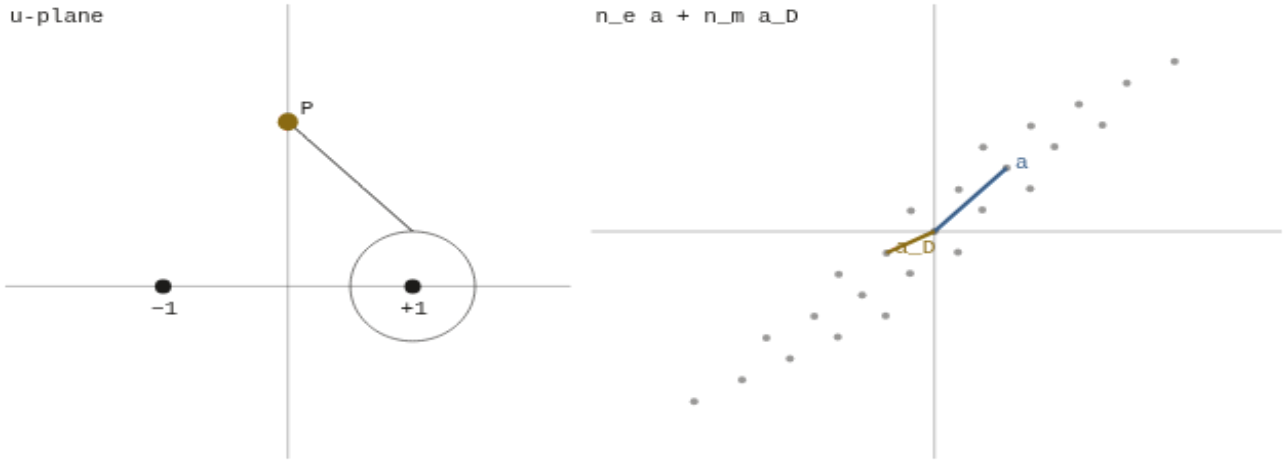
If every monodromy commuted with M_∞, then a^2 would be a global coordinate and we would be back at the contradiction of § 04. SW therefore proposed that there are two singularities at finite places (choosing units, $u = \pm 1$), where the mass of a monopole or a dyon becomes 0 Physics . Near $u = 1$ the effect of the light monopole gives $a \approx a_0 + (i/\pi) a_D \ln a_D$, and one turn gives $a \to a - 2a_D$.

$$\mathsf{M}_1 = \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}, \qquad \mathsf{M}_{-1} = \begin{pmatrix} -1 & 2 \\ -2 & 3 \end{pmatrix}, \qquad \mathsf{M}_1 \mathsf{M}_{-1} = \mathsf{M}_\infty \quad (10)$$

(The formulas of SW. In the same Figure 1, SW write the product in the order that puts the matrix of the loop traversed first on the right.) The charge $q = (n_m, n_e)$ of the

particle that becomes massless satisfies $q\mathsf{M} = q$; for M_1 it is $q = (1,0)$, that is, the monopole.

Figure 1 With "Play", u goes once round a singularity from $P = 1.5i$ (left), and a (blue) and a_D (yellow) move (right; the points are $n_e a + n_m a_D$). After one turn the pair is rewritten by a matrix



u = 0.000+1.500i, a = 1.256+1.256i, a_D = −0.832−0.424i

Computed The numbers in this figure were computed on this machine (the equation (PF) of § 07 solved along the path from P by the Runge–Kutta method; the matrices agree with Table 1).

Table 1 The matrices for one anticlockwise turn from the base point $P = 1.5i$ (deviation from integers at most 10^{-12})

Point encircled	Matrix	$q = (n_m, n_e)$ with $q\mathsf{M} = q$
$u = 1$: $\mathsf{M}(1)$	$\begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}$	$(1,0)$ monopole
$u = -1$: $\mathsf{M}(-1)$	$\begin{pmatrix} 3 & 2 \\ -2 & -1 \end{pmatrix}$	$(1,1)$ dyon
Both: $\mathsf{M}(\infty)$	$\begin{pmatrix} -1 & 2 \\ 0 & -1 \end{pmatrix}$	None

In Table 1 the matrix for going round path A and then path B is written $\mathsf{M}_A\mathsf{M}_B$ (the order of analytic continuation). Seen from P, the large loop is "-1 and then $+1$", so $\mathsf{M}(-1)\mathsf{M}(1) = \mathsf{M}(\infty)$, and $\mathsf{M}(-1) = \mathsf{M}_\infty \mathsf{M}_1^{-1} = \mathsf{M}_\infty \mathsf{M}_{-1} \mathsf{M}_\infty^{-1}$ Computed . It is conjugate to the M_{-1} of SW, and the charge of the massless dyon changes from $(1,-1)$ to $(1,1)$. This is an instance of the original's 'complete democracy among dyons' (conjugating by M_∞ shifts the charge by 2).

07

Writing with the periods of a curve — the periods of a torus and $\operatorname{Im}\tau > 0$

SW § 6 assigns to each u the curve

$$E_u:\ y^2=(x-1)(x+1)(x-u) \quad (11)$$

Gluing the two sheets of $y=\pm\sqrt{\cdots}$ along two cuts joining the branch points $-1,1,u,\infty$ gives a surface with one hole (a **torus**). A torus is the plane divided by a lattice $\mathbb{Z}\omega_1+\mathbb{Z}\omega_2$, and with the orientation of $\tau=\omega_2/\omega_1$ chosen consistently, $\operatorname{Im}\tau>0$ Known . The branch points coincide only at $u=\pm1,\infty$, which agree with the singularities. Set the periods to be

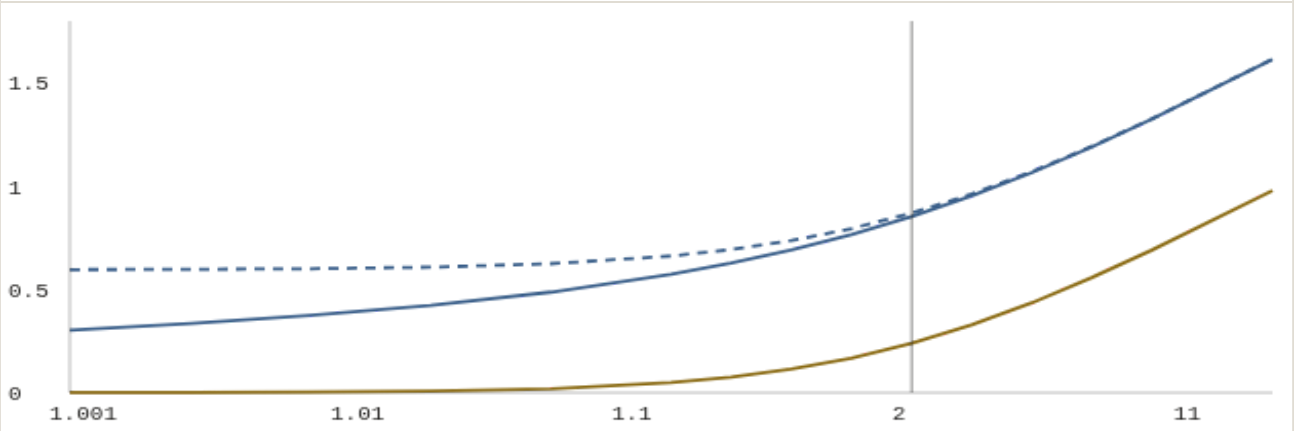
$$a=\frac{\sqrt{2}}{\pi}\int_{-1}^{1}\frac{\sqrt{x-u}}{\sqrt{x^2-1}}\,dx,\qquad a_D=\frac{\sqrt{2}}{\pi}\int_{1}^{u}\frac{\sqrt{x-u}}{\sqrt{x^2-1}}\,dx,\qquad \tau=\frac{da_D/du}{da/du} \quad (12)$$

Differentiating in u, the integrand becomes a constant multiple of dx/y, and da/du and da_D/du are integrals along the two loops of the torus, that is, periods of the lattice (times the same constant). τ is their ratio, so $\operatorname{Im}\tau>0$ holds automatically (the properties of the curve are Known ; the correspondence with the theory is Physics). The two periods satisfy the same equation $4(u^2-1)\Pi''+\Pi=0$ (the Picard–Fuchs equation, (PF) below) (relative residual at most 10^{-15} on the real axis $u>1$ Computed). Figure 1 integrates this equation numerically along the path. Near $u=1$,

$$a_D\approx\frac{i(u-1)}{2},\qquad a\approx\frac{4}{\pi}-\frac{(u-1)\ln(u-1)}{2\pi} \quad (13)$$

and the $a\to a-2a_D$ of (10) comes out. Example: the numerical integration of (12) agreed with the closed form in hypergeometric functions to within 10^{-8}, and $a(1)=4/\pi$ Computed .

Figure 2 Choosing u (>1) on the real axis with the knob lets you read the mass ratio of monopole to W, $|a_D|/a$ (yellow), and the coupling $\operatorname{Im}\tau=4\pi/g^2$ (blue; the dashed line is one loop, $(2/\pi)\ln 2a$)



u = 1.989, W: √2a = 2.773, monopole: √2|a_D| = 0.662, ratio 0.239, Im τ = 0.853 (one loop 0.870)

Computed The numbers in this figure were computed on this machine (15 points of the numerical integration of (12) embedded).

Table 2 Periods and coupling on the real axis $u > 1$ (in units that put the singularities at $u = \pm 1$)

u	a	$\lvert a_D\rvert/a$	$\operatorname{Im}\tau$	One loop
1.001	1.2750	0.0004	0.303	0.596
1.5	1.6776	0.1448	0.735	0.771
3	2.4320	0.3727	1.000	1.007
100	14.1421	1.4912	2.128	2.128

The one-loop column is $(2/\pi)\ln 2a$. As $u \to 1$, $\operatorname{Im}\tau$ goes to 0 and the electric coupling diverges.

Detour: at $u = 3$ we get $\tau = i$

A subsection you may skip.

Under $x \mapsto (x+1)/2$ the branch points move to $0, 1, 2, \infty$ and the curve becomes $y^2 = x(x-1)(x-2)$. Under $x \mapsto 2 - x$ the right-hand side changes sign, so $(x, y) \mapsto (2 - x, iy)$ carries the curve to itself and returns to the start after 4 steps. This is a symmetry that turns the lattice by 90° onto itself, and such a lattice is square Known . If it is square, τ is a value carried to i by $SL(2, \mathbb{Z})$, and in the normalisation of (12) it was exactly $\tau(3) = i$ Computed . Since $-1/\tau = \tau$, it is the point where the electric and magnetic couplings are equal.

