Volume 1 Preparation

Part 1 Mathematical preparation

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

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

Chapter		Page
1-01	Vector spaces and linear maps	2
1-02	What a group is	13
1-03	Lie groups and Lie algebras	26
1-04	SU(2) and SU(3)	39
1-05	Manifolds and tangent spaces	52
1-06	Differential forms and the exterior derivative	63
1-07	Vector bundles and connections	76
1-08	Probability and measure	89
1-09	Probability in infinite dimensions	101
1-10	Hilbert spaces and self-adjoint operators	114
1-11	Fourier analysis and distributions	126
1-12	Detour: what happens only in four dimensions	139

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

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

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

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

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

Order of this chapter

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

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

01

Why start from linear algebra

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

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

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

02

Vector spaces

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

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

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

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

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

03 Linear maps and matrices

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

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

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

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

04 Eigenvalues and eigenvectors

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

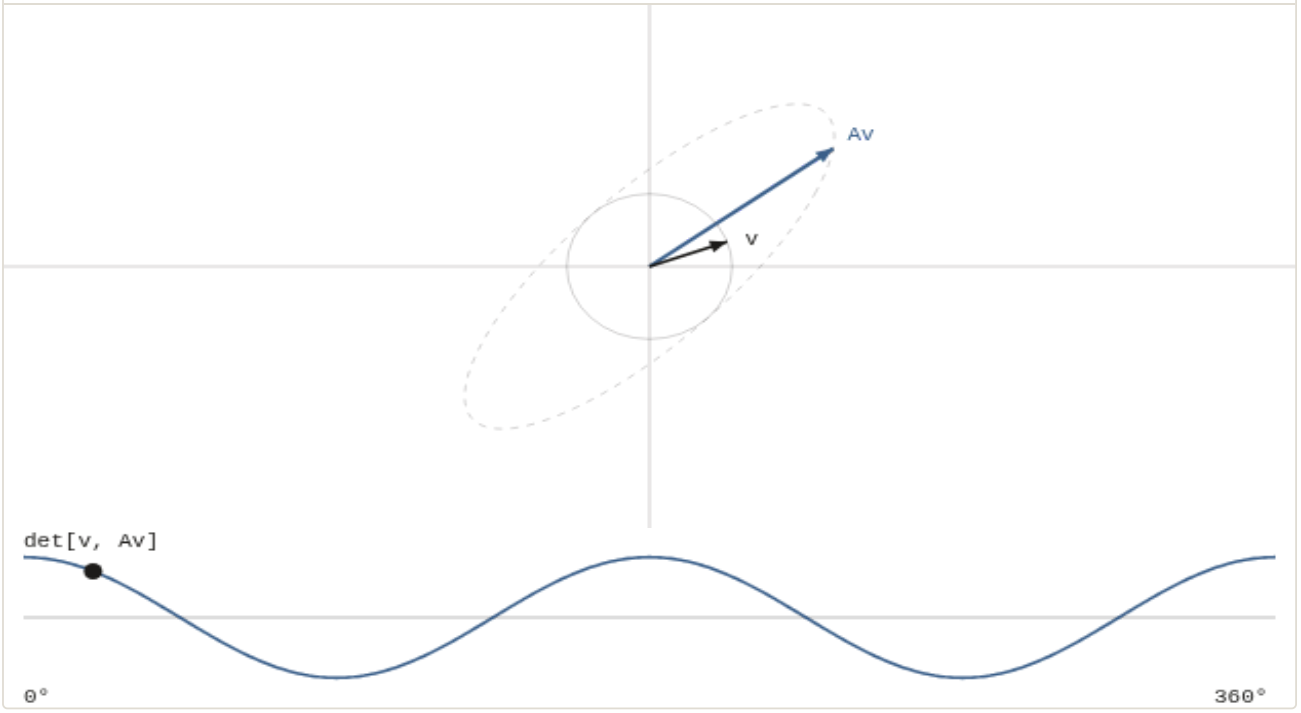
$$Av = \lambda v \ \ (v \neq 0) \qquad \Longleftrightarrow \qquad \det(A - \lambda I) = 0$$

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

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

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

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



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

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

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

Detour: rotations have no real eigenvectors

A subsection that may be skipped.

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

05

Inner products and adjoints

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

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

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

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

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

06

Hermitian matrices

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

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

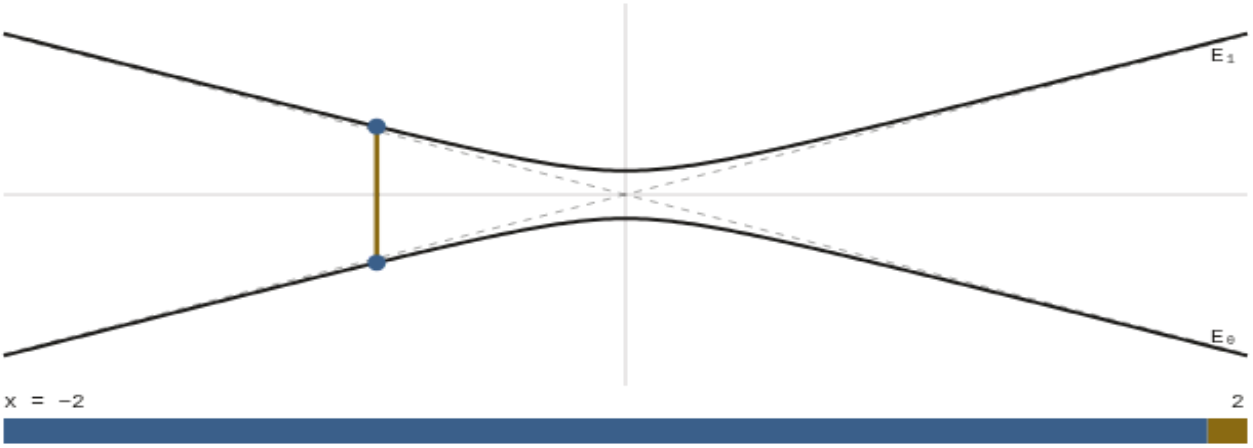
Pushing this further gives the **spectral theorem** Known . A Hermitian matrix has an orthonormal basis of eigenvectors, and the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$ are real. With

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

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

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

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



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

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

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

Detour: levels avoid each other

A subsection that may be skipped.

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

07

Unitary matrices

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

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

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

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

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

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

08

The matrix exponential

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

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

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

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

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

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

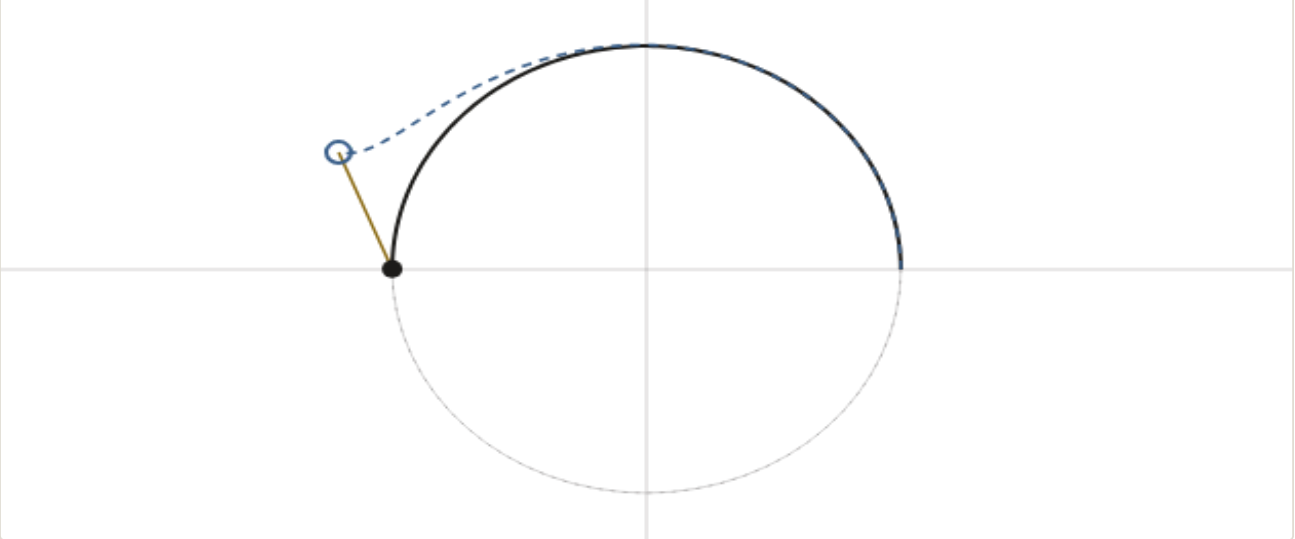
$$e^A e^B = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \qquad e^B e^A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix},$$

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

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

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

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



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

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

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

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

A subsection that may be skipped.

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

```
09
```

Joining Hermitian and unitary

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

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

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

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

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

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

What this chapter can say

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

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

Sources and reproduction

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

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

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

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

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

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

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

Order of this chapter

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

01

From symmetry to groups

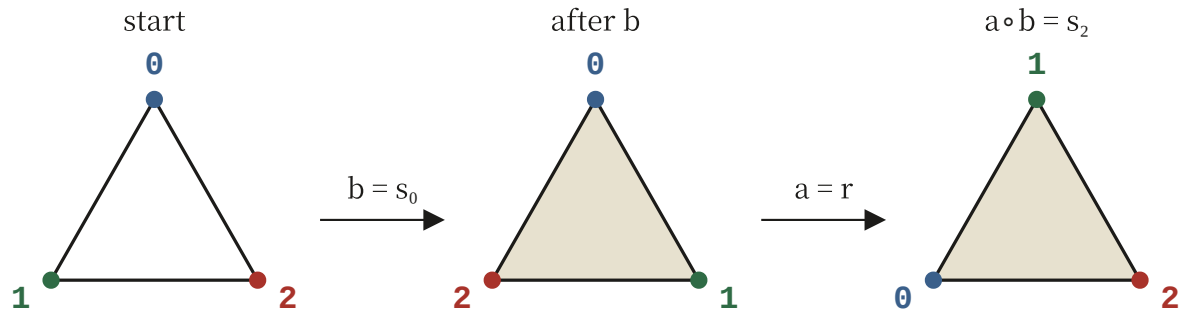
There are six ways to put an equilateral triangular plate back **so that it fits exactly over its original place**: three rotations by 0°, 120° and 240°, and three flips about an axis through a vertex. **Doing two in succession gives one of the six again**, and there is an operation that does nothing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

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

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

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

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



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

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

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

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

02

Definition of a group, and examples

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

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

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

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

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

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

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

03

Homomorphisms, normal subgroups, quotients

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

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

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

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

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

04

The rotation group SO(3)

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

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

Every rotation has an axis

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

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

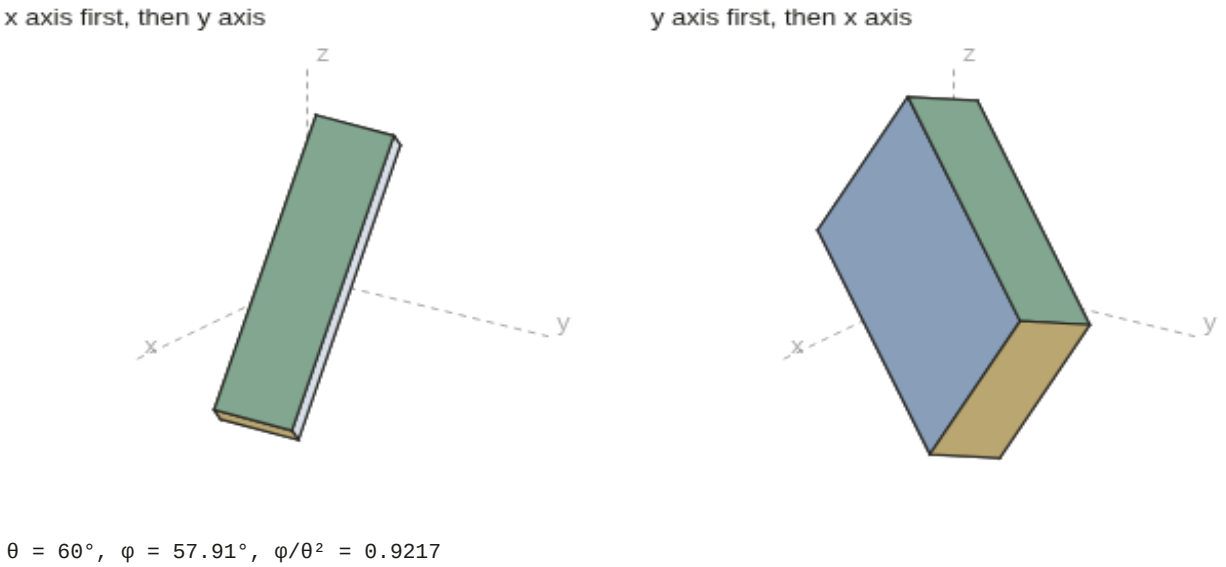
$$R(\theta, n)\, x = \cos\theta\, x + \sin\theta\, (n \times x) + (1 - \cos\theta)(n \cdot x)\, n$$

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

The mismatch from the order of rotations

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

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



θ = 60°, φ = 57.91°, φ/θ² = 0.9217

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

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

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

05

SU(2)

Next is a group of complex 2×2 matrices.