08

Reading the instanton coefficients from the exact solution — a series with rational coefficients, checked against one loop

Expanding (12) at large u and solving inversely for a gives a series with rational coefficients c_k Computed .

$$u = \frac{a^2}{2} + \sum_{k\geq 1} c_k\, a^{2-4k}, \qquad c1 = \tfrac{1}{4}, \quad c2 = \tfrac{5}{64}, \quad c3 = \tfrac{9}{128}, \quad c4 = \tfrac{1469}{16384} \tag{14}$$

Matone (1995) introduced $\mathcal{G} = \pi i\big(\mathcal{F} - \tfrac{1}{2} a\, a_D\big)$, and the abstract says it is modular invariant. Putting $\mathcal{G} = u$ into (4), the one-loop term gives $a^2/2$ and the k-th term gives $2\pi i k$ times the term in a^{2-4k}, so the coefficient of a^{2-4k} in (4) reads as $c_k/(2\pi i k)$. The equation of the abstract, $(1 - \mathcal{G}^2)\mathcal{G}'' + \tfrac{1}{4} a\, \mathcal{G}'^3 = 0$, also becomes the same equation as (PF) on setting $\mathcal{G} = u$ Computed .

We check this reading with a_D. Example: at $u = 2$ the numerical integration of a_D/i gives 0.473434. One loop alone gives 0.462621, up to $k = 1$ 0.473083, and up to $k = 3$ 0.473432; each added term shrinks the difference by more than one digit Computed .

Dorey–Khoze–Mattis (1996) computed one and two instantons directly and reported 'precise agreement' Physics .

09

The perturbation to N=1 — monopole condensation and the mass gap

In the language of N=1, the N=2 set splits into the set (A_μ, λ) (the vector multiplet) and the set Φ of ϕ, ψ. Adding a term $m\,\mathrm{tr}\,\Phi^2$ that gives mass to Φ, an N=1 theory remains at low energies. The original describes this theory as 'strongly believed to generate a mass gap', and lists confinement, the breaking $\mathbb{Z}_4 \to \mathbb{Z}_2$ and that there are exactly two vacua (citing Witten 1982) Physics . $\mathbb{Z}_4$ is the symmetry generated by the transformation that multiplies the phase of λ by i, and the remaining $\mathbb{Z}_2$ is multiplication by -1.

Near $u = 1$ we use the dual photon set A_D and the monopole fields $M, \tilde{M}$ (names of fields, not masses). The holomorphic function that fixes the interactions (the superpotential) and the conditions for the vacuum are

$$\widehat{W} = \sqrt{2}\,A_D M\tilde{M} + m\,U(A_D); \qquad \sqrt{2}\,M\tilde{M} + m\,\frac{du}{da_D} = 0, \quad a_D M = a_D\tilde{M} = 0 \tag{15}$$

where $U(A_D)$ is the coordinate u written as a function of A_D for the whole set. If $m \neq 0$ and $du/da_D \neq 0$, then $M, \tilde{M} \neq 0$ is required, and $a_D M = 0$ gives $a_D = 0$, that is, $u = 1$. From (13), $du/da_D = -2i$, so the solution with $|M| = |\tilde{M}|$ has $|M|^2 = \sqrt{2}\,|m|$ (in the units of § 06) Computed .

The monopoles condense. The dual photon acquires mass by the Higgs mechanism, and the original writes 'Expanding around this vacuum, it is easy to see that there is a mass gap'. What condenses is magnetic charge, so the electric field is squeezed into tubes and electric charges are confined (exactly the picture of 17-02) Physics . At $u = -1$ a dyon condenses, and there are two vacua. Example: for $m = 0.01$, $|M| = 0.119$ Computed . There are two scales, Λ and m, so the situation differs from "a single scale" of pure Yang–Mills (17-12).

10

The distance from pure Yang–Mills — the border between arguments of physics and theorems of mathematics

To reach pure Yang–Mills, two limits are needed: decoupling Φ with $m \to \infty$, and sending the mass of the fermion λ to infinity as well. The weak-coupling description of § 09 is for small m, and that the gap and confinement are not interrupted along the

limits remains an expectation; there is no means to control it (17-10) Physics . The whole argument also presupposes that the N=2 theory exists. The example of SW is the typical case where, in four dimensions, confinement was derived from a duality at the level of physics (semiclassical calculations also exist, such as Polyakov's in three dimensions; 17-02), but what remained on the side of mathematics took another form.

Table 3 Claims and levels

Claim	Level
The integral over the space of instantons with $\varepsilon 1, \varepsilon 2$ (the two parameters of rotations in four dimensions) put in has, as $\varepsilon 1, \varepsilon 2 \to 0$, a limit that agrees with the SW prepotential (Nekrasov's conjecture)	Known Theorem: Nakajima–Yoshioka, Nekrasov–Okounkov ($SU(N)$), Braverman–Etingof (general simple groups)
Invariants of four-manifolds made by counting the solutions of the Seiberg–Witten equations (a mathematical definition)	Known (3-06, the four-dimensions article)
That these invariants are equivalent to the Donaldson invariants (the abstract of Witten 1994: 'Recent developments ... suggest a new point of view', 'one can define equivalent four-manifold invariants')	Physics (the current state of proofs on the side of mathematics was not investigated in this chapter)
(11) and (12) describe the low energies of the N=2 theory, and for $m \neq 0$ condensation, confinement and a gap appear	Physics
The gap of pure Yang–Mills	This chapter says nothing

The mathematics in Table 3 consists of theorems and definitions about integrals over the space of instantons and about solutions of equations, not claims about the spectrum (4-11). Example: the coefficients of (14), once the conventions are aligned, are numbers guaranteed by a theorem as integrals over the space of instantons. That the same numbers are "properties of the N=2 quantum field theory" cannot be said, since the construction of the theory (7-01) is missing. It is accurate to read this as an instance, at the level of physics, of the "duality transformations" of § 6.6 of the problem statement (17-13, 16-05).

What this chapter can say

What is said	What is not said
Known (1), the curve (11) and the properties of the torus, the proof of Nekrasov's conjecture, the definition of the SW invariants, § 6.6 of the problem statement.	That the N=2 and N=1 theories exist as quantum field theories (they have not been constructed).

What is said	What is not said
Computed Tables 1 and 2, Figures 1 and 2, (PF), (14), the one-loop constant, $\tau(3) = i$, the $\lvert M\rvert^2 = \sqrt{2}\lvert m\rvert$ of the monopole field. All start from (12).	That (11) and (12) are derived from the N=2 theory (the determination by holomorphy and duality is an argument of physics). How far the equivalence of the SW invariants and the Donaldson invariants has been shown as mathematics (not investigated in this chapter).
Physics The points where the mass of the monopole becomes 0, condensation, confinement and the gap for $m \neq 0$, agreement with the direct instanton calculations.	Control of the two limits back to pure Yang–Mills, and what they imply for the gap. This book itself shows neither the mass gap nor the construction of the continuum limit (the published construction of the continuum limit is in Volume 2).

Sources and reproduction

Item	Kind	Source / tools
The formulas and quotations of SW ((2)–(4), (7)–(13), (15))	Physics	N. Seiberg, E. Witten, Nucl. Phys. B426 (1994) 19–52 (Erratum B430 (1994) 485), hep-th/9407087. Text of the TeX manuscript checked. From the reference list of the same paper (bibliographic data only): Seiberg, Phys. Lett. 206B (1988) 75; Witten–Olive, Phys. Lett. 78B (1978) 97; Witten, Nucl. Phys. B202 (1982) 253.
§ 6.6 of the problem statement	Known	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory” (the Clay problem description) § 6.6. Text of the PDF checked.
(1) and (5)	Known	The source in 17-01. $b = 4$ checked against the $(\Lambda/a)^{4k}$ and $\mathbb{Z}_{4N_c}$ of SW.
Matone in § 08	Physics	M. Matone, Phys. Lett. B357 (1995) 342–348, hep-th/9506102. Abstract only (the definition of $\mathcal{G}$, modular invariance, the equation for $\mathcal{G}$, 'we find u=u(a)'). $\mathcal{G} = u$ is from the check in § 08; the main text was not read.
Dorey–Khoze–Mattis in § 08	Physics	N. Dorey, V. V. Khoze, M. P. Mattis, Phys. Rev. D 54 (1996) 2921, hep-th/9603136. Abstract checked.
Nekrasov's conjecture and its proofs in Table 3	Known	N. Nekrasov, Adv. Theor. Math. Phys. 7 (2004) 831, hep-th/0206161; H. Nakajima, K. Yoshioka, Invent. Math. 162 (2005) 313–355, doi:10.1007/s00222-005-0444-1, math/0306198; N. Nekrasov, A. Okounkov, in *The Unity of Mathematics* (Progress in Mathematics, Birkhäuser) 525–596, hep-th/0306238; A. Braverman, P. Etingof, math/0409441. Abstracts checked.
The Seiberg–Witten invariants in Table 3	Known, Physics	E. Witten, “Monopoles and Four-Manifolds”, Math. Res. Lett. 1 (1994) 769–796, hep-th/9411102. Abstract checked. Mathematical textbook: J. W. Morgan, *The Seiberg–Witten Equations and Applications to the Topology of Smooth Four-Manifolds*, Princeton University Press, doi:10.1515/9781400865161 (bibliographic data only; year of publication not checked).

Item	Kind	Source / tools
Tables 1 and 2, Figures 1 and 2, (PF), (14), the numbers in §07–§08	Computed	Python 3 on this machine (numpy, scipy): `sw_periods.py` (the check of (12), the one-loop constant, Table 2, Figure 2), `sw_monodromy.py` (Table 1, the starting values of Figure 1, $\tau(3) = i$), `sw_pf_check.py` (the residual of (PF), agreement with Matone's equation), `sw_instanton.py` ((14) and the check of a_D).

Next chapter to read: 17-12 The sentence that plays the role of an equivalence principle — cutting out "a single scale" as an assumption in one sentence.

The sentence that plays the role of an equivalence principle — cutting out "a single scale" as one assumption

> This chapter is this book's own investigation. § 01, § 02, § 08 and § 09 are established knowledge; § 10 is a map of open questions.

What this chapter shows — We take out of the path of pure Yang–Mills one sentence that stands in the same place as the equivalence principle at the start of general relativity: "the tube of confinement has no length parameter other than the string tension". We follow in formulas how joining four lengths by three chains gives "the diffusive-relaxation-time assumption", and check it with the lattice tables and with examples where one scale breaks and where it holds.

Prerequisite chapters — 17-01 How physics sees it and 4-07 Dimensional transmutation and Λ. The Poincaré constant (1-10) and the correlation length (2-08) are defined on the spot as far as needed.

Said first — The correspondence with the equivalence principle is an analogy, not a mathematical equivalence. "A single scale" is not a theorem of the four-dimensional continuum theory, and neither the relaxation-time assumption nor any of the three chains has been proved. This chapter does not show the mass gap; it names the one sentence that remains unproved. The lattice numbers are values from primary sources.

The order of this chapter

1. General relativity began from one sentence — there is no second kind of mass
2. Detour: measuring a second kind of mass — the 10^{-15} of MICROSCOPE
3. A single scale — two kinds of length — the ball length and the string length line up
4. Looking at the lattice tables — the form of the extrapolation comes in from outside
5. Four lengths and three chains — splitting the relaxation-time assumption into a product of ratios

6. Which chain plays the role of the equivalence principle — a table of correspondences (an analogy)
7. Dropping candidates — the other three candidates
8. Examples where one scale breaks — U(1), supergravity, the leading strong-coupling term
9. Examples where one scale holds, and the input needed — three-dimensional $\mathbb{Z}_2$, and I and R
10. Where it differs from the equivalence principle — it cannot be put as an assumption

---

01

General relativity began from one sentence

General relativity is sometimes explained as a combination of the **equivalence principle**, the **constancy of the speed of light**, **general covariance** (the form of the laws is the same in any coordinates) and **Riemannian geometry**. Riemannian geometry is a tool of mathematics, the constancy of the speed of light is a premise inherited from special relativity (setting $c = 1$ is a convention of units), and general covariance is a requirement on how laws are written. On top of these, the one sentence of physics newly added about gravity is the equivalence principle. We write falling in Newtonian mechanics.

$$m_{\mathrm{i}}\, a = m_{\mathrm{g}}\, g \quad \Longrightarrow \quad a = \frac{m_{\mathrm{g}}}{m_{\mathrm{i}}}\, g \tag{1}$$

m_{i} is the **inertial mass** (how hard a body is to move), and m_{g} is the **gravitational mass** (how strongly gravity pulls on it). They are defined by separate experiments, and the theory gives no reason for them to be equal. The equivalence principle says that the ratio $m_{\mathrm{g}}/m_{\mathrm{i}}$ is the same for every body; that is, **there is no second kind of mass**. Then gravity disappears locally inside a freely falling box, and gravity can be written as the geometry of spacetime (said to begin with Einstein 1907; the spacetime article) Physics. Example: if the ratio differs between materials by η, the difference in acceleration is about ηg (§ 02).

The question is: **does the path to the mass gap also contain "one sentence that is obvious physically and unproved mathematically"?** It does: "the tube of confinement has no length parameter other than the string tension". Its role, however, differs from that of the equivalence principle (§ 10).

02

Detour: measuring a second kind of mass

A section you may skip.

The difference in how two bodies fall is measured by the signed quantity $\eta = 2(a_1 - a_2)/(a_1 + a_2)$ (a_1, a_2 are the accelerations of the two bodies). Long ago there was the torsion balance of Eötvös and others (said to be a paper of 1922), and the satellite MICROSCOPE reported this final result with titanium and platinum Known .

$$\eta(\mathrm{Ti}, \mathrm{Pt}) = \left[-1.5 \pm 2.3\,(\mathrm{stat}) \pm 1.5\,(\mathrm{syst}) \right] \times 10^{-15} \quad (2)$$

Combining the errors in quadrature gives $\pm 2.75 \times 10^{-15}$, compatible with 0 Computed . More precisely, experiment only narrows the band and cannot show $\eta = 0$. Physics puts this one sentence as an axiom and keeps checking it.

03

A single scale — two kinds of length

Classical four-dimensional Yang–Mills has no constant with the dimension of mass, and in the quantum theory a single scale Λ appears (4-07). Physics takes a quantity whose mass dimension is k to be a constant multiple of Λ^k (for example the mass of the lightest glueball $m = c_1 \Lambda$ and the string tension $\sigma = c_2 \Lambda^2$; 17-01 § 03) Physics . Then the ratio $m/\sqrt{\sigma} = c_1/\sqrt{c_2}$ is a pure number fixed by the group alone. On the lattice (5-01) what is measured is am and $a\sqrt{\sigma}$ in lattice units; as $\beta \to \infty$ both go to 0, and in the ratio a cancels. Below, m, σ are numbers in lattice units. Making "a single scale" into an inequality uniform in β gives the following sentence (as the record of this book's study writes it).

$$c\,\sigma(\beta) \;\le\; m(\beta)^2 \;\le\; C\,\sigma(\beta) \qquad (\beta \ge \beta_0) \quad (3)$$

The constants $c, C > 0$ do not depend on β. It assumes neither $m > 0$ nor $\sigma > 0$; it only says that if they go to 0, they do so at the same rate. In terms of length, we call the length $\xi = 1/m$ of the lightest glueball (glue plus ball) the **ball length** below (the correlation length: the distance over which correlations decay as $e^{-|x|/\xi}$; 2-08 § 03). $\xi_\sigma = 1/\sqrt{\sigma}$ is the **string length**. $A \asymp B$ means that the ratio A/B is bounded above and below by positive constants independent of β. (3) is the same as $\xi \asymp \xi_\sigma$. It is a claim of the same form as "two kinds of mass", and in the language of physics it is the statement that **no second dimensionless parameter appears**. If there were a second parameter z, the ratio would be $F(z)$, and could go to 0 or to infinity.

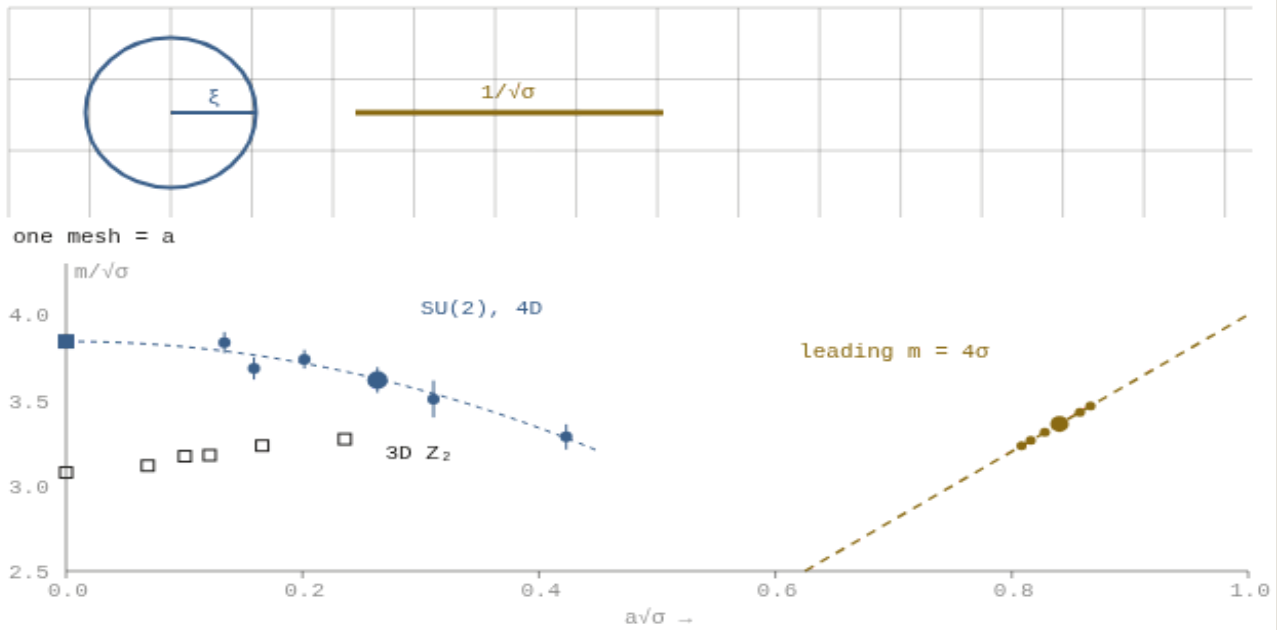
Example: for SU(2) at $\beta = 2.60$ (24^4), $a\sqrt{\sigma} = 0.13395$ and $am = 0.514$ (Lucini–Teper), so in lattice units $\xi_\sigma = 7.47$ and $\xi = 1.95$, and the ratio is 3.837 Computed . That this ratio stays bounded is the content of (3).

04

Looking at the lattice tables — the form of the extrapolation comes in from outside

These are the six SU(2) points of Lucini–Teper (2001). σ is the mass am_l of a tube of glue wound once round the periodic direction, plus the string-fluctuation correction $\pi/(3L)$, divided by the length L (in lattice steps) (equation (4) of the original paper: $am_l = a^2\sigma L - \pi/(3L)$). In using this formula the authors state 'we are assuming that we have linear confinement' (17-01 § 06).

Figure 1 The top is drawn in physical units (the string length $1/\sqrt{\sigma}$ fixed as the length of the bar). Raising β with the knob makes only the lattice mesh (spacing a) finer, and the ball length ξ (the radius of the circle) hardly changes. The bottom is the ratio $m/\sqrt{\sigma}$ (the square is three-dimensional $\mathbb{Z}_2$). The button overlays the leading strong-coupling term (dashed)



β = 2.4 (12⁴) a√σ = 0.2634 am = 0.953, in lattice units 1/√σ = 3.80、ξ = 1.05 m/√σ = 3.618, leading term 3.361

Table A $m/\sqrt{\sigma}$ for SU(2) (from Lucini–Teper Table 1. The master copy of the six values of am is 16-05 Table 3, and of $a\sqrt{\sigma}$ 5-09 Table 1)

β	Lattice	$a\sqrt{\sigma}$	am	$m/\sqrt{\sigma}$
2.25	8^4	0.4231(25)	1.390(30)	3.285(74)
2.30	10^4	0.3108(17)	1.090(33)	3.507(108)
2.40	12^4	0.2634(14)	0.953(19)	3.618(75)

β	Lattice	$a\sqrt{\sigma}$	am	$m/\sqrt{\sigma}$
2.475	16^4	0.2016(13)	0.754(10)	3.740(55)
2.55	20^4	0.15896(63)	0.586(10)	3.686(65)
2.60	24^4	0.13395(62)	0.514(8)	3.837(62)

Computed The numbers in this figure and table were computed on this machine (ratios of values from primary sources; the errors are a rough guide, treated as independent).