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

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

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

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

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

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

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

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

06

The double cover SU(2) → SO(3)

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

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

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

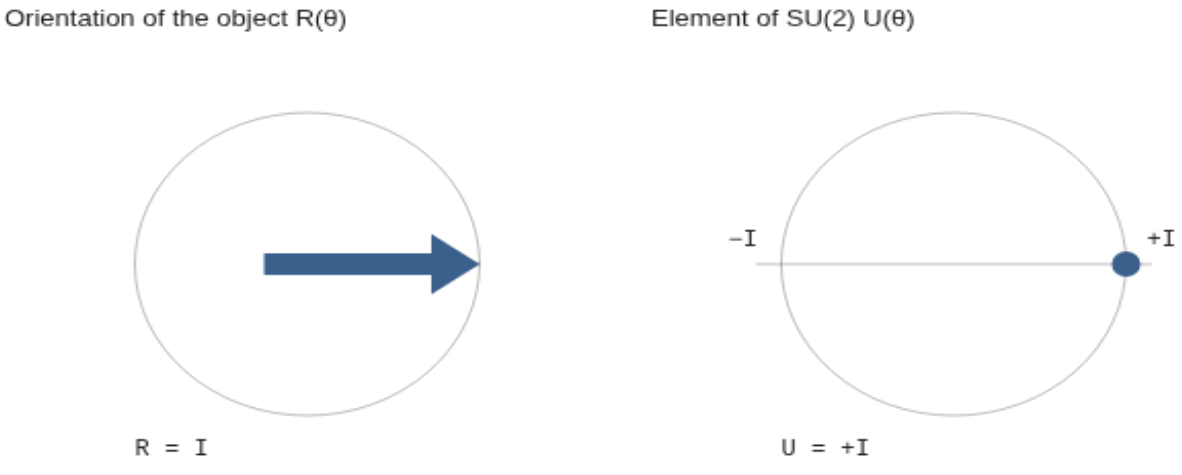
1. **It preserves length**: $|x'| = |x|$ follows from $\det(UXU^\dagger) = \det X$. SU(2) is connected and ρ is continuous, so the determinant cannot move from +1, and $\rho(U) \in \mathrm{SO}(3)$.
2. **It is a homomorphism**: $\rho(UV) = \rho(U)\rho(V)$ follows from $(UV)X(UV)^\dagger = U(VXV^\dagger)U^\dagger$.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

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

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

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

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



θ = 0°, tr R = 3.000, tr U = 2.000

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

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

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

07

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

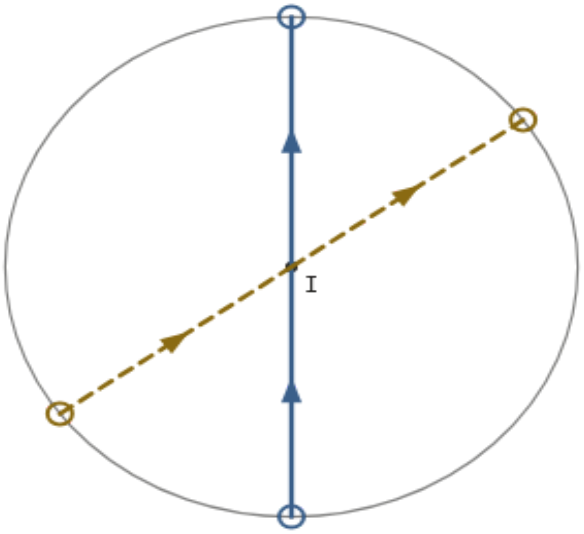
A section that may be skipped. It is not a prerequisite for later chapters. It looks at the double cover as a "shape".

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

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

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

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



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

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

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

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

08

A first look at representations

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

Representations of U(1) are labelled by integers

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

Irreducible and completely reducible

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

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

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

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

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

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

09

The group inside Yang–Mills

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

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

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

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

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

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

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

What this chapter can say

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

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

Sources and reproduction

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

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

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

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

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

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

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

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

Order of this chapter

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

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

---

01

Looking at infinitesimal elements

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

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

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

02

Matrix Lie groups and their Lie algebras

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

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

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

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

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

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

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

03

How far the exponential map reaches

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

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

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

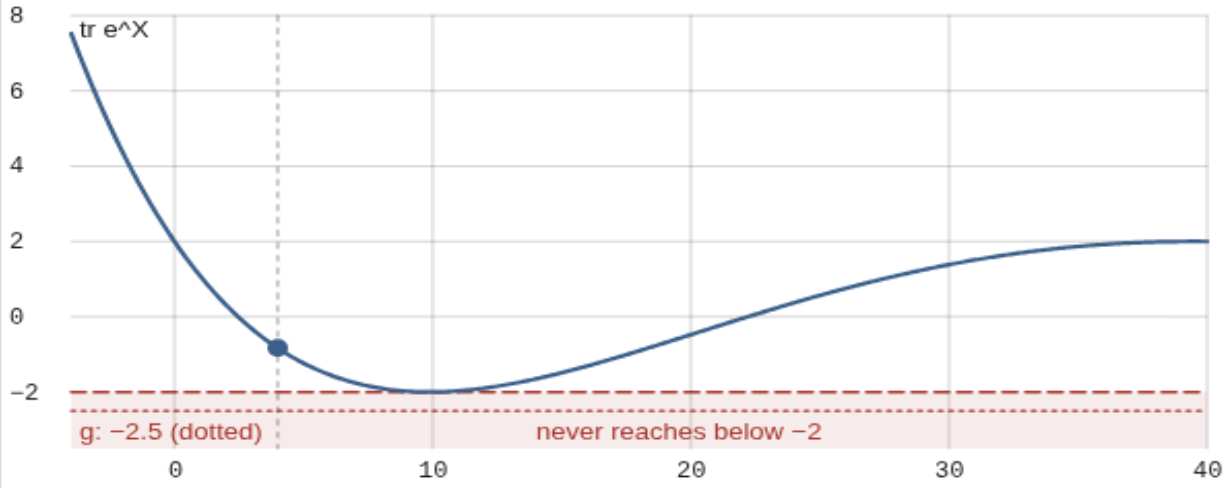
A subsection that may be skipped. It is not a prerequisite for later chapters. It is an example of a non-compact group.

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

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

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

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



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

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

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

04

The commutator is the second-order term of the mismatch

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

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

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

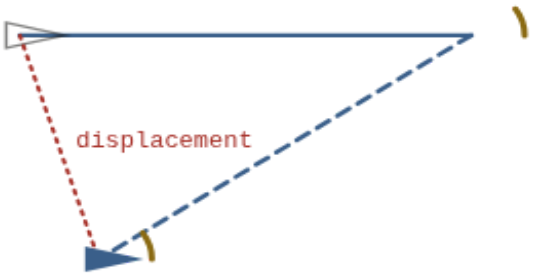
Example: parallel parking. Write the group SE(2) of "points with an orientation" in the plane (position (x, y), orientation α) as 3×3 matrices.

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

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

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

forward → turn → back → turn back



solid: forward, dashed arm back turning on the spot

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

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

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

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

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

05

The structure called a Lie algebra

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

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

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

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

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

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

06

Structure constants

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

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

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

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

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

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

07

The Baker–Campbell–Hausdorff formula

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

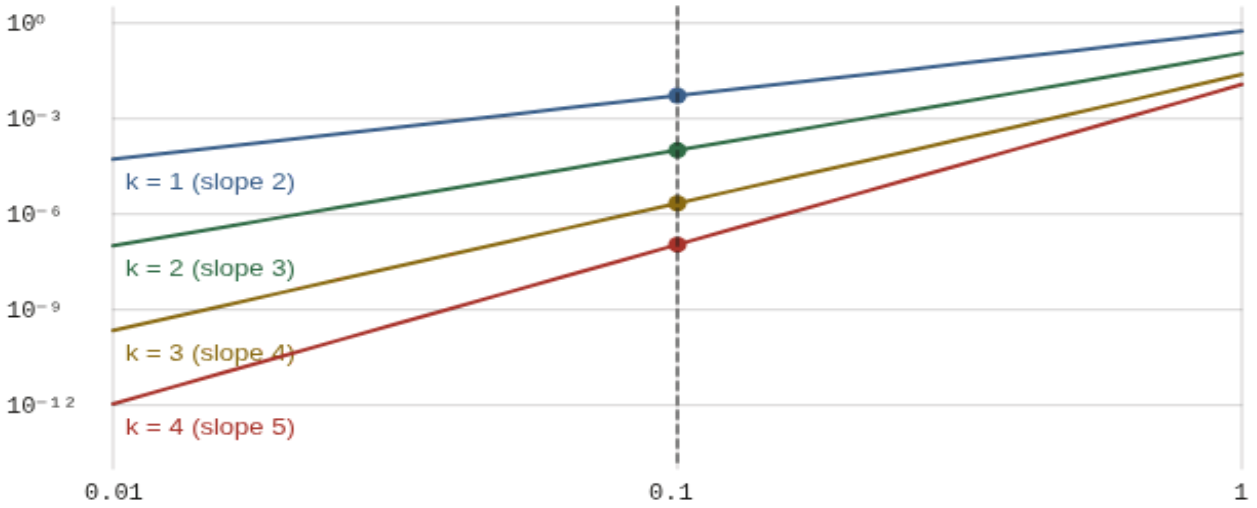
$$Z = X + Y + \tfrac{1}{2}[X,Y] + \tfrac{1}{12}\big([X,[X,Y]] + [Y,[Y,X]]\big) - \tfrac{1}{24}[Y,[X,[X,Y]]] + \cdots \quad (2)$$

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

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

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

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



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

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

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

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

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

```
08
```

The Killing form

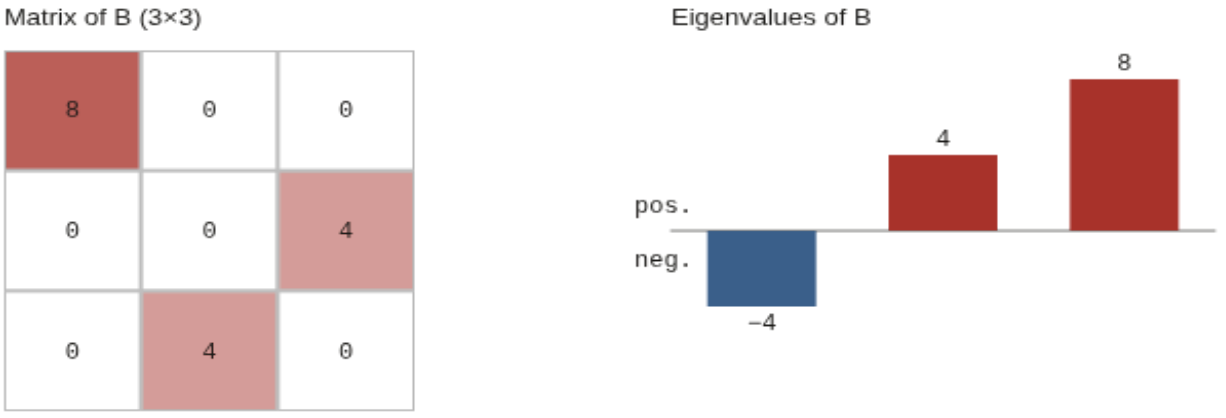
The Yang–Mills action needs a "squared length" on the Lie algebra that does not change under Ad_g (§ 09). The candidate that can be made from the Lie algebra alone is the **Killing form**.