The string length in lattice units, $1/(a\sqrt{\sigma})$, grows from 2.36 to 7.47, but the ratio stays within 3.29–3.84, and broadly increases, approaching from below (it falls once from 2.475 to 2.55, within the errors). A fit with a constant independent of a is rejected with $\chi^2 = 38.6$ (5 degrees of freedom), so the ratio depends on a Computed . The continuum value is extrapolated after fixing the form of this dependence.

$$\frac{m}{\sqrt{\sigma}}(a) = r_0 + c\,(a\sqrt{\sigma})^p \quad (4)$$

Table B Fits of (4) (least squares weighted by the errors, all 6 points)

p	Continuum value r_0	χ^2 / dof
2	3.844(43)	2.39 / 4
1	4.053(73)	2.73 / 4
Primary source (Table 5)	3.844(61)	—

$p = 2$ agrees with the primary source in the central value, but $p = 1$ gives a χ^2 just as good, with a continuum value of 4.05 Computed . **The data alone cannot choose the form of the extrapolation.** The authors write 'the functional form of such corrections is known' and use $O(a^2)$, that is, $p = 2$ (5-05). 3.844 is a number resting on the premise that the limit exists and is approached in this form (17-01 § 08) Physics .

05

Four lengths and three chains — the diffusive-relaxation-time assumption

There are four lengths, each counted in lattice steps: the **bookkeeping length** ξ_{RG}, the string length ξ_σ, the ball length ξ, and the **length over which fluctuations are smoothed out** $\sqrt{CP}$.

The bookkeeping length is the length fixed by writing down, for each step that coarsens the lattice by a factor b (one step of the renormalisation group, 7-05), how much the coupling β falls, and piling up the steps. After k steps lengths are multiplied by b^k, so the number of steps needed to fall to a coupling of a fixed strength decides "how many lattice steps make one physical unit". We call that the bookkeeping length, and in this unit set the lattice spacing to $a := 1/\xi_{\mathrm{RG}}$ (the definition in the record of this

book's study; 23-02). Counting by the one-loop running alone (4-06), for SU(2), up to a constant factor, $\xi_{\mathrm{RG}} \approx e^{3\pi^2\beta/11}$ ($\beta = 4/g^2 = 2N/g^2$ of the table of notation; the coefficient in the exponent is a rewriting of the one-loop coefficient b_0) Known .

C_P is, for the probability distribution μ of the lattice gauge field (SU(2), Wilson action, infinite volume, no gauge fixing),

$$\mathrm{Var}_\mu(f) \;\le\; C_P\,\mathcal{E}(f), \qquad \mathcal{E}(f) = \mathbb{E}_\mu\Big[\sum_\ell |\nabla_\ell f|^2\Big] \quad (5)$$

the best constant for which this holds for every f (the Poincaré constant). Var_μ is the variance under μ, and ∇_ℓ is the derivative in the SU(2) directions of the link ℓ. Depending on whether the derivative is taken in the angle ϑ of the link ($U = e^{i\vartheta\cdot\sigma}$) or in the field $A = \sqrt{\beta}\,\vartheta$, $\mathcal{E}$ changes as $\mathcal{E}_A = \mathcal{E}/\beta$, and C_P differs by a factor β. In the free approximation $C_P = (g^2/4)\xi^2$ for the former and $C_P = \xi^2$ for the latter. Whether the band (6) below holds depends on this choice (with the angle normalisation the lower side breaks logarithmically even in the free approximation: 18-05 § 03). The best constant is the inverse of the gap (1-10 § 07). The dynamics at work is **Langevin dynamics**, a stochastic time evolution that moves each link little by little in the direction that lowers the action while adding a small random shake. Run for long, the distribution settles to μ. $\lambda_1 = 1/C_P$ is the gap of this dynamics, and C_P is the time of the slowest return. More precisely, C_P equals the return time only for reversible dynamics in continuous time (Langevin); for discrete-time updates or HMC it is a different quantity (18-03 § 06). The assumption the record of this book's study makes is that the length of fluctuations is on the same scale as the bookkeeping length, which in band form is (6) below. The names follow A-06 Index of names. **The relaxation-time assumption** is the statement at the level of powers $\lim \ln C_P / \ln \xi = 2$, and (6) is the band statement H_{band} (naming the normalisation, $H_{\mathrm{band}}[\mathcal{E}_\vartheta]$ or $H_{\mathrm{band}}[\mathcal{E}_A]$) combined with the chains (ii)(iii).

$$a^2 C_P \asymp 1 \qquad (\text{i.e. } \sqrt{C_P} \asymp \xi_{\mathrm{RG}}) \quad (6)$$

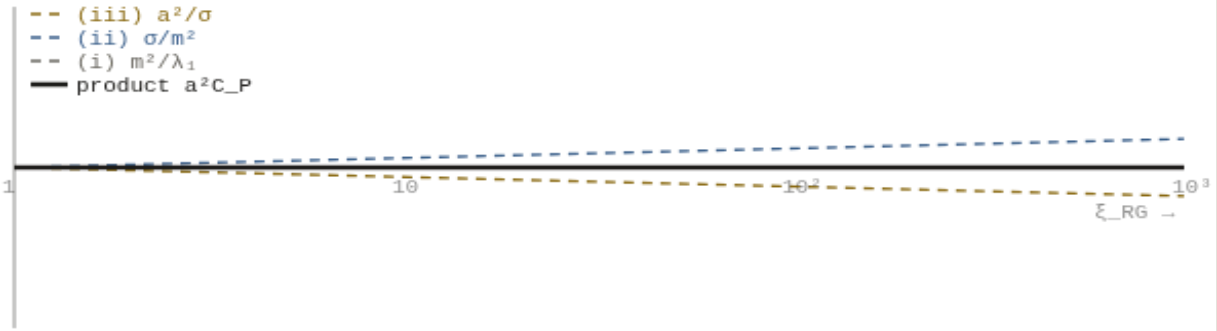
We split this into a product of three ratios.

$$a^2 C_P \;=\; \underbrace{\frac{a^2}{\sigma}}_{\text{(iii)}} \cdot \underbrace{\frac{\sigma}{m^2}}_{\text{(ii)}} \cdot \underbrace{\frac{m^2}{\lambda_1}}_{\text{(i)}} \quad (7)$$

That each factor is $\asymp 1$ we call one **chain**. (i) $\lambda_1 \asymp m^2$: the fluctuation length and the ball length line up (the relaxation chain). Langevin dynamics is diffusive, so smoothing out fluctuations of length ξ takes a time of order ξ^2 (dynamical exponent $z = 2$) Physics . That is why λ_1 is compared not with m but with m^2. (ii) $m^2 \asymp \sigma$: the ball length and

the string length line up (one scale, (3)). (iii) $\sigma \asymp a^2$: the string length and the bookkeeping length line up (confinement happens at the bookkeeping scale). If all three hold, (6) follows. Under (ii)(iii), (6) is the same as the band statement $C_P \asymp \xi^2$, and the double-well relaxation-time power statement follows from it (A-06). The converse does not follow: if the factors cancel each other, the product is constant even when chains are broken (Figure 2). The algebra in the direction from (7) to (6) is machine-checked Lean (`chain_band`, `sq_times_poincare_band`, `LatticeGaugeOneScale.lean`. The hypotheses are three band inequalities, and the conclusion is that $a^2 C_P$ lies in a positive band. The list is in 23-13), but none of the three chains themselves has been proved.

Figure 2 Changing the powers p of the three factors (factor $\propto \xi_{\mathrm{RG}}^{p}$) with the knobs tilts the three dotted lines and the solid line of the product $a^2 C_P$. If the product is horizontal, (6) holds. In the initial position ((iii) −0.3, (ii) +0.3) two chains are broken, yet the product is horizontal



Product a²C_P ∝ ξ_RG^(+0.00) (6) : holds, chains: (i) ○ (ii) × (iii) ×

Computed A schematic drawing of the sum of the logarithms in (7), drawn on the spot (not a measurement).

06

Which chain plays the role of the equivalence principle

(i) is a relation between the dynamics chosen as a tool and the gap of the theory (the gap of the transfer matrix, 5-03); change the dynamics and the statement changes too. (iii) depends on how the steps are booked. If we call the small error dropped at each coarsening step the **shortfall**, (iii) is governed by the one-loop coefficient and by the shortfalls summed over the steps (23-02). In the record's arrangement, only (ii) is a sentence of the theory that depends neither on the tool nor on the bookkeeping.

The one sentence on this path

The tube of confinement has no length parameter other than the string tension ($m^2 \asymp \sigma$, equation (3)).

In the same pattern as "there is no second kind of mass", the record made the following table of correspondences.

General relativity	This path (analogy)
Riemannian geometry (the stage)	Lattice gauge theory (a probability measure in finite dimensions)
General covariance (a requirement on how laws are written)	One step of the renormalisation group (coarsening the lattice leaves the content the same)
The convention of units coming from the constancy of the speed of light ($c = 1$)	The convention of setting the scale (the convention of units that sets $\sqrt{\sigma} = 1$; 17-01 § 07)
The equivalence principle (one sentence of physics)	(ii) One scale (there is no second length parameter)

As a metaphor, just as general relativity made "weight" one kind, this sentence makes "length" one kind. More precisely, the table is an analogy and no theorem joins its left and right; the equivalence principle is an equality at each point, while (3) is a uniform inequality.

07

Dropping candidates — dual superconductivity, centre vortices, asymptotic scaling

(a) Dual superconductivity and **(b) centre vortices** (17-02) are statements about configurations after fixing a gauge and projecting. In raw form they depend on how the gauge is fixed (Gribov copies, 17-02 § 04), and do not become a sentence about the probability measure. The core that can be written gauge-invariantly ('t Hooft's vortex free energy) is a statement about σ, and merges into chain (iii). **(c) Non-perturbative asymptotic scaling** is the sentence that lengths in lattice units grow as perturbation theory predicts. The two-loop prediction is the following f_2 (SU(N), $g^2 = 2N/\beta$) Known .

$$f_2(\beta) = (b_0 g^2)^{-b_1/(2b_0^2)} \exp\Big(-\frac{1}{2b_0 g^2}\Big), \qquad b_0 = \frac{11N}{48\pi^2}, \quad b_1 = \frac{34N^2}{3(16\pi^2)^2} \tag{8}$$

The coefficients agree with the values in the PDG review (without quarks) (checked at $N = 3$ Computed). (c) says "the inverse of every length in lattice units, for example am/f_2 or $a\sqrt{\sigma}/f_2$, converges to a positive constant". If f_2 is read as the scale marking of physical units, the convergence of am/f_2 is itself the statement that the mass gap in physical units is positive and finite. Since it contains the conclusion we want to show, the record set it aside as circular.