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

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

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

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



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

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

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

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

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

09

The Lie algebra inside Yang–Mills

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

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

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

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

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

What this chapter can say

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

Sources and reproduction

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

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

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

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

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

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

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

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

Order of this chapter

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

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

---

01

Why SU(2) and SU(3)

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

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

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

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

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

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

02

The Pauli matrices and su(2)

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

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

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

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

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

03

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

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

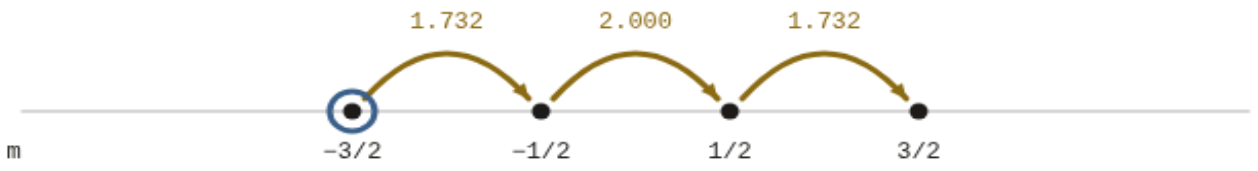
$$J_\pm = J_1 \pm i J_2, \qquad [J_3, J_\pm] = \pm J_\pm, \qquad [J_+, J_-] = 2J_3 \quad (4)$$

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

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

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

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



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

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

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

04

The Gell-Mann matrices and su(3)

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

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

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

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

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

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

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

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

05

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

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

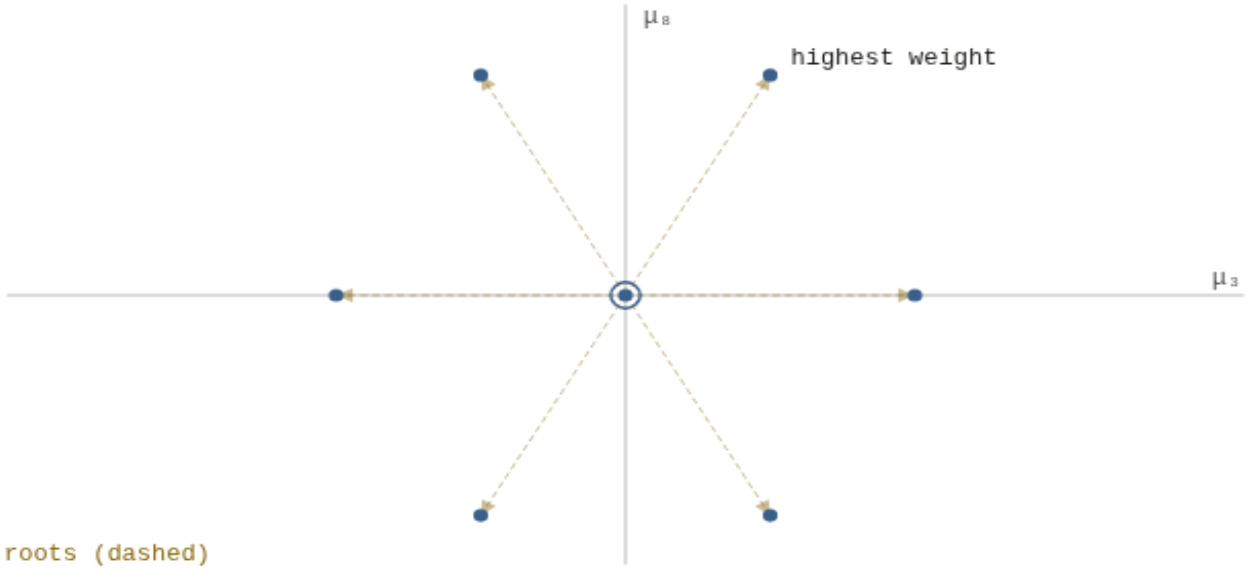
$$\mu(e_1) = \left(\tfrac{1}{2},\ \tfrac{1}{2\sqrt{3}}\right), \qquad \mu(e_2) = \left(-\tfrac{1}{2},\ \tfrac{1}{2\sqrt{3}}\right), \qquad \mu(e_3) = \left(0,\ -\tfrac{1}{\sqrt{3}}\right) \quad (6)$$

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

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

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

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



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

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

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

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

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

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

A subsection that may be skipped.

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

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

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

06

The adjoint representation: the gauge field changes in this representation

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

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

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

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

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

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

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

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

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

> 07

The centre and N-ality

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

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

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

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

08

Root systems: only four possible angles

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

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

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

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

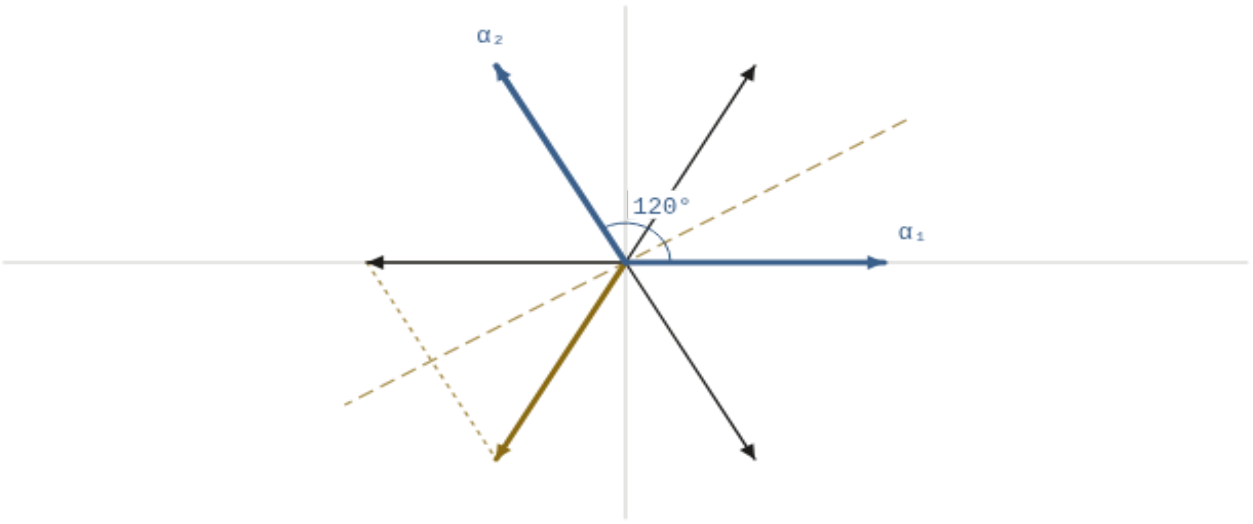
There are many roots, so we choose representatives. Draw a line through the origin containing no root (a hyperplane, for rank 3 or more), and call the roots on one side the **positive roots**. The positive roots that cannot be written as a sum of two positive roots are the **simple roots**; their number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

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

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

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

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



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

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

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

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

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

09

The list of compact simple Lie groups

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

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

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

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

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

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

What this chapter can say

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

Sources and reproduction

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

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

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

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

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

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

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

Order of this chapter

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

01

Why manifolds are needed

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

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

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

02

Coordinate neighbourhoods and atlases

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

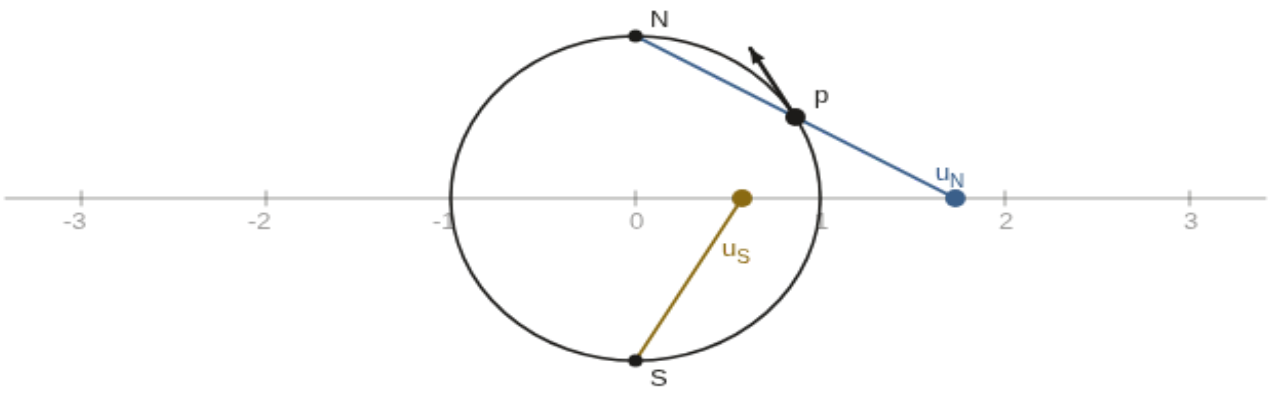
$$\varphi_\alpha \circ \varphi_\beta^{-1} : \ \varphi_\beta(U_\alpha \cap U_\beta) \longrightarrow \varphi_\alpha(U_\alpha \cap U_\beta) \quad \text{is } C^\infty \ (1)$$

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

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

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

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



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

θ (degrees)	u_N	u_S	v_N	v_S	v_S/v_N	$-1/u_N^2$
30	1.7321	0.5774	2.0000	−0.6667	−0.3333	−0.3333
89	114.5887	0.0087	6565.7794	−0.5000	−0.0001	−0.0001
−135 (= 225)	−0.4142	−2.4142	0.5858	−3.4142	−5.8284	−5.8284

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

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

03

Definition and examples of manifolds

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

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

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

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

04

Cutting out by equations — the regular value theorem

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

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

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

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

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

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

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

05

Smooth functions and maps

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

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

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

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

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

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

06

Tangent vectors — three views

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

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

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

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

07

The transformation law of components and the differential of a map

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

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

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

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

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

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

08

Vector fields and flows

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

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

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

Detour: the hairy ball theorem

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

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

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

09

The bracket — can flows be interchanged?

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

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

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

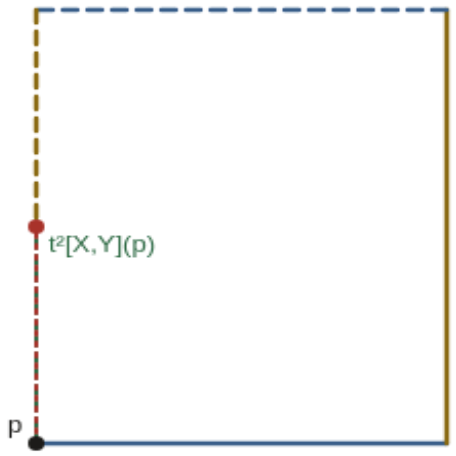
For coordinate basis fields, $[\partial_i, \partial_j] = 0$. Conversely, if any two of n fields forming a basis commute, then locally there is a chart whose coordinate basis they are (a theorem on commuting frames, a form of the Frobenius theorem) Known .