It differs numerically too. For SU(2) with the bare coupling $g^2 = 4/\beta$, as $\beta = 2.40 \to 2.60$, $a\sqrt{\sigma}/f_2$ goes 57.26→48.23 (−15.8 %), while the ratio goes 3.618→3.837 (+6.1 %) Computed . The ratio moves less, and does not use perturbation theory either.

The bare coupling β carries large corrections peculiar to the lattice. Lucini–Teper use the mean-field improved coupling $\beta_I = \beta \cdot \frac{1}{N}\langle \mathrm{Re}\,\mathrm{Tr}\,U_p\rangle$ (U_p is the matrix of the plaquette) and write 'Transforming this to β_I is known to remove a large part of the lattice corrections at small a.' (§ 5 of the original paper, the sentence after equation (11); as the origin of this coupling the original paper cites earlier literature) Physics . The values obtained by finding the two-loop Λ with this coupling and extrapolating to the continuum limit are in Table 6 of the original paper, but for the SU(2) extrapolation it says 'the best continuum fit is very poor'.

08

Examples where one scale breaks

That (3) is not obvious is seen from examples where it breaks. **Three-dimensional compact U(1)**: Göpfert–Mack proved $\sigma > 0$ at all couplings Known , and also state that in the continuum limit, with the mass fixed in physical units, the string tension goes to infinity ($\sigma/m^2 \to \infty$) (from the abstract of the same paper; Villain action; the main text not read). If that description is right, there is confinement and yet the ratio $m/\sqrt{\sigma}$ goes to 0 (17-02 § 03). **The supergravity regime of holography**: the ratio goes to 0 as $\lambda^{-1/2}$, and $\lambda = g^2 N$ is the second parameter (17-09 § 08) Physics .

The leading term of the strong-coupling expansion is not a counterexample but an approximation with the opposite slope. With the action $S = \beta \sum_p \left(1 - \frac{1}{2}\,\mathrm{Tr}\,U_p\right)$ (the sum over all plaquettes, 5-01), the leading term for SU(2) gives, with $u = I_2(\beta)/I_1(\beta)$ (a ratio of modified Bessel functions), $\sigma = -\ln u$, and the lightest state is $m = -4\ln u = 4\sigma$ (5-02) Known , so

$$\frac{m}{\sqrt{\sigma}} = 4\,a\sqrt{\sigma} \quad (9)$$

which is the dashed line in Figure 1 (a straight line that would pass through the origin if extended). Extended to $a\sqrt{\sigma} \to 0$ the ratio falls to 0, the opposite direction to the lattice numbers. The leading term at $\beta = 2.40$ gives $a\sqrt{\sigma} = 0.840$ (measured 0.2634) Computed , outside the region where the expansion converges.

09

Examples where one scale holds, and the input needed

An example where it holds: the three-dimensional $\mathbb{Z}_2$ gauge model. The group is abelian, but by a duality transformation it maps to the three-dimensional Ising model; the glueball mass becomes the inverse of the correlation length, and the string tension becomes the interface tension. Agostini–Carlino–Caselle–Hasenbusch (1997), taking it

that $\sigma\xi^2$ should be constant in the scaling limit, found the continuum value $0.1056(19)$, checked that it agrees with the prediction $0.1024(88)$ of ϕ^4 theory, and obtained $m/\sqrt{\sigma} = 3.08(3)$ Physics .

The ratio $1/\sqrt{\sigma\xi^2}$ is 3.272(11) at $a\sqrt{\sigma} = 0.236$, 3.117(52) at 0.069, and 3.077(28) in the continuum, approaching from above (from Table 6 of the original paper Computed). The physical explanation is this. A **relevant direction** is a direction of couplings whose effect grows as the lattice is coarsened. Within the range that keeps the $\mathbb{Z}_2$ symmetry (in the Ising model, with no magnetic field), there is only one relevant direction near the critical point, so m and $\sqrt{\sigma}$ are fixed by one and the same ξ (2-08 § 07) Physics . It is not a theorem.

If the description of the continuum limit of three-dimensional U(1) is right, one scale breaks for the abelian U(1) and holds for the abelian $\mathbb{Z}_2$, so **it is not decided by whether the group is non-abelian**. What is needed is "an input that forbids a second small parameter". The record takes this input in four-dimensional SU(2) to be the positivity of the one-loop coefficient (asymptotic freedom, 4-06), and rewrote it in two forms. I: there is one relevant direction, and m and $\sqrt{\sigma}$ are fixed by the same ξ. R: the rotational symmetry broken by the lattice comes back at long distances (the shape of the tube matters; 17-03 § 08). Whether these two are enough is not known.

10

Where it differs from the equivalence principle — a sentence that cannot be put as an assumption

In general relativity the equivalence principle is an axiom. The prize problem asks us to *prove* existence and the gap from the definitions, so (ii) cannot be added as an axiom; it has to be made a theorem or bypassed.

The problem statement defines the gap Δ as there being no spectrum in $(0, \Delta)$, and continues 'The supremum of such Δ is the mass m, and we require m < ∞.' Known . Splitting the form $C_P \asymp \xi^2$ of (6) under (ii)(iii) (the band statement H_{band}) into upper and lower halves, in the record's reading $C_P \lesssim \xi^2$ (the relaxation-time upper-bound assumption) corresponds to the side of $\Delta > 0$, and $C_P \gtrsim \xi^2$ (the relaxation-time lower-bound assumption) to the side where the mass does not collapse to infinity (the form $m < \infty$). The precise treatment that splits the relaxation-time assumption itself into upper and lower halves is in 23-01. The lower-bound assumption is a different thing from "non-trivial" in the Clay sense (that there are interactions).

The lower half of (ii), $m^2 \gtrsim \sigma$, reduces to an operator inequality saying that the deficit of the tube is at least c/ℓ for perimeter ℓ (called the tube transfer-operator inequality in

the record) Lean (`mass_sq_ge_of_tube`, `LatticeGaugeTubeEntropy.lean`. Hypothesis: the norm of the transfer operator is at most $e^{-2\sqrt{c\sigma}}$, that is, a consequence of the tube inequality. Conclusion: $m^2 \geq 4c\sigma$. The list is in 23-13). There is no pattern of proof for the tube inequality itself (23-02). From finitely many measurements only an upper bound on m follows (a restatement of a known relation; 23-02). The one sentence can be named, but there are not yet tools to make it a theorem.

What this chapter can say

What is said	What is not said
Physics That "a single scale" can be written as (3), and that (ii) is the only chain that depends neither on the tool nor on the bookkeeping (the record's arrangement).	That the table of correspondences is a mathematical equivalence (it is an analogy). That (ii) is a sufficient input for the relaxation-time assumption.
Computed Tables A and B, Figures 1 and 2, the numbers in § 02, § 07, § 08 and § 09.	These are ratios and fits of values from primary sources, not measurements. The data alone cannot choose the form of the extrapolation.
Lean The algebra of the product in the direction from (7) to (6) (`chain_band`, `sq_times_poincare_band`). The reduction of the lower half of (ii) to the tube inequality (`mass_sq_ge_of_tube`).	Proofs of the chains (i)(ii)(iii), the relaxation-time assumption, (3) and the tube inequality. This chapter shows nothing about one scale, the mass gap or the construction of the four-dimensional continuum theory.
Known The area law at all couplings for three-dimensional U(1), the form of the leading term.	The continuum limit of three-dimensional U(1) (main text not checked). One scale for $\mathbb{Z}_2$ and SU(2) is at the level of numbers and physics.

Sources and reproduction

Item	Kind	Source / tools
The definitions of the gap and of $m < \infty$	Known	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory" (the Clay problem statement) § 4. Text checked.
The starting point of the equivalence principle, the torsion balance	Physics	A. Einstein, Jahrb. Radioakt. Elektron. 4 (1907) 411–462. R. v. Eötvös, D. Pekár, E. Fekete, Ann. Phys. (Leipzig) 68 (1922) 11. Both bibliographic data only (secondary sources).
Equation (2)	Known	P. Touboul et al. (MICROSCOPE), PRL 129 (2022) 121102, arXiv:2209.15487. Abstract checked.
The inputs of Tables A and B and § 07, quotations, the mean-field improved coupling	Physics	B. Lucini, M. Teper, JHEP 0106 (2001) 050, hep-lat/0103027 Tables 1, 5 and 6, equations (4)(5) of the original paper, the sentence after equation (11) in § 5. Text checked.

Item	Kind	Source / tools
The $\mathbb{Z}_2$ inputs of § 09, the duality transformation, the ϕ^4 prediction 0.1024(88), 3.08(3)	Physics	V. Agostini, G. Carlino, M. Caselle, M. Hasenbusch, NPB 484 (1997) 331, hep-lat/9607029, equations (4)(15)(16) and Table 6 of the original paper. Text checked.
The leading term of the strong-coupling expansion ($u = I_2/I_1$, $m = 4\sigma$)	Known	J.-M. Drouffe, J.-B. Zuber, Phys. Rep. 102 (1983) 1–119. Bibliographic data only. The derivation is in 5-02.
't Hooft's vortex free energy (§ 07)	Known	G. 't Hooft, NPB 138 (1978) 1–25. Tomboulis–Yaffe, CMP 100 (1985) 313. Both bibliographic data only. 17-02.
The one-loop coefficient	Known	D. J. Gross, F. Wilczek, PRL 30 (1973) 1343. H. D. Politzer, PRL 30 (1973) 1346. Bibliographic data only.
Check of the coefficients of (8)	Known	J. Huston, K. Rabbertz, G. Zanderighi, "Quantum Chromodynamics", PDG, PRD 110 (2024) 030001, equation (9.3). Text checked.
Three-dimensional compact U(1)	Known	M. Göpfert, G. Mack, CMP 82 (1982) 545–606. Abstract (INSPIRE). Main text not read. The year is Crossref's year of publication; INSPIRE gives 1981.
The supergravity regime	Physics	The table of sources in 17-09.
(3), the relaxation-time assumption, the three chains, the table of correspondences, the judgement of candidates, I and R	Analogy, assumption	The arrangement of the record of this book's study. The chains and the relaxation-time assumption are assumptions, and the table of correspondences is an analogy. Lean covers only the algebra of (7) (`chain_band`, `sq_times_poincare_band`, `LatticeGaugeOneScale.lean`) and the reduction of the lower half of (ii) (`mass_sq_ge_of_tube`, `LatticeGaugeTubeEntropy.lean`). The list of theorems is in 23-13.
Tables A and B, Figures 1 and 2, the numbers in § 07, § 08 and § 09, the checks	Computed	Python 3 on this machine (numpy, scipy): `onescale.py` (propagation of errors, the least squares of (4), (8), I_2/I_1, the ratio for $\mathbb{Z}_2$).