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

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

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

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



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

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

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

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

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

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

10

Manifolds inside Yang–Mills

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

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

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

What this chapter can say

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

Sources and reproduction

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

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

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

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

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

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

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

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

Order of this chapter

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

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

---

01

Why differential forms

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

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

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

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

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

02

1-forms

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

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

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

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

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

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

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

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

03

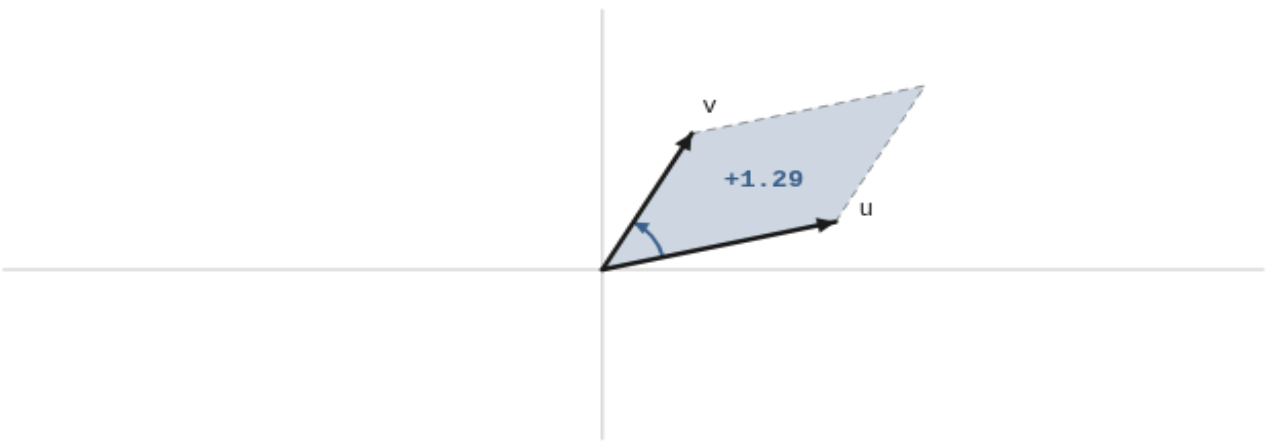
k-forms and the wedge product

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

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

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

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




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

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

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

04

The exterior derivative

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

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

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

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

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

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

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

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

05

Integration and orientation

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

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

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

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

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

$$\int_{\text{upper hemisphere}} z\,dx \wedge dy = \int_0^{2\pi}\int_0^{\pi/2} \cos^2\vartheta\,\sin\vartheta\,d\vartheta\,d\varphi = \frac{2\pi}{3}$$

(numerical integration and a check of the orientation Computed). The value is the same as the volume of the solid hemisphere. This is not a coincidence but an instance of

Stokes' theorem in the next section, § 06: $d(z\,dx \wedge dy) = dx \wedge dy \wedge dz$, and on the base disc, part of the boundary of the solid hemisphere, $z = 0$.

06

Stokes' theorem

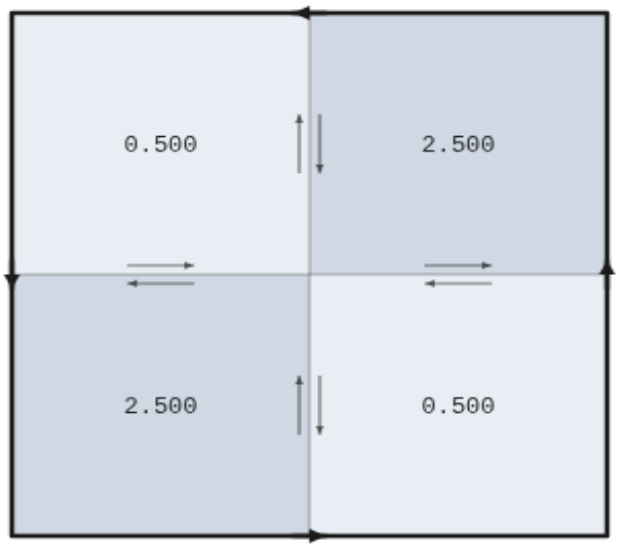
"The total of the derivative inside is the total of the values on the boundary." It is the form of (3) gathered for all dimensions. If M is a compact oriented n-dimensional manifold (it may have a boundary ∂M) and ω an (n−1)-form, then

$$\int_M d\omega = \int_{\partial M} \omega \quad (9)$$

The orientation of ∂M is taken so that putting the outward normal first gives the orientation of M. For an interval it is (3); for a planar region it is Green's theorem via (7); for a surface in $\mathbb{R}^3$ it is Stokes' theorem for the curl; for a solid it is Gauss's divergence theorem Known .

Sketch of the proof. On a rectangle $[a, b] \times [c, d]$, if $\omega = Q\,dy$ then $\iint \partial_x Q\,dx\,dy = \int_c^d (Q(b, y) - Q(a, y))\,dy$, which is the sum of the line integrals along the right edge (upwards) and the left edge (downwards). A general region is divided into small squares. Adjacent squares traverse their common edge in opposite directions, so the inner edges cancel and only the outer boundary remains. Meanwhile, the circulation around a small square approaches "$d\omega$ at the centre × area".

Figure 2 However the number of divisions N is changed, the total of the circulations of the small squares (numbers and shading) stays at the value 6 of the outer boundary. "Circulation ÷ area" approaches $d\omega$ at the centre. For N ≤ 4, the opposite arrows traversed by the two neighbours are drawn on the inner edges.



N = 2, Σ circulation = 6.000000, ∮ on the outer boundary = 6.000000, max|circulation/area − dω| = 0.250000 = 1/4

$\omega = (-\frac{y}{4} - xy^2 - y^3)\,dx + (\frac{x}{4} + x^2 y)\,dy$, $d\omega = (\frac{1}{2} + 4xy + 3y^2)\,dx \wedge dy$, region $[-1,1]^2$. Computed The circulations are computed on the spot from closed formulas for the line integrals along the four edges (agreeing with Table 1).

Computed **Table 1** For the ω of Figure 2, the circulation of each small square found as the sum of the line integrals along its four edges (no surface integrals used).

N	Total of circulations	max |circulation/area − dω(centre)|
1	6.000000000000	1.000000
2	6.000000000000	0.250000
8	6.000000000000	0.015625
64	6.000000000000	0.000244

For every N the total agrees with the $\oint \omega = 6$ of the outer boundary (difference at most 3×10^{-15}), and the deviation decreases exactly as $1/N^2$ (on a square of side $h = 2/N$ the average of $3y^2$ is larger than the value at the centre by $h^2/4$). On the unit disc too both sides agree at $5\pi/4$ Computed .

07

Closed and exact

If $d\omega = 0$ the form is called **closed**, and if there is an η with $\omega = d\eta$ it is called **exact**. By (6), exact implies closed. The converse depends on the shape of the space. On the plane with the origin removed:

$$d\theta := \frac{x\,dy - y\,dx}{x^2 + y^2} \quad (10)$$

By (7), $\partial_x \frac{x}{r^2} + \partial_y \frac{y}{r^2} = \frac{2r^2 - 2x^2 - 2y^2}{r^4} = 0$, so it is closed. On the unit circle, as in Example 1, $\oint d\theta = 2\pi$, so by (3) it is not exact. The argument θ increases by 2π in one turn, so it is a function only on a region with a cut (for example the negative x axis). **dθ is defined without a cut.**

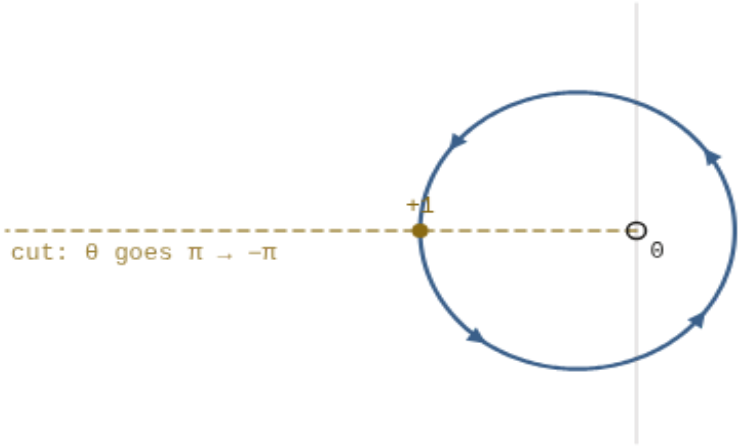
The Poincaré lemma. On a star-shaped region (one containing the segment joining some point to each point) every closed form is exact Known . If the region is star-shaped with respect to the origin, then for a closed 1-form α

$$\eta(x) = \int_0^1 x^i\, \alpha_i(tx)\, dt \quad (11)$$

satisfies $d\eta = \alpha$ (checked with $d(\sin x\, e^y + x^3 y)$; moving the base point to (1, 0), θ is recovered from dθ on the right half-plane Computed).

The winding number. $\frac{1}{2\pi}\oint_\gamma d\theta$ is the number of times the closed curve γ goes round the origin, and does not change under deformations that do not cross the origin (dθ is closed on the band swept by the deformation, so by (9) the difference is 0). θ jumps from π to −π each time it crosses the negative x axis from above to below, so the winding number is also **the oriented number of crossings of the cut**.

Figure 3 Moving the centre c_x of the closed curve makes $\frac{1}{2\pi}\oint d\theta$ jump through the integers 0, 1, 2. The value changes only when the curve crosses the origin, and equals the oriented number of crossings of the cut (dashed line).




```
Circle, centre (-0.375, 0), ∮dθ/2π = 1.000000, crossings of the cut = 1, distance from the origin 0.625
```

Computed The sum of differences of the argument at 720 points is computed on the spot (agreeing with the computation on this machine at 16 positions). The limaçon is the curve $r = \frac{1}{2} + \cos t$ shifted by $(c_x, 0)$.

Computed **Table 2** $\frac{1}{2\pi}\oint d\theta$. Computed in two ways at 2×10^5 points; the difference from an integer is below 10^{-9}.

Closed curve	Winding number
Circle of radius 1, centre (0, 0) / reversed / twice round	1 / −1 / 2
Circle of radius 1, centre (2, 0) (does not enclose the origin)	0
Limaçon $c_x = -0.25 / -1 / 0.5$	2 / 1 / 0

The index of the zeros in the detour of 1-05 (the hairy ball theorem) is also this winding number. Outside the solenoid of 2-03 § 06, $A = \frac{\Phi}{2\pi}\,d\theta$, and although $dA = 0$ the circulation is Φ. The Aharonov–Bohm effect is the physics of a form that is closed but not exact.

08

An entrance to de Rham cohomology

Among closed forms, regard the exact ones as 0, and count what remains.

$$H^k(M) = \frac{\{\omega : d\omega = 0\}}{\{d\eta\}}, \qquad b_k(M) = \dim H^k(M) \quad (12)$$

b_k is called a **Betti number**. **Example.** H^0 consists of functions constant on each connected component, so b_0 is the number of connected components. $H^1(\mathbb{R}^2 \setminus \{0\})$ is the one dimension spanned by [dθ] (subtracting $c\,d\theta$ from a closed form ω so that the circulation on the unit circle is 0 makes it exact. Every closed curve can be deformed, without crossing the origin, into "a curve that goes round the unit circle some number of times", and deformation does not change circulation (the argument of § 07), so every circulation is 0. Then the line integral from a base point does not depend on the path and defines a function [Known]). The area form σ of the sphere is closed and $\int_{S2} \sigma = 4\pi$, but if $\sigma = d\eta$, then since S^2 has no boundary, (9) would make the integral 0. Hence it is not exact.

De Rham's theorem (1931), in today's language: H^k agrees with cohomology (with real coefficients) defined from topology alone, without forms [Known] . Examples: $\mathbb{R}^n$ is (1, 0, 0, ...), $\mathbb{R}^3 \setminus \{0\}$ is (1, 0, 1, 0), S^4 is (1, 0, 0, 0, 1) (the stage of 3-05).