Next chapter to read: 17-13 Where the way physics digs parts from mathematics — picking up, from the whole of Part 17, the places where physics and mathematics part. For the detailed form of the relaxation-time assumption, go to 23-01.

Where the way physics digs parts from mathematics — translating a sentence of physics into a sentence of mathematics brings the assumptions into the open

> This chapter organises established knowledge. § 04 is this book's own investigation.

What this chapter shows — The premises that come attached when a sentence of physics is translated into a mathematical statement. The seven moves through which premises enter, and where the theorems stop in each. How much of the string picture is in the denominator of the lattice ratio $m/\sqrt{\sigma}$ (recomputed from the tables of the primary source).

Prerequisite chapters — 17-10 (together with 17-01–17-03, 5-01 and 4-10).

Said first — This is an arrangement to avoid mistaking the level of results of physics for theorems; it does not discuss whether the methods of physics are good or bad. This chapter shows neither the existence of the mass gap nor the construction of the continuum limit. The numbers are recomputations from the tables of the primary sources.

The order of this chapter

1. Translating a sentence of physics into a sentence of mathematics
2. Measuring after putting that it exists
   - Detour: why the gap in a finite box is easy
3. The string picture enters the denominator
4. How much the ratio moves when the string picture is removed (Figure 1)
5. Extrapolating from finite β
6. Looking from the side of a small parameter
7. Moving to a neighbouring theory and coming back with a knob (Figure 2)
8. Looking after choosing a gauge
9. What physics regards as one

10. A list of the places where they part

01

Translating a sentence of physics into a sentence of mathematics

Table 5 of Lucini–Teper 2001 gives 3.844(61) as the continuum value of the ratio of the mass $m_{0^{++}}$ of the lightest glueball of $SU(2)$ to the string tension σ Physics . Translated into a mathematical statement, this sentence is

$$\lim_{\beta\to\infty} \frac{m(\beta)}{\sqrt{\sigma(\beta)}} = 3.844\ldots \quad (1)$$

($\beta = 2N/g^2$, and m and σ are quantities in units of the lattice spacing a; 5-01). (1) brings in four premises.

a. For each β the infinite-volume limit exists and is unique. Subsequential limits always exist by compactness, so the question is uniqueness.

b. For each β, $m(\beta) > 0$ and $\sigma(\beta) > 0$.

c. The limit $\beta \to \infty$ exists.

d. The limit is a ratio of quantities of a continuum theory satisfying the axioms (4-10).

For four-dimensional $SU(2)$ at large β, none of (a)–(d) is a theorem (§ 05) Known . "(61)" is an estimated numerical error, not an error bound given by inequalities. What this chapter calls "parting" is a place where some stage of the procedure of physics has no corresponding theorem and can only be put as an assumption.

02

Measuring after putting that it exists

The calculations of physics presuppose that a four-dimensional pure Yang–Mills theory exists, and is one. In a finite box B (a finite cube of lattice points) and at finite β, the probability measure on the link variables $U_\ell \in G$,

$$d\mu_{B,\beta}(U) = \frac{1}{Z_{B,\beta}} \exp\Big(\frac{\beta}{N} \sum_{P\subset B} \operatorname{Re}\operatorname{tr} U_P\Big) \prod_{\ell\subset B} dU_\ell \quad (2)$$

is properly defined (U_P is the product of the link variables once round the plaquette P, and dU_ℓ is Haar measure) Known . The premises enter in the limits. The problem statement does not accept weak existence by compactness (convergence of

subsequences) alone as a solution, unless the properties of the limit (the gap and the axioms) are shown by other means (footnote 2 of § 6.5 of the problem statement, 17-01). Further, in § 5 the problem statement names a **uniform** gap: 'the existence of a uniform gap for finite-volume approximations may play a fundamental role in the proof of existence of the infinite-volume limit' Known . The numbers of physics are "predictions made after assuming that it exists". For the construction published on 2026-11-10, which builds a theory satisfying the Wightman axioms as a limit of Wilson's lattice gauge theory, the committee's assessment is that it is proved in Lean. However, the infinite-volume limit is taken along a subsequence, and neither its uniqueness nor the mass gap, on which the numbers of physics rest, is shown ((a) above is a statement at each β of the lattice and lies outside this construction) (6-04).

Detour: why the gap in a finite box is easy

A subsection you may skip.

The β of this detour is the inverse temperature of the Ising chain (not the β of lattice gauge theory). The transfer matrix of the one-dimensional Ising chain (weight $e^{\beta s_i s_{i+1}}$, $s_i = \pm 1$) is

$$T = \begin{pmatrix} e^{\beta} & e^{-\beta} \\ e^{-\beta} & e^{\beta} \end{pmatrix}, \qquad \lambda_0 = 2\cosh\beta, \;\; \lambda_1 = 2\sinh\beta, \;\; \ln\frac{\lambda_0}{\lambda_1} = \ln\coth\beta \tag{3}$$

and correlations decay as $(\tanh\beta)^{|i-j|}$. The largest eigenvalue of a matrix whose entries are all positive is simple and strictly larger than the others (Perron–Frobenius) Known . Example: at $\beta = 1, 2, 3$ the gaps are 0.272, 0.0366 and 0.00496 Computed . It is positive at every finite β.

In lattice gauge theory too, the gap in a finite box is positive. More precisely, the link variables take values in a continuous group, so the transfer matrix is not a matrix but an integral operator with a positive continuous kernel. In a finite box the kernel is continuous and the group is compact, so this operator is compact. Then by an infinite-dimensional version of Perron–Frobenius (a generalisation of Jentzsch's theorem; the source is in 5-03) the largest eigenvalue is simple and isolated (the construction of the transfer matrix is Lüscher 1977 and Osterwalder–Seiler 1978; 5-03) Known . The question is whether, as $a \to 0$, $(\text{gap})/a$ remains positive uniformly in the volume (for two dimensions, 2-09).

03

The string picture enters the denominator

The σ in the denominator of (1) cannot be measured directly. What is measured is the mass m_l of a flux tube closed once round a periodic box (side L lattice points), and Lucini–Teper regard the tube as a string and subtract the Lüscher term (17-03 § 03) Physics .

$$am_l = a^2\sigma L - \frac{\pi}{3L} \quad (4)$$

The authors write 'in using eqn(4) we are assuming that we have linear confinement ..., that the effective string theory ... is in the Nambu-Gotto universality class, and that our string is in fact long enough', and also note 'with only two values of L ... we cannot claim to have evidence for both the 1/L functional form and for its coefficient'. The full form of the Nambu–Goto ground state (equation (23) of Athenodorou–Bringoltz–Teper) is

$$E_0(l)^2 = (\sigma l)^2 - \frac{\pi\sigma(d-2)}{3}, \qquad d = 4 \quad (5)$$

($l = aL$ is the length of the tube and d the dimension of spacetime; the primary source writes D). Taking the square root and expanding in $1/l$, the first two terms are (4) Known . That is, (4) is a long-string approximation, and the shorter the tube, the more the later terms matter. On the lattices of the primary source the tube lengths are $L\sqrt{\sigma} \approx 3.1$ to 3.4 (lengths in units of $1/\sqrt{\sigma}$) Computed , and that (4) is enough at these lengths is, as in the quotation above, put as an assumption by the authors too.

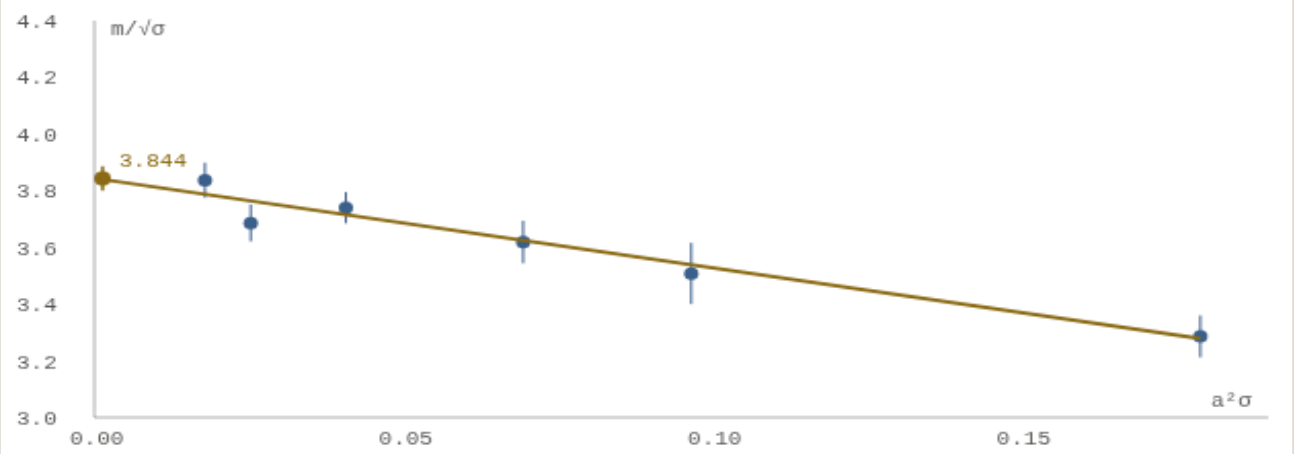
04

How much the ratio moves when the string picture is removed

From the tube masses in Table 1 of the primary source we extracted σ in three ways (no correction $a^2\sigma = am_l/L$, (4) and (5)) and recomputed the ratio Computed . No correction is the ruler that measures the contribution of the string picture.

How to read the fits — The ratio at each β is placed against the horizontal axis $a^2\sigma$ (the coarseness of the lattice), each point is weighted by the inverse square of its error, and a straight line is drawn (weighted least squares; the form of the line is (6) of § 05). The value at the left end $a^2\sigma = 0$ is the continuum value. χ^2 is the sum of the squares of the deviations between points and line, each divided by the error of the point; if it is of about the same size as the degrees of freedom (number of points − 2), the line fits the points within the errors. The errors of the ratios in Table 1 are the square root of the sum of the squared relative errors of $am_0{}^{++}$ and $a\sqrt{\sigma}$, treated as independent.

Figure 1 Changing how σ is taken and how many coarse points are dropped changes the fitted line and the continuum value (left end). The dashed line is equation (4), all points



Equation (4), points dropped: 0 → continuum value 3.844(43), χ²/dof 2.38/4, difference from equation (4), all points: 0.0 %

Computed The numbers in this figure were computed on this machine (the results of the same calculation as Table 1 embedded).

Table 1 $m_0{}^{++}/\sqrt{\sigma}$ for $SU(2)$, by how σ is taken (from Lucini–Teper Table 1. The master copy of the six values of am is 16-05 Table 3. How the errors are obtained is in "How to read the fits" above)

β	$L\sqrt{\sigma}$	**No correction**	**(4)**	**(5)**
2.25	3.38	3.447(78)	3.286(74)	3.278(73)
2.30	3.11	3.715(115)	3.508(108)	3.496(108)
2.40	3.16	3.825(80)	3.619(75)	3.608(74)
2.475	3.23	3.943(59)	3.740(55)	3.729(54)
2.55	3.18	3.894(69)	3.686(65)	3.676(64)
2.60	3.21	4.048(66)	3.837(62)	3.826(62)
Continuum value (all points)		4.059(45)	3.844(43)	3.833(42)

The central value 3.844 for (4) with all points agrees with Table 5 of the primary source. With or without the Lüscher term the continuum value moves by 5.6% (about 3.5 times the primary source's error 0.061), and the difference between (4) and (5) is 0.3% Computed . On the lattices of Table 1 the correction is 9–11% of the leading term σL. The precision of the ratio rests on **the leading correction of the string picture being right** (Athenodorou–Bringoltz–Teper report that many of the low states are well described by Nambu–Goto) Physics .

Translated into mathematics, the starting point of the string picture is a long tube of confinement, that is, confinement. Chatterjee 2018 writes that for four-dimensional non-abelian groups the area law at large β 'remains elusive', and that the area law at small β holds for all groups including $U(1)$, so it is no evidence Known .

> 05

Extrapolating from finite β

The primary source expects the lattice corrections of the plaquette action to start at a^2, and

$$\frac{m(a)}{\sqrt{\sigma(a)}} = \frac{m(0)}{\sqrt{\sigma(0)}} + c\,a^2\sigma \quad (6)$$

fits this (taking c as a constant), dropping the coarsest point if the fit is poor Physics. The knobs of Figure 1 reproduce this procedure; with (4) the continuum value stays within 3.84–3.86 Computed (for movement due to the form of the fit, see 17-12).

On the side of mathematics, the lattice correlation-length problem of Chatterjee 2018 (Chatterjee's Problem 5.1) asks for the following to be shown. Set $W_P = \operatorname{tr} U_P$, and let $f_\beta(x) = \langle W_{P_0} W_{P_x}\rangle - \langle W_{P_0}\rangle\langle W_{P_x}\rangle$ be the correlation (covariance) between the plaquette P_0 at the origin and the plaquette P_x nearest to the point $x \in \mathbb{R}^4$. For any infinite-volume limit of four-dimensional non-abelian lattice gauge theory, at every $\beta > 0$,

$$\lim_{|x|\to\infty} \frac{\log f_\beta(x)}{|x|} = -\frac{1}{\xi(\beta)}, \quad \xi(\beta) \in (0,\infty), \qquad \lim_{\beta\to\infty} \xi(\beta) = \infty \quad (7)$$

holds (ξ is the correlation length in lattice units; the continuum string-tension problem (Chatterjee's Problem 5.2) is the continuum limit of Wilson loops). In words: correlations decrease far away like $e^{-|x|/\xi}$, the length ξ is finite and positive at every β, and moreover grows without bound compared with the lattice spacing as the continuum limit is approached. The author writes 'Since none of the above has been proved', and states that at large β in dimensions higher than 2 even the uniqueness of the infinite-volume limit is unknown, and that $\beta_c = \infty$ in four dimensions is at the level of 'It is believed' Known. Example: even at the finest point ($\beta = 2.6$, 24^4), $\xi = 1/(am_{0^{++}}) = 1.95$ lattice spacings, and the ξ of the six points are 0.72–1.95 Computed. Extrapolation is the assumption that extends the slope over this range all the way to $\xi \to \infty$.

06

Looking from the side of a small parameter

Perturbation theory and semiclassics are both expansions around $g \to 0$, but the scale of dimensional transmutation (4-07) at one loop is

$$\Lambda = \mu\,\exp\Big(-\frac{1}{2b_0 g^2(\mu)}\Big), \qquad b_0 = \frac{11N}{48\pi^2} \quad (8)$$

$e^{-1/x}$ approaches 0 as $x \to 0^+$ faster than any power, and all its Taylor coefficients are 0 Known . If $\Delta = c\Lambda$, the gap appears at no order of the series in g^2 (an expectation) Physics (7-12, 4-12). Example: for $SU(3)$ at $g^2 = 0.5$, $\Lambda/\mu = 5.8 \times 10^{-7}$, while g^{20} is still 9.8×10^{-4} Computed .

The examples where semiclassics reaches the gap are none of them the theory of the problem statement. In three-dimensional compact $U(1)$, monopoles cause confinement (Polyakov 1977, semiclassical), and Göpfert–Mack proved the area law at all couplings on the lattice with the Villain action Known (the abstract of the same paper says that in the continuum limit the string tension in physical units goes to infinity; main text not read). On $\mathbb{R}^3 \times S^1$, Ünsal–Yaffe 2008 derived a gap and linear confinement for a deformed $SU(N)$ theory from a semiclassical analysis in the regime of small N where the circumference L is small enough (at the level of physics) Physics . That there is no phase transition on the way to $L \to \infty$ is a conjecture, based on the argument that there is no order parameter distinguishing small L from large L. What mathematics needs is an estimate non-perturbative in g, and the constructive methods too stop where they cross over to strong coupling (7-10).

07

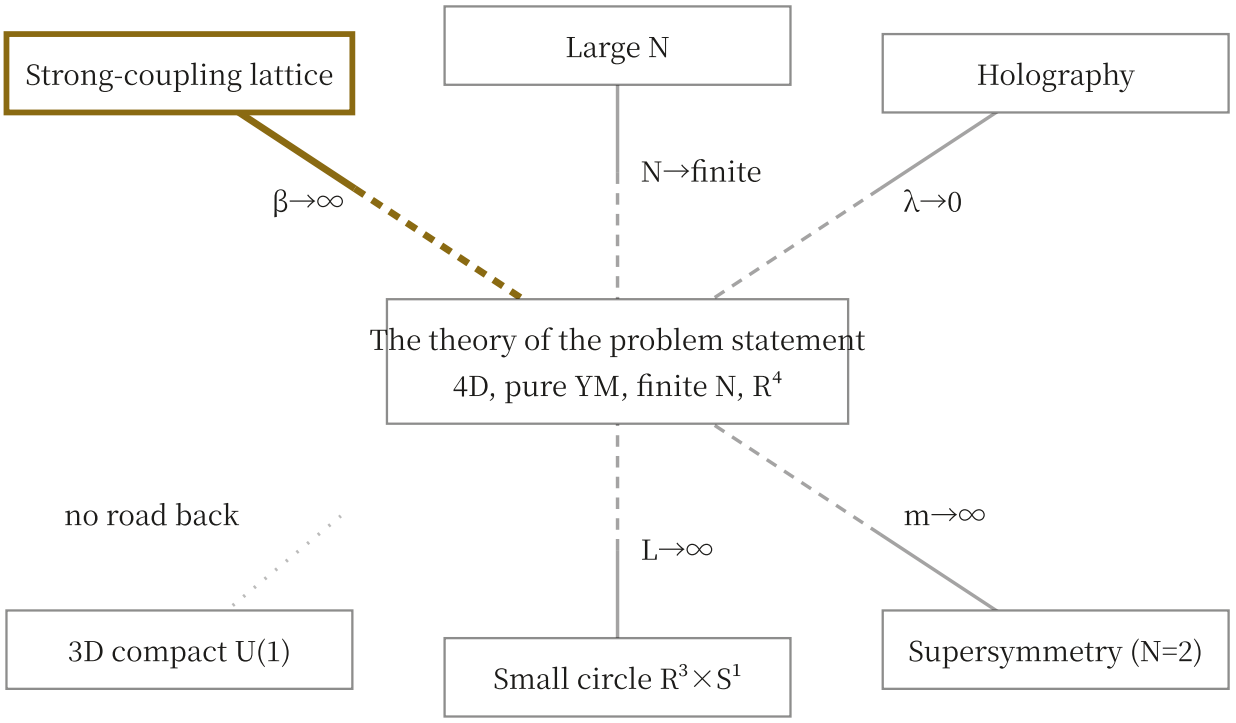
Moving to a neighbouring theory and coming back with a knob

In the move of going to a computable **neighbouring theory** and turning a knob to come back, the assumption is to put that nothing happens on the way. Translated into mathematics this is an exchange of limits, and uniformity is needed (17-10 § 06). For large N, for a dimensionless ratio F,

$$\lim_{a\to 0}\lim_{N\to\infty} F(N,a) \overset{?}{=} \lim_{N\to\infty}\lim_{a\to 0} F(N,a) \quad (9)$$

and what the problem statement needs is the limit at fixed N.

Figure 2 Choosing a shore (button, box, or row of Table 3) highlights the road back. Solid lines are the parts the calculation controls, dashed lines the roads back, and dotted lines shores with no road back



Strong-coupling lattice: The infinite-volume limit and the area law at small β (Osterwalder–Seiler). U(1) included Known , knob to come back: β→∞, needed on the way: The region of convergence does not reach weak coupling

Figure 2 is a picture of the content of Table 3, and has no numbers.

Table 3 Neighbouring theories and the knobs to come back

Shore	What can be said as mathematics	Knob to come back	What is needed on the way
Strong-coupling lattice	The infinite-volume limit and the area law at small β (Osterwalder–Seiler). U(1) included Known	β→∞	The region of convergence does not reach weak coupling
Large N	Sums over strings for strongly coupled SO(N) (Chatterjee). A representation as sums over surfaces at all β (Cao–Park–Sheffield) Known	N to finite	Uniformity of the 1/N corrections
Holography	The eigenvalue problem after fixing the background Known . The correspondence is a conjecture (17-08)	λ→0	Beyond supergravity; the definition of string theory
Supersymmetry (N=2)	Instanton integrals give the prepotential (17-11) Known	Mass→∞	That the gap is not interrupted
Small circle $R^3\times S^1$	Semiclassics at small N and small L (Ünsal–Yaffe) Physics	L→∞	That there is no phase transition
3D compact U(1)	The area law at all couplings (Göpfert–Mack, Villain action) Known	None	The scale does not become one (the abstract of the same paper)

For large N, 't Hooft 1983 is said to have constructed planar field theory under the condition "all particle masses non-zero and the coupling small enough" (bibliographic data and search summaries only), and the gap is an input. In § 6.6 the problem statement writes of 1/N as 'perhaps a good starting point for a rigorous proof (for sufficiently large N)' Known . Example: the lattice ratio is 3.405(21) for $SU(3)$ (Athenodorou–Teper 2020, 2021) and 3.072(14) for $SU(\infty)$ (same, 2021) Physics . Physics reads this as "large N is close", but what mathematics needs is a proof of the uniformity in (9).