The point magnetic charge. The 2-form $\beta = (x\,dy \wedge dz + y\,dz \wedge dx + z\,dx \wedge dy)/r^3$ on $\mathbb{R}^3 \setminus \{0\}$ is the magnetic field $\hat{\mathbf{r}}/r^2$, and $d\beta = 0$, but its integral over a closed surface enclosing the origin is 4π (difference below 10^{-9} for a sphere and a shifted cube, and 0 for a sphere not enclosing it [Computed]). Hence there is no A with $dA = \beta$ on the whole space. Off the negative z axis $A_N = (1 - \cos\vartheta)\,d\varphi$ can be used, and off the positive z axis $A_S = -(1 + \cos\vartheta)\,d\varphi$ (spherical coordinates); the difference $2\,d\varphi$ is a gauge transformation by the multivalued function 2φ [Computed] . Taking potentials separately on the two regions, north and south, and gluing them on the overlap by a gauge transformation is in § III of Wu–Yang (1975). The language of this gluing is bundles and connections (1-07).

If the total magnetic flux is Φ, the difference is $\frac{\Phi}{2\pi}\,d\varphi$ (here Φ = 4π). The phase factor $e^{iq\chi/\hbar}$ of a particle of charge q (2-03) is single-valued under $\chi = \frac{\Phi}{2\pi}\varphi$ when $q\Phi/h$ is an integer. The quantisation of the point magnetic charge is due to Dirac (1931), and the reading in the language of bundles is in 1-07 § 08.

Detour: differential forms on a lattice

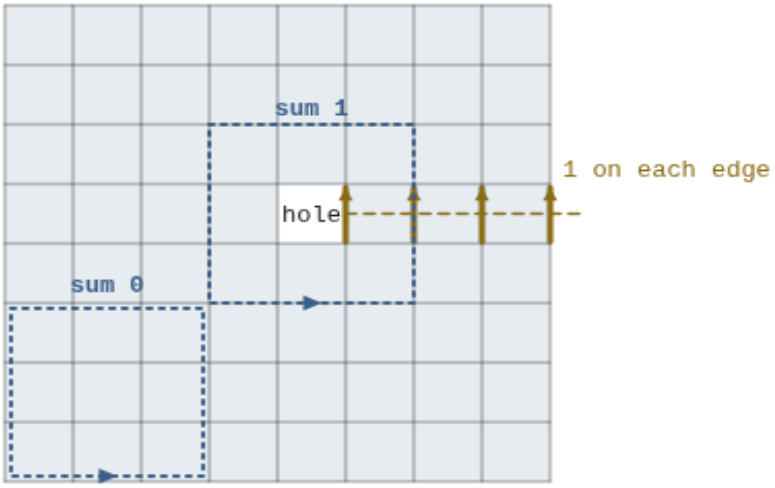
A subsection you may skip.

Values placed on the vertices, the oriented edges and the faces of a square lattice are taken as 0-, 1- and 2-forms; d_0 is the difference between the two ends of an edge, and d_1 the sum around the boundary of a face traversed anticlockwise. Each vertex appears once with + and once with − on the boundary of a face, so $d_1 d_0 = 0$ holds exactly as a matrix of integers (the reason the

gauge function disappeared from the face values on the lattice of 2-03). Table 3 shows the computation of $b_0 = \dim\ker d_0$, $b_1 = \dim\ker d_1 - \operatorname{rank} d_0$, $b_2 = F - \operatorname{rank} d_1$ (F is the number of faces).

Computed **Table 3** Betti numbers on lattices (side L = 8; the same for L = 4) and the known values for the corresponding smooth surfaces. The table sits inside Figure 4; choosing a row changes the figure.

Figure 4 Choosing a row of the table below shows that lattice (side 8), the numbers of vertices, edges and faces V, E, F, and the Betti numbers. For "Remove one face", the 1-form with 1 on the edges crossing the cut (dashed line) is drawn (thick arrows), together with its sums along two closed paths (dotted). It is 1 on a closed path around the hole and 0 on one that does not go around it.



`Remove one face  V = 81  E = 144  F = 63  (b₀, b₁, b₂) = (1, 1, 0)  V − E + F = 0 = b₀ − b₁ + b₂`

Lattice	(b₀, b₁, b₂)	Smooth case
Square	(1, 0, 0)	Disc (1, 0, 0)
Remove one face	(1, 1, 0)	Disc with one hole (1, 1, 0)
Remove two faces	(1, 2, 0)	Disc with two holes (1, 2, 0)
Periodic lattice	(1, 2, 1)	Torus T^2 (1, 2, 1)
Surface of a cube	(1, 0, 1)	Sphere S^2 (1, 0, 1)

Computed The numbers in this figure were computed on this machine (from the ranks of the integer matrices d_0, d_1). The figure only embeds and draws those numbers.

Regardless of fineness, the values of the smooth case come out (a small check of the theorem that the two cohomologies agree Known). In the numbers of Figure 4, $V - E + F = b_0 - b_1 + b_2$ also holds. It holds for every lattice, because the dimension of the kernel plus the rank is the number of columns Known . On the lattice with one hole, putting 1 only on the edges crossing a cut from the hole to the right edge gives a closed form; its sum around the hole is 1, and on a closed path not around the hole 0 Computed . It is the lattice version of dθ/2π. Putting group elements rather than numbers on the edges is Wilson's lattice gauge theory (5-01).

09

The Hodge dual and Maxwell's equations

Up to here we have not used a metric. The Hodge dual needs one, so in this section we define on the spot only as much as we use. A metric is given by the table of inner products η_{ij} between coordinate directions (here diagonal, with entries ± 1). In Minkowski spacetime (2-04), for coordinates (t, x, y, z), $\eta = \mathrm{diag}(-1, 1, 1, 1)$ (signature $(-,+,+,+)$), $c = \varepsilon_0 = \mu_0 = 1$); in Euclidean $\mathbb{R}^4$, $\eta = \mathrm{diag}(1, 1, 1, 1)$.

- **Volume form**: vol $= dt \wedge dx \wedge dy \wedge dz$ (in the Euclidean case $dx^1 \wedge dx^2 \wedge dx^3 \wedge dx^4$).
- **Inner product of forms**: for basis elements $\langle dx^{i1} \wedge \cdots \wedge dx^{ik},\ dx^{i1} \wedge \cdots \wedge dx^{ik} \rangle = \eta_{i1 i1} \cdots \eta_{ik ik}$, and basis elements with different sets of indices give 0. For general forms extend bilinearly in the coefficients.
- **Hodge dual**: the linear map taking a k-form β to an (n−k)-form $*\beta$ such that $\alpha \wedge *\beta = \langle \alpha, \beta \rangle$ vol for every k-form α.

Example. In Minkowski, from $\langle dt \wedge dx, dt \wedge dx \rangle = \eta_{tt}\eta_{xx} = -1$ and $dt \wedge dx \wedge dy \wedge dz =$ vol, $*(dt \wedge dx) = -dy \wedge dz$. From $\langle dy \wedge dz, dy \wedge dz \rangle = +1$ and $dy \wedge dz \wedge dt \wedge dx =$ vol, $*(dy \wedge dz) = dt \wedge dx$ (checked for all 6 basis 2-forms Computed). A −1 enters only for basis elements containing time, and this affects the sign in (14) and the sign of $**$ below.

We use it for the electromagnetic field. Gathering the potentials into a 1-form gives

$$A = -\phi\, dt + A_x\, dx + A_y\, dy + A_z\, dz, \qquad F = dA = -E_i\, dt \wedge dx^i + B_x\, dy \wedge dz + B_y\, dz \wedge dx$$

($\mathbf{E} = -\nabla\phi - \partial_t \mathbf{A}$, $\mathbf{B} = \nabla \times \mathbf{A}$). The four Maxwell equations become two.

$$dF = 0, \qquad d{*}F = *j, \qquad j = -\rho\, dt + J_x\, dx + J_y\, dy + J_z\, dz \quad (14)$$

$dF = 0$ is automatic from (6) and is the two source-free Maxwell equations of 2-03; $d{*}F = *j$ is the two Maxwell equations with sources (checked with random potentials Computed). That F does not change under the gauge transformation $A \mapsto A + d\chi$ is also (6) (the formula for gauge transformations in 2-03). What enters the action is $F \wedge *F = (|\mathbf{B}|^2 - |\mathbf{E}|^2)$ vol, and also $F \wedge F = -2(\mathbf{E} \cdot \mathbf{B})$ vol Computed .

What happens only in four dimensions

$*F$ is an (n−2)-form, and has the same degree as F only for n = 4. The $**$ on 2-forms is +1 in Euclidean $\mathbb{R}^4$ (the eigenvalues ± 1 of $*$ three each) and −1 in Minkowski (eigenvalues $\pm i$) (agreeing with the formula $(-1)^{k(n-k)+s}$, where s is the number of negative signs, for every k Computed). In the Euclidean case it splits as $F = F^+ + F^-$ (self-dual and anti-self-dual), and $|F|^2 = |F^+|^2 + |F^-|^2 \geq \big||F^+|^2 - |F^-|^2\big| = |\langle F, *F \rangle|$.

Integrating gives $\int F \wedge *F \geq |\int F \wedge F|$, with equality when $F = \pm * F$; this is the skeleton of the instanton inequality (3-05).

In Minkowski, $** = -1$ on 2-forms, so if $*F = \lambda F$ then $-F = \lambda^2 F$, and the only real F satisfying $*F = \pm F$ is 0. In physics, instantons are considered in imaginary time (2-04, 4-02). For U(1), $F \wedge F = d(A \wedge F)$ is exact, so its integral is fixed by the boundary term (at infinity) alone. In the non-abelian case this becomes an integer-valued topological quantity (3-05).

See also the article on four dimensions and 1-12. In the non-abelian case $F = dA + A \wedge A$ (A is the geometric way of writing with values in anti-Hermitian matrices; the translation to the physicists' way is in 3-02), and in place of $dF = 0$ we have $dF + A \wedge F - F \wedge A = 0$ (Bianchi) (1-07, 3-03).

What this chapter can say

Item	What is said	What is not said
Exterior derivative, Stokes	(4)–(8), the reason for $d^2 = 0$, the example of the hemisphere, Table 1 Computed	That pullback commutes with d, the proof of the general Stokes theorem Known
Closed and exact, de Rham	Why dθ, the area form and the point magnetic charge are not exact, Tables 2 and 3, the 1-form of the cut in Figure 4 Computed	The general form of the Poincaré lemma, de Rham's theorem Known
Maxwell and the Hodge dual	The values of $*$ on the basis, the sign in (14), $F \wedge *F$, $F \wedge F$, the sign and eigenvalues of $**$ Computed	The non-abelian case (1-07, 3-03), the existence of instantons (3-05)
Yang–Mills	Why gauge fields can be written as forms, one of the special features of four dimensions	Shows nothing about the mass gap or the construction of the theory

Sources and reproduction

Item	Kind	Source / tools
Tables 1–3, the numbers in the text, Figures 1–4	**Computed on this machine**	Python 3.12, NumPy 1.26: `forms_algebra.py`, `stokes_cells.py`, `winding.py`, `cochains.py`, `monopole.py`, `maxwell_forms.py`, `hodge_basis.py`, `hemisphere.py`, `lattice_fig.py`. The figures are plain JavaScript (Figure 1 computes the determinant on the spot; the functions of Figures 2 and 3 were checked with node against Table 1 and `winding.py`; Figure 4 embeds the numbers from `lattice_fig.py`)

Item	Kind	Source / tools
Exterior derivative, Stokes, the Poincaré lemma, de Rham's theorem	Known (bibliographic data only)	M. Spivak, *Calculus on Manifolds* (1965); R. Bott, L. W. Tu, *Differential Forms in Algebraic Topology* (1982)
The origin of the exterior derivative	Primary source (bibliographic data only)	É. Cartan, Ann. Sci. ÉNS (3) 16 (1899) 239–332. Its placing as the origin of the exterior derivative relies on secondary sources
De Rham's theorem	Primary source (bibliographic data only; the attribution of the theorem checked in secondary sources)	G. de Rham, J. Math. Pures Appl. (9) 10 (1931) 115–200
Treating the point magnetic charge by gluing two regions (§ III)	Primary source (text checked)	T. T. Wu, C. N. Yang, Phys. Rev. D 12 (1975) 3845–3857
Quantisation of the point magnetic charge	Primary source (bibliographic data only; attribution following the text of Wu–Yang)	P. A. M. Dirac, Proc. R. Soc. Lond. A 133 (1931) 60

Next chapter: 1-07 Vector bundles and connections (the gluing of § 08).

Vector bundles and connections — a rule for comparing separate spaces at each point along a path

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

What this chapter shows — We write the rule for comparing the vector spaces at each point along a path (a connection) as a matrix-valued 1-form A, and derive in one sequence parallel transport, curvature $F = dA + A \wedge A$, holonomy and the gauge transformation $A \mapsto gAg^{-1} - (dg)g^{-1}$. From the gluing of the monopole an integer (the first Chern number) comes out.

Prerequisite chapters — 1-04 (Lie algebras, the adjoint representation) and 1-06 (1-forms, the exterior derivative, the point magnetic charge of § 08). Knowing the phase factor of 2-03 makes for quicker reading (what § 08 needs is defined on the spot).

Said first — This is a preparatory chapter, and shows nothing about the mass gap. Gauss–Bonnet, Ambrose–Singer and the integrality of Chern numbers are cited without proof (Known). The space of gauge orbits is in 3-04.

Order of this chapter

1. Why bundles and connections — you cannot subtract until you have carried
2. Vector bundles — transition functions, trivial bundles and twisted bundles
3. Connections and covariant derivatives — the formula for gauge transformations
4. Parallel transport — once round the sphere (Figure 1). Detour: Foucault's pendulum
5. Curvature — the mismatch around a small loop, a connection that curves even when constant (Figure 2)
6. Holonomy — the Wilson loop
7. The geometry of gauge transformations — the Bianchi identity
8. Gluing and integers — phase factors from north and south (Figure 3), the first Chern number

01

Why bundles and connections — you cannot subtract until you have carried

Suppose a field ψ takes values in $\mathbb{C}^N$ at each point ($N = 3$ for the colour of quarks). In gauge theory we allow the basis of $\mathbb{C}^N$ at each point (which direction to call "the first colour") to be chosen afresh point by point (3-01). Then $\psi(x+h) - \psi(x)$ is a difference of components written in separate bases, and its value changes when the bases are rechosen. Tangent vectors of the sphere are the same: the tangent planes at the north pole and at the equator are different things. So we prepare a rule P for carrying one space to the other along a path, and subtract after carrying (e_i is a coordinate direction).

$$D_i\psi(x) = \lim_{h\to 0} \frac{P_{x+he_i\to x}\,\psi(x+he_i) - \psi(x)}{h} \quad (1)$$

The carrying rule is called a **connection**, and the spaces at each point bundled together a **vector bundle**. **Example.** In electromagnetism, $N = 1$, and the carrying rule was the phase factor $e^{\frac{iq}{\hbar}\int \mathbf{A}\cdot d\mathbf{l}}$ (2-03). This chapter extends that to matrices, and treats the "gluing of two sheets" of 1-06 § 08 in the same language.

02

Vector bundles — locally a product, possibly twisted globally

A **vector bundle** $E \to M$ attaches to each point x of a manifold M (1-05) an N-dimensional space E_x (the **fibre**), so that over a small open set U_α it can be identified with $U_\alpha \times \mathbb{C}^N$. N is called the **rank**, and one with $N = 1$ a **line bundle**. It can be defined in the same way with fibre $\mathbb{R}^N$, and this is called a **real vector bundle**. A basis giving the identification is a **local frame**, and on an overlap the components in the two frames are related by a matrix.

$$\psi_\alpha = g_{\alpha\beta}\,\psi_\beta, \qquad g_{\alpha\beta}\,g_{\beta\gamma} = g_{\alpha\gamma} \quad (2)$$

$g_{\alpha\beta}$ is called the **transition function**, and the formula on the right, where three overlap, the cocycle condition; when the transition functions take values in a group G (U(1), SU(N) and so on), G is called the **structure group**. Given transition functions satisfying (2), a bundle can be made by gluing Known . A smooth choice of an element of E_x at each point is a **section**, and the field ψ is a section.

Trivial bundles and twisted bundles

A bundle that can be identified with $M \times \mathbb{C}^N$ ($M \times \mathbb{R}^N$ if real) over the whole space is called **trivial**. The following three are not trivial Known : the Möbius strip (a real line bundle with fibre $\mathbb{R}$; covering the circle with two arcs, the transition function is +1 on one part of the overlap and −1 on the other), the tangent bundle TS^2 (a real bundle of rank 2), and the monopole bundle of § 08 ($n \neq 0$). A trivial bundle would have a section that is nowhere 0. On the Möbius strip, the value of a section written in the frame changes sign in one turn, so by the intermediate value theorem it is 0 somewhere. On TS^2 the hairy ball theorem (1-05) forbids it.

Making the fibre complex untwists it with the same transition functions. $-1 = e^{i\pi}$ can be moved continuously inside U(1) to 1 through e^{it} (t from π to 0), and moving transition functions continuously keeps the bundle isomorphic Known , so the complex line bundle with transition functions +1 and −1 on the overlap is trivial. On the circle, whether a twist remains is decided by whether the transition function can be moved continuously to 1 inside the group where it takes values. Inside $\{\pm 1\}$ it cannot; inside U(1) it can. In § 08 the winding number in U(1) carries this impossibility of moving. Bundles over $\mathbb{R}^4$ are trivial Known , but on S^4, with infinity regarded as one point, and on periodic lattices, non-trivial bundles appear (3-05).

03

Connections and covariant derivatives — $D = d + A$

Once a frame is fixed, the first-order term of the P of (1) becomes a matrix-valued 1-form $A = A_i\, dx^i$. $A_i(x)$ is an element of the Lie algebra $\mathfrak{g}$ of the structure group (1-03), and for U(N), SU(N) it is an anti-Hermitian matrix.

$$D\psi = d\psi + A\,\psi, \qquad D_i\psi = \partial_i\psi + A_i\,\psi \quad (3)$$

This is what we get by putting $P_{x+he_i \to x} = 1 + hA_i + O(h^2)$ in (1). For $D'\psi' = g\,D\psi$ to hold even when the frame is changed to $\psi' = g\psi$ ($g(x) \in G$), since $(d + A')(g\psi) = dg\,\psi + g\,d\psi + A'g\psi$ equals $g(d\psi + A\psi)$,

$$A' = g\,A\,g^{-1} - (dg)\,g^{-1} \quad (4)$$

is needed. This is the **gauge transformation**. The A_α and A_β of an overlap are also related by (4) with $g = g_{\alpha\beta}$: **gluing and gauge transformation are the same formula**.

Example (U(1)). Putting $A = -\frac{iq}{\hbar}\,\mathbf{A}\cdot d\mathbf{x}$, $g = e^{iq\chi/\hbar}$, (3) is the covariant derivative of 2-03 and (4) is $\mathbf{A} \mapsto \mathbf{A} + \nabla\chi$. Physics books write $A = -i\,A^a_\mu T_a\,dx^\mu$ with Hermitian T_a (1-04), and call the real A^a_μ the gauge field (for where the coupling constant is put, see 3-

02). The difference of two connections changes as $g(A_1 - A_2)g^{-1}$ (the dg term cancels), and is a quantity that can be added and subtracted. The set of all connections is then an **affine space** — a set with no fixed origin, in which the difference of two points is an element of a vector space. Choosing one reference connection A_0, every connection can be written as A_0 + (a $\mathfrak{g}$-valued 1-form) (3-04).

04

Parallel transport — an ordinary differential equation along a path

Along a path $\gamma : [0,1] \to M$, carry v so that the covariant derivative is 0.

$$\frac{dv}{dt} + A\big(\dot\gamma(t)\big)\, v = 0, \qquad v(1) = P_\gamma\, v(0) \tag{5}$$

The solution is unique, and P_γ is called **parallel transport**. If A is anti-Hermitian, $\frac{d}{dt}|v|^2 = -v^*(A + A^*)v = 0$, so P_γ is unitary. The $A(\dot\gamma)$ at different times do not commute, so it cannot be written $\exp(-\int A)$; instead the solutions over short intervals are multiplied in order of time.

$$\begin{aligned} P_\gamma &= \lim_{n\to\infty} e^{-A_n \Delta t} \cdots e^{-A_1 \Delta t} \\ &=: \mathcal{P}\exp\Big(-\int_\gamma A\Big) \end{aligned} \tag{6}$$

$A_k = A(\dot\gamma(t_k))$ is the **path-ordered product**, with later times placed on the left. Joining paths gives $P_{\gamma_2\gamma_1} = P_{\gamma_2} P_{\gamma_1}$, and under gauge transformations it only undergoes the change of frame at the two ends.

$$P'_\gamma = g\big(\gamma(1)\big)\, P_\gamma\, g\big(\gamma(0)\big)^{-1} \tag{7}$$

Example: once round the sphere

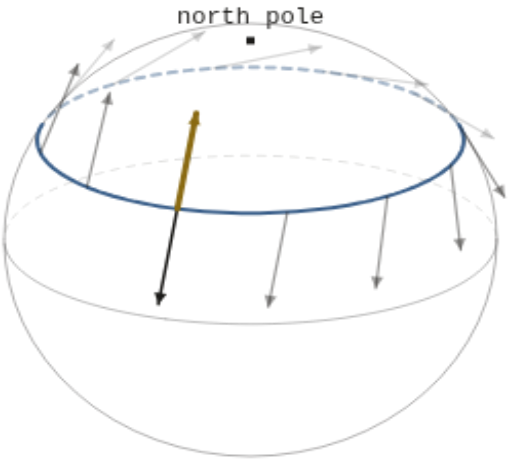
Carrying a tangent vector so that it does not bend as seen from inside the sphere is the Levi-Civita parallel transport; on the unit sphere it is $\dot v = -(v\cdot\dot\gamma)\,\gamma$ (allowing only change in the normal direction). A curve that keeps its own velocity vector $\dot\gamma$ under this parallel transport is called a **geodesic**; on the sphere these are great circles. Going east round the circle of latitude at colatitude θ_0, the angle relative to the south-pointing and east-pointing frame (e_θ, e_φ) changes by $\alpha(\varphi) = \alpha_0 - \varphi\cos\theta_0$, and in one turn

$$\Delta\alpha \equiv 2\pi(1 - \cos\theta_0) \pmod{2\pi} \tag{8}$$

The right-hand side is the area of the spherical cap enclosed by the circle of latitude. Circles of latitude other than the equator are not great circles (not geodesics), so the

vector turns relative to the frame. (8) is a special case of the Gauss–Bonnet theorem, "on a surface of curvature 1 the angle turned in one circuit equals the area enclosed" Known . For a **geodesic triangle**, whose three sides are geodesics, the angle turned also equals the area; the bottom row of Table 1 is an example.

Figure 1 Change the colatitude θ_0 with the slider (or a row of Table 1), and go east once round the circle of latitude with "Carry". After one turn the thick arrow is displaced from the thin arrow at the start (the faint arrows are in between), and the angle equals the area of the cap (8).



θ_0 = 60°, φ = 360°, α = -180.0°, (8) = 180.0°

θ_0	Turn in one circuit (RK4)	Difference from (8)
10°	5.469209°	2×10^{-12}
60°	**180.000000°**	**4×10^{-11}**
90°	0° (= 360°)	6×10^{-14}
Geodesic triangle (area $\pi/2$)	90.000000°	—

Computed **Table 1** The turn in one circuit obtained by integrating (5) with RK4 (angle relative to the frame, mod 360°). The bottom row is the geodesic triangle joining the north pole and the points at longitude 0° and 90° on the equator. The figure computes $\alpha = -\varphi\cos\theta_0$ in the browser. If JS is disabled, this table is the content of the figure.

Detour: Foucault's pendulum

A subsection you may skip.