08

Looking after choosing a gauge

Many mechanisms of confinement are told in terms of variables defined after choosing a gauge (monopoles, centre vortices, the Landau-gauge propagator) (17-02) Physics . On the side of mathematics, there is in general no globally continuous gauge fixing (Singer, 3-04), and even local conditions leave several solutions (Gribov, 4-08) Known . On the lattice, the expectation of a local quantity equals the expectation of that quantity averaged over gauge transformations (its gauge-invariant part), and components that change under gauge transformations have expectation 0 (Elitzur; for example $\langle U_\ell \rangle = 0$) Known .

The target of the problem statement is gauge-invariant local fields. For example, for $O = \operatorname{tr} F_{\mu\nu}F_{\mu\nu}$ (summed over μ, ν),

$$\left| \langle O(x)\, O(0) \rangle - \langle O \rangle^2 \right| \leq C\, e^{-\Delta |x|} \quad (10)$$

is one way the gap Δ shows itself (4-11). For the propagator, $D(0)$ is reported to look finite and non-zero on large lattices ('seems to') (Cucchieri–Mendes), and Zwanziger argues that copies inside the region do not affect expectations Physics . There is no theorem deriving (10) from $D(p)$.

There are gauge-invariant bridges too. The **centre** of a group G is the set of elements that commute with every element; for $SU(N)$ it is the identity matrix times an N-th root of 1, $e^{2\pi i k/N} I$. **Centre symmetry** is the symmetry under the transformation that multiplies all links of one layer by the same element of the centre (which does not change the values of the plaquettes), and we say it is "unbroken" when no measure changes under this transformation (precisely, on a slab of finite thickness, for every boundary condition; 17-02). Tomboulis–Yaffe 1985 are said to have bounded the Wilson loop from above by the free energy of centre vortices for $SU(2)$ (search summaries only). Chatterjee 2021 proved two things on the lattice for Wilson loops in representations acting non-trivially on the centre (representations in which some

element of the centre maps to something other than the identity matrix; for example the fundamental representation of $SU(N)$) Known : that if centre symmetry is unbroken there is confinement (in the sense $V(R) \to \infty$, weaker than the area law), and that if exponential decay of correlations for any boundary condition is assumed, centre symmetry is unbroken and the area law follows. What is missing is that this assumption holds at weak coupling in four dimensions (the claim of Tomboulis 2007 has been pointed out to have 'missing links'; Ito–Seiler).

09

What physics regards as one

Physics regards the scale after dimensional transmutation as one, and treats separate quantities as faces of the one scale (17-01 § 03). Translated into mathematics, the sentence splits.

Table 4 Sentences of physics and the translated statements

Sentence of physics	Mathematical statement and its current state
The mass and the string tension are constant multiples of the same Λ	That m^2/σ is bounded above and below uniformly in β (the diffusive-relaxation-time assumption of 17-12). Examples where it breaks: three-dimensional compact $U(1)$, the supergravity regime (17-09) Physics
Confinement and the gap are one phenomenon	Separate claims. The problem statement asks only for $\Delta > 0$; confinement is an extension problem ('natural extensions' in § 5 of the problem statement). There is no theorem deriving confinement from a gap in the sense of the problem statement (17-02 § 02). On the lattice, from the strong form, exponential decay for any boundary condition, the area law follows for groups with a non-trivial centre (Chatterjee 2021 Thm 2.4) Known . For finite groups, exponential decay of correlations of gauge-invariant quantities has been proved at weak coupling (Adhikari–Cao; a form on a box with no condition imposed at the boundary, different from the strong form) Known . That they do not confine is at the stage of belief (Chatterjee 2021 writes 'believed'; 5-10) Physics
The gap is the mass of the lightest glueball	The gap and an isolated one-particle state above it are different things. The latter is an extension problem (4-11) Known

10

A list of the places where they part

In every row, what is missing is not the line of argument of physics but the corresponding theorem.

Table 5 The seven moves, and where the translated statements stand

Move (section)	Translated into a mathematical statement	Current state

Move (section)	Translated into a mathematical statement	Current state
Placing it as existing (§ 02)	Construction of a limit satisfying the axioms	The construction published on 2026-11-10 (the committee's assessment is that it is proved in Lean; the infinite-volume limit is along a subsequence, with no uniqueness and no mass gap; 6-01)
Taking σ from the string picture (§ 03, § 04)	The area law at large β	'remains elusive' Known
Extrapolating (§ 05)	(7) and the continuum limit of Wilson loops	At the stage of being posed as problems (the lattice correlation-length problem, the continuum string-tension problem) Known
Expanding in a small parameter (§ 06)	An estimate non-perturbative in g	All Taylor coefficients of $e^{-1/x}$ are 0 Known . $\Delta = c\Lambda$ is an expectation of physics Physics
Coming back from a neighbouring theory (§ 07)	The uniformity in (9)	Within the range we searched, no justification found
Choosing a gauge (§ 08)	Gauge-invariant correlations (10)	Up to conditional lattice theorems Known
Regarding as one (§ 09)	Separate statements and a uniform comparison	At the stage of being cut out as an assumption Physics

Most of the places where they part lie not in divergences but in **limits and uniformity** (7-11), and for mathematics the numbers of physics are not steps of a proof but **targets and checks**. Where the translation succeeded, the result is **conditional theorems** on the lattice. The overall picture set beside the side of mathematics is in 16-05, and the parallel list of names is the parallel table in 16-05.

What this chapter can say

What is said	What is not said
Computed From Table 1, that the continuum value moves by 5.6% with or without the Lüscher term, and that the difference between (4) and (5) is 0.3%. That the central value for (4) with all points agrees with 3.844.	Whether the string picture is right. Confinement itself. The breakdown of the primary source's error 0.061 (0.043 in the fit on this machine).
Known For the seven moves, the translated statements and the current state as stated by the primary sources.	That the seven are exhaustive (the division is this chapter's arrangement). That the hypotheses of the conditional theorems hold.

What is said	What is not said
Physics The readings of semiclassics, large N, holography and supersymmetry in § 06 and § 07.	The existence of the limits taken by turning the knobs. The mass gap and the construction of the continuum limit (this book does not show them).

Sources and reproduction

Item	Kind	Source / tools
(1)(4)(6), the inputs of Table 1	Physics	Lucini–Teper, JHEP 0106 (2001) 050, hep-lat/0103027. Text (Table 1, Table 5, § 4) checked.
(5), the reports in § 03 and § 04	Physics	Athenodorou–Bringoltz–Teper, JHEP 1102 (2011) 030, arXiv:1007.4720. Text (equation (23), abstract) checked.
The problem statement in § 02, § 07 and Table 4	Known	Jaffe–Witten, "Quantum Yang–Mills Theory" (the Clay problem description) § 5 ('uniform gap', 'natural extensions'), § 6.5 (footnote 2), § 6.6 (1/N). Text checked.
(7), the area law, uniqueness	Known	Chatterjee, arXiv:1803.01950 § 4, § 5 (Problems 5.1 and 5.2). Text checked.
The transfer matrix of the detour, strong coupling in Table 3	Known	Lüscher, CMP 54 (1977) 283 (construction of a self-adjoint positive transfer matrix); Osterwalder–Seiler, Ann. Phys. 110 (1978) 440 (both bibliographic data only). The theorems of Perron–Frobenius and Jentzsch: O. Perron, Math. Ann. 64 (1907) 248–263 / R. Jentzsch, J. reine angew. Math. 141 (1912) 235–244 (bibliographic data only). The generalisation: H. H. Schaefer, *Banach Lattices and Positive Operators*, Springer (1974) p. 337 (bibliographic data only; the claim checked in secondary sources; details in the table of sources in 5-03).
Centre symmetry and finite groups in § 08 and Table 4	Known	Chatterjee, CMP 385 (2021) 1007–1039, arXiv:2006.16229. Text (Thms 2.2 and 2.4, the 'believed' paragraph on finite groups) checked. Adhikari–Cao, Ann. Probab. 53 (2025) 140–174, arXiv:2202.10375. Text (Theorem 1.1) checked.
Semiclassics in § 06	Known / Physics	Göpfert–Mack, CMP 82 (1982) 545–606 (Villain action; abstract (INSPIRE); main text not read; the year is Crossref's year of publication, INSPIRE gives 1981). Polyakov, NPB 120 (1977) 429 (bibliographic data only). Ünsal–Yaffe, PRD 78 (2008) 065035, arXiv:0803.0344 (abstract).
Large N in § 07	Known / Physics	't Hooft, CMP 88 (1983) 1 (bibliographic data and search summaries only). Chatterjee, arXiv:1502.07719; Cao–Park–Sheffield, arXiv:2307.06790 (abstracts). Athenodorou–Teper, arXiv:2007.06422 ($SU(3)$; Tables 14 and 17 of the text checked), arXiv:2106.00364 (Tables 24 and 34 [$SU(3)$] and Tables 31 and 38 [$SU(\infty)$] of the text checked).

Item	Kind	Source / tools
Gauges in § 08	Known / Physics	Singer, CMP 60 (1978) 7; Gribov, NPB 139 (1978) 1; Elitzur, PRD 12 (1975) 3978 (all bibliographic data only); Tomboulis–Yaffe, CMP 100 (1985) 313 (bibliographic data and search summaries only). Cucchieri–Mendes, arXiv:0710.0412; Zwanziger, hep-ph/0303028; Tomboulis, arXiv:0707.2179; Ito–Seiler, arXiv:0711.4930 (abstracts).
Table 1, Figure 1, (3), the numbers in § 03– § 06	Computed	Python 3 on this machine (standard library): string_sigma.py, extra.py. The inputs are only the values in the tables of the primary sources. (3) and Perron–Frobenius are standard facts.

Next chapter to read: 18-01 How to read the dictionary (the entrance to Part 18, which sets the words of physics and of mathematics side by side in two columns of standing, definition and reality).