The plane of oscillation appears to turn relative to the ground, but to be exact what turns is the frame of the ground (north, east), and the plane of oscillation is parallel-transported along the circle of latitude (in the approximation of slow rotation of the Earth) Known . At latitude λ north, put $\theta_0 = 90^\circ - \lambda$ in (8): in one sidereal day $-2\pi\sin\lambda$ (clockwise in the northern hemisphere). At latitude 48.85° north (near Paris) it is 271.08° Computed .

05

Curvature — the mismatch around a small loop

Whether parallel transport depends on the path can be measured by the mismatch after going round a small loop. In the difference of covariant derivatives taken in swapped order, the derivatives of ψ cancel, and only matrix multiplication remains.

$$[D_i, D_j]\,\psi = F_{ij}\,\psi, \qquad F_{ij} = \partial_i A_j - \partial_j A_i + [A_i, A_j] \quad (9)$$

Gathered into forms, it is $F = dA + A \wedge A$. dA is the exterior derivative component by component, and the wedge product of matrix-valued forms multiplies the component matrices and takes the wedge product of dx.

$$A \wedge A = \sum_{i,j} A_i A_j\, dx^i \wedge dx^j = \tfrac{1}{2} \sum_{i,j} [A_i, A_j]\, dx^i \wedge dx^j$$

The second equality is the average with the formula with the names i and j swapped ($dx^j \wedge dx^i = -dx^i \wedge dx^j$). For a scalar 1-form α, $\alpha \wedge \alpha = 0$ (1-06), but matrices do not commute, so it is not 0. The coefficient of $F = \sum_{i<j} F_{ij}\, dx^i \wedge dx^j$ is the F_{ij} of (9), exactly the formula of the Jaffe–Witten problem description (1-03). F is called the **curvature**, and in physics the **field strength**.

Checking with a small loop

Going once anticlockwise round a square of side ε in the x and y directions from the point p,

$$P_{\square} = 1 - \varepsilon^2 F_{xy}(p) + O(\varepsilon^3) \quad (10)$$

Putting (4) into (9) gives $F' = gFg^{-1}$. For U(1), F does not change under gauge transformations and is (apart from the factor $-iq/\hbar$) the electric and magnetic fields themselves (1-06 § 09), but in the non-abelian case F turns with the frame, and what does not change are combinations such as $\operatorname{tr}(F_{ij}F_{kl})$ (the action of 3-03).

Computed **Table 2** Integrating (5) with RK4 for a smooth $\mathfrak{su}(2)$-valued connection and comparing with (10) (operator norm, $|F_{xy}| = 0.369$).

ε	$\lvert \frac{P_{\square} - 1}{\varepsilon^2} + F_{xy} \rvert$	With dA in place of F
0.2	0.076538	0.348489
0.1	0.038844	0.324000
0.05	0.019532	0.311139

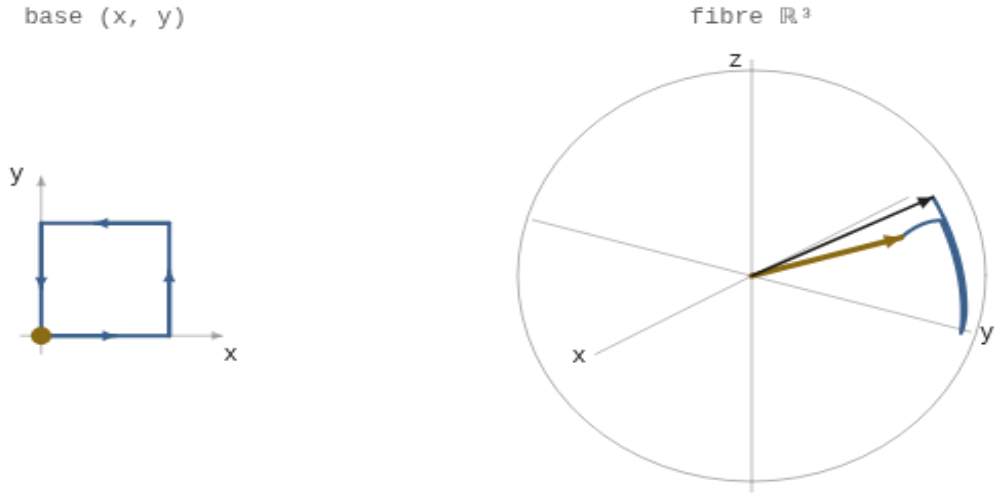
ε	$\lvert\frac{P_\square - 1}{\varepsilon^2} + F_{xy}\rvert$	With dA in place of F
0.025	0.009790	0.304592

Halving ε halves the error too ($O(\varepsilon)$). With dA, where the commutator is dropped, the error approaches $\lvert[A_x, A_y]\rvert = 0.298$ and does not vanish. For a random g, (7) differs by 1.5×10^{-12}, and $F' = gFg^{-1}$ by 3×10^{-7} (the error of differencing) Computed .

A connection that curves even when constant

In the non-abelian case F is not 0 even when A is constant. Take the group to be SO(3) (the adjoint representation of SU(2), 1-04) and the fibre $\mathbb{R}^3$ (the real vector bundle of § 02). Here L_x, L_y, L_z are the real antisymmetric matrices $(L_a)_{bc} = -\varepsilon_{abc}$, the Hermitian L_a of 1-04 § 02 multiplied by $-i$. They satisfy $L_a v = e_a \times v$, $[L_x, L_y] = L_z$, and e^{tL_z} is the rotation by angle t about the z axis (the same as the e^{-itL3} of 1-04). Putting $A_x = aL_x$, $A_y = aL_y$, we have $dA = 0$ and yet $F_{xy} = a^2 L_z$. A constant A for U(1) can be removed by $g = e^{A_x x + A_y y}$ (pure gauge), but here $F' = gFg^{-1}\ /= 0$, so no g can remove it.

Figure 2 On the left is the square in the base, on the right the fibre $\mathbb{R}^3$. With "Carry", while the point on the left goes once round, the vector on the right moves by (5), and after one turn the thick arrow is displaced from the thin arrow at the start. The slider (or a row of Table 3) is $s = \varepsilon a$, and "Make commutative" is $A_y = aL_x$.



s = 0.60, angle between v_0 and v 19.52°, about the axis (-0.28, 0.28, -0.92) by 20.04° (s^2 = 20.63°)

s	Rotation angle θ (rad)	θ / s²	Rotation axis
0.1	0.009992	0.999168	(−0.050, 0.050, −0.998)
0.4	0.157919	0.986994	(−0.195, 0.195, −0.961)
1.2	1.297942	0.901348	(−0.492, 0.492, −0.719)

Computed **Table 3** The rotation angle and axis of the composition of the rotations along the four sides $e^{sL_y} e^{sL_x} e^{-sL_y} e^{-sL_x}$ ($\lvert P_\square - 1\rvert < 10^{-15}$ if commutative). The figure computes the same composition in the browser. If JS is disabled,

this table is the content of the figure.

θ/s^2 approaches 1 as s → 0, and the axis is almost −z, agreeing with the $1 - s^2 L_z$ of (10) (a positive rotation about −z). This is the geometric face of "the gauge field interacting with itself" (3-02).

06

Holonomy — what a closed path leaves behind

The parallel transport along a path γ that leaves the point x_0 and returns is called the **holonomy** Hol_γ. By (7), since both ends are the same point, it changes by conjugation under gauge transformations, and its trace does not change.

$$\begin{aligned} \mathrm{Hol}'_\gamma &= g(x_0)\,\mathrm{Hol}_\gamma\, g(x_0)^{-1}, \\ W(\gamma) &= \tfrac{1}{N}\,\mathrm{tr}\,\mathrm{Hol}_\gamma \end{aligned} \qquad (11)$$

The gauge-invariant $W(\gamma)$ is called the **Wilson loop**.

- **It remains even when flat.** If $F = 0$, the holonomy does not change under continuous deformation of the loop Known . In a space with a hole, the holonomy of a loop that cannot be shrunk can be other than 1; this is the Aharonov–Bohm effect (2-03).
- **Curvature is the infinitesimal version of holonomy.** If M is connected, the Lie algebra of the set (subgroup) of holonomies of loops at x_0 is spanned by the curvatures $F_{ij}(y)$ at all points y carried to x_0 along every path joining y and x_0 (Ambrose–Singer 1953) Known .
- **The lattice.** Placing parallel transports $U_\ell \in G$ on the edges and using the plaquette product U_p in place of curvature is Wilson's lattice (5-01). From (10), with spacing a, $U_p \approx \exp(-a^2 F_{\mu\nu})$, hence $\mathrm{Re}\,\mathrm{tr}(1 - U_p) \approx -\frac{a^4}{2}\,\mathrm{tr}\,F_{\mu\nu}^2 \geq 0$ (μ, ν fixed and not summed; $F_{\mu\nu}^2 = -F_{\mu\nu}^* F_{\mu\nu} \leq 0$ because $F_{\mu\nu}$ is anti-Hermitian) Known .

07

The geometry of gauge transformations — changing frames, and what does not change

A gauge transformation is a smooth map $g : M \to G$, which moves a section to $g\psi$, the connection by (4), and the curvature to gFg^{-1}. The whole $\mathcal{G}$ is an infinite-dimensional group. As a metaphor, it is a rechoosing of the orientation of the ruler at each point; to be exact, it can be read either as a change of frame (passive) or as an **automorphism** of the bundle — a map of the bundle that maps each fibre linearly and isomorphically

and does not move the points of the base — (active), and in both readings the formula is (4). What has meaning in physics are quantities that do not change under $\mathcal{G}$: Wilson loops, the $\mathrm{tr}(F \wedge *F)$ of the action, the topological quantity $\mathrm{tr}(F \wedge F)$ (3-05).

Curvature satisfies an identity independent of the gauge. $A \wedge F$ and $F \wedge A$ are also fixed by the same rule as in § 05 (multiply the component matrices and take the wedge product of dx), and the order of the matrices matters.

$$DF := dF + A \wedge F - F \wedge A = 0 \quad (12)$$

This is the **Bianchi identity**. Putting in $F = dA + A \wedge A$, it vanishes by $d(A \wedge A) = dA \wedge A - A \wedge dA$, $d^2 = 0$ and the associativity of the wedge product alone. For U(1) it is $dF = 0$ (the two Maxwell equations without sources). For a $\mathfrak{su}(2)$-valued connection in three dimensions, the left-hand side is within the differencing error 3.5×10^{-7}; dropping the commutator gives 0.553 Computed .

The $\mathcal{A}/\mathcal{G}$ obtained by dividing the set of all connections $\mathcal{A}$ by $\mathcal{G}$ is the space of configurations that physics can distinguish, and choosing a representative continuously from each orbit (a global gauge fixing) is in general impossible (Singer, 3-04, 4-08). In the translation table of Wu–Yang (1975), gauge potential = connection, phase factor = parallel transport and field strength = curvature are listed, and only the entry for sources is "?".

08

Gluing and integers — the monopole and the first Chern number

In 1-06 § 08, the potential of the point magnetic charge $\hat{\mathbf{r}}/r^2$ could be written only with two sheets, north and south, and the difference on the overlap was $2\,d\varphi$. Multiplying the magnetic charge by g_m to get the magnetic field $g_m\hat{\mathbf{r}}/r^2$ (total flux $\Phi = 4\pi g_m$), the difference is $2g_m\,d\varphi$, and multiplying by $-iq/\hbar$ to make a connection as in the example of § 03 gives

$$A_N - A_S = -\frac{iq}{\hbar}\,2g_m\,d\varphi$$

On the other hand, (4) in the commutative case is $A_N - A_S = -g_{NS}^{-1}dg_{NS}$, so the transition function on the equator is $g_{NS} = e^{in\varphi}$, $n = 2qg_m/\hbar$. In the language of bundles, it is the U(1) bundle obtained by gluing the north and south sheets along the equator by this g_{NS}. g_{NS} must return to itself after going once round the equator, so **n is an integer**.

Seen through phase factors

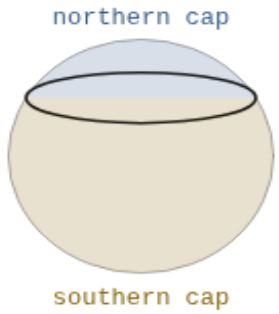
Seen through phase factors, the meaning of n takes a measurable form. The **phase factor** of a closed path C is the parallel transport (the holonomy of C, § 06) obtained by putting $A = -\frac{ig}{\hbar}\,\mathbf{A}\cdot d\mathbf{x}$ in (6). For U(1) no path ordering is needed, and by Stokes' theorem (1-06)

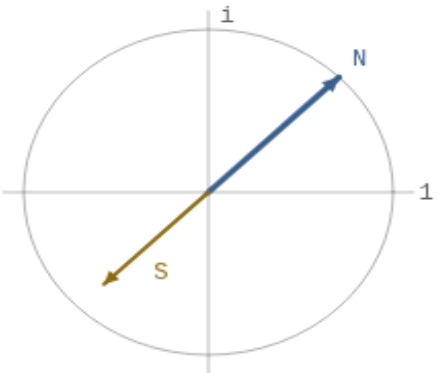
$$\mathrm{Hol}_C = \exp\!\left(\frac{iq}{\hbar}\oint_C \mathbf{A}\cdot d\mathbf{l}\right) = e^{iq\Phi C/\hbar} = e^{2\pi i\,\Phi C/\Phi 0}, \qquad \Phi_0 = \frac{h}{q}$$

(ΦC is the magnetic flux enclosed by C, and we take q > 0; for details see 2-03). $\Phi 0$ is called the **flux quantum**, and writing $f = \Phi/\Phi_0$, we have $n = 2qg_m/\hbar = q\Phi/(2\pi\hbar) = f$. **From here on n and f are the same quantity**; when we write f we also consider values that are not necessarily integers.

Computing the phase factor of the circle of latitude C at colatitude ϑ with A_N, usable on the northern cap, the enclosed flux is $\Phi(1-\cos\vartheta)/2$, giving $e^{i\pi f(1-\cos\vartheta)}$. Computing with A_S, usable on the southern cap, C goes round the boundary of the southern cap in the opposite direction, giving $e^{-i\pi f(1+\cos\vartheta)}$. The ratio is $e^{2\pi i f}$, and they agree only when f is an integer. This is the condition for the two descriptions to be a connection on one bundle, and is the reading, in the language of bundles, of Dirac's quantisation condition (1931).

Figure 3 Changing f and the colatitude ϑ with the sliders (or a row of Table 4) moves two arrows on the unit circle — the phase factor computed from the northern cap and the one computed from the southern cap. The two coincide only when f is an integer, and then they move together even when ϑ is changed. "Move f" increases f from 0 to 3.





f = 0.50, ϑ = 60°, north 45.0°, south -135.0°, mismatch 180.0°

f	From the north	From the south	Winding number of the transition function
0.5	0.707+0.707i	−0.707−0.707i	0.5 (does not return after one turn)
1	i	i	1
2	−1	−1	2

Computed **Table 4** On the circle of latitude at colatitude 60°. The winding number is the increase of the argument of $e^{if\varphi}$ once round the equator ÷ 2π. $\frac{i}{2\pi}\int F$ is f in every row. The figure computes the closed formulas of the two phase factors in the browser, and the angle of mismatch is $360^\circ \times f$ (integer multiples of 360° removed). If JS is disabled, this table is the content of the figure.

Read from curvature

The integer can also be read from curvature. Using Stokes' theorem on the northern and southern caps, and (4) on the equator (commutative, so $A_N - A_S = -g^{-1}dg$),

$$c_1 = \frac{i}{2\pi}\int_{S^2} F = \frac{1}{2\pi i}\oint_{\text{equator}} g_{NS}^{-1}\, dg_{NS} = n \quad (13)$$

The left-hand side is $\frac{i}{2\pi}\cdot\left(-\frac{ig}{\hbar}\right)\Phi = f$, and the right-hand side is the winding number of the transition function (1-06 § 07); the n = f seen above appears here too. c_1 is called the **first Chern number**. Changing the connection continuously cannot make an integer jump, so it is fixed by the bundle alone Known .

In four dimensions

SU(2) bundles over S^4 are classified by an integer $c_2 = \frac{1}{8\pi^2}\int \mathrm{tr}(F\wedge F)$ (the sign depends on the book) Known . On $\mathbb{R}^4$ too, solutions of the Yang–Mills equations with finite action approach a pure gauge $-(dg)g^{-1}$ ($g : S^3 \to SU(2)$) far away, and the winding number of g is the instanton number Known . This is not an unconditional fact. To be exact, it rests on the theorem that such solutions extend to connections on a bundle over S^4, obtained by adding one point at infinity (Uhlenbeck's removable singularity theorem); for general fields that are not solutions of the equations, the same is said by assuming suitable decay far away. For details see 3-05, 1-12 and the article on four dimensions.

What this chapter can say

Item	What is said	What is not said
Bundles, connections, parallel transport	Derivation of (2)–(8). Table 1, Figure 1 Computed	The non-triviality of the examples of § 02, that moving transition functions continuously keeps the bundle isomorphic, the general form of Gauss–Bonnet Known
Curvature and holonomy	(9)–(12). Tables 2 and 3, Figure 2, checks of (7) and $F' = gFg^{-1}$ Computed	Proofs of Ambrose–Singer and of the properties of flat connections Known

Item	What is said	What is not said
The monopole and the Chern number	That an integer comes out of the gluing, $n = \Phi/\Phi_0$, (13), Table 4, Figure 3 Computed	The general theory of Chern numbers, the classification by c_2, the extension of fields on $\mathbb{R}^4$ to S^4 Known
Yang–Mills	The geometric meaning of gauge fields, field strengths and gauge transformations	Shows nothing about the mass gap or the construction of the theory

Sources and reproduction

Item	Kind	Source / tools
Tables 1–4, the numbers of § 04–§ 07, Figures 1–3	**Computed on this machine**	Python 3.12, NumPy, SciPy (`sphere_holonomy.py` and 3 others). The figures are plain JavaScript, computing closed formulas on the spot
The ordering of the path-ordered product, the translation table (Table I), the monopole	Primary source (text checked)	T. T. Wu, C. N. Yang, Phys. Rev. D 12 (1975) 3845–3857
$F = dA + A \wedge A$	Primary source (text checked)	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory” (Clay)
Parallel transport, holonomy	Known (text checked)	S. M. Carroll, arXiv:gr-qc/9712019 (1997), Chapter 3
Levi-Civita, Ambrose–Singer, Dirac, Chern classes	Primary source (bibliographic data only)	T. Levi-Civita, Rend. Circ. Mat. Palermo 42 (1916) 173–204; W. Ambrose, I. M. Singer, Trans. AMS 75 (1953) 428–443; P. A. M. Dirac, Proc. R. Soc. A 133 (1931) 60–72; S. S. Chern, Ann. of Math. 47 (1946) 85–121, doi:10.2307/1969037 (page range checked in the zbMATH bibliographic data)
Four dimensions in § 08: removable singularities	Primary source (bibliographic data only)	K. Uhlenbeck, “Removable singularities in Yang–Mills fields”, Comm. Math. Phys. 83 (1982) 11–29, doi:10.1007/BF01947068
Seeing Foucault's pendulum as parallel transport	Paper (bibliographic data only), checked in secondary sources	J. B. Hart, R. E. Miller, R. L. Mills, “A simple geometric model for visualizing the motion of a Foucault pendulum”, Am. J. Phys. 55 (1987) 67–70, doi:10.1119/1.14972 (bibliographic data only); Wikipedia “Foucault pendulum” (text checked)
The examples of § 02, Gauss–Bonnet, c_2	Known (bibliographic data only)	Kobayashi–Nomizu, *Foundations of Differential Geometry* I (1963); Nakahara, *Geometry, Topology and Physics* (1990)

Next chapter: 1-08 Probability and measure. This chapter continues in 3-01–3-04.

Probability and measure — the mean settles down, and the fluctuations become Gaussian

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

What this chapter shows — The language that writes probability as "a rule that distributes weight over sets (a measure)". That the Gaussian distribution is fixed by its covariance alone, and higher moments are a count of pairings. That the law of large numbers and the central limit theorem say "the mean settles down, and the fluctuations become Gaussian", and that the rate is $1/\sqrt{n}$.

Prerequisite chapters — None (the example of § 03 is easier to read with 1-02).

Said first — We do not prove measure theory. Probability in infinite dimensions is the subject of 1-09. Constructing four-dimensional Yang–Mills theory as a probability measure is unsolved, and this site does not show it either. What the construction published in 2026-11 (6-00) made is correlation functions and a field theory reconstructed from them, not a probability measure (6-04).

Order of this chapter

1. Why probability
2. Probability spaces
3. Random variables and distributions — the uniform distribution on a group and the semicircle (Figure 1)
4. Expectation, variance, characteristic function
5. Independence
6. The Gaussian distribution — covariance and counting pairings (Figure 2, Detour: counting the Catalan numbers in two ways)
7. Gaussian fields on a lattice — correlation length and mass (Figure 3)
8. The law of large numbers — the mean settles down (Figure 4)
9. The central limit theorem — the fluctuations become Gaussian (Figure 4)

10. Detour: what does not become Gaussian

01

Why probability — field theory is written with probability measures

Moving quantum field theory to imaginary time (Euclideanisation, 4-02) attaches a weight e^{-S} (S is the action) to each configuration of the field, and after normalisation it can be read as a **probability**. The problem statement too lists 'non-Gaussian, Euclidean-invariant measures' among the by-products of constructing field theories Known . In lattice Yang–Mills a group element U_ℓ sits on each edge ℓ (5-01), and the probability has the following form.

$$d\mu(U) = \frac{1}{Z}\, e^{-S(U)} \prod_{\ell} dU_\ell, \qquad Z = \int e^{-S(U)} \prod_{\ell} dU_\ell \tag{1}$$

dU_ℓ is the uniform weight on the group (§ 03), and Z the constant that makes the total 1 (2-08). **Example:** if there is one edge, the group is the circle $U(1)$ and $S = -\beta \cos\theta$ (β is the inverse of the lattice coupling; the same letter as the inverse temperature of statistical mechanics), then (1) has density $e^{\beta\cos\theta}/Z$, and the larger β is, the more it gathers near $\theta = 0$. This chapter assembles the language for this.

02

Probability spaces — distributing weight over sets

A probability is a rule that distributes weights from 0 to 1 over subsets (**events**) of the whole Ω of what can happen. The family $\mathcal{F}$ of events that receive weights is assumed to contain Ω and to be closed under complements and countable unions (a **σ-algebra**). The triple $(\Omega, \mathcal{F}, P)$ is a **probability space**, and on P we impose the following Known .

$$P(\Omega) = 1, \qquad P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i) \quad (A_i \text{ pairwise disjoint}) \tag{2}$$

A non-negative set function with $\mu(\emptyset) = 0$ satisfying the second half of (2) (**countable additivity**) is a **measure**, and a measure with total 1 is a probability (Durrett § 1.1; the axiomatisation is said to begin with Kolmogorov's book of 1933). $\mathcal{F}$ is not taken to be all subsets because a construction using the axiom of choice produces subsets to which no length invariant under rotation and satisfying (2) can be assigned (Vitali 1905) Known . **Example:** a die is $P(A) = |A|/6$. The uniform distribution on $[0,1]$ is $P([a,b]) = b - a$, that is, length (**Lebesgue measure**). The probability of each point is 0, yet some

point always comes out; but (2) cannot be used for uncountable unions, so there is no contradiction. Something that holds with probability 1 is said to hold "**almost surely**".

03

Random variables and distributions — the uniform distribution on a group and the semicircle

A **random variable** is a real-valued function X on Ω such that every $\{X \le a\}$ is an event (**measurable**). Its **distribution** is the measure obtained by pushing the weight of Ω onto the number line by X.

$$\mu_X(B) = P\big(X^{-1}(B)\big), \qquad F_X(a) = P(X \le a), \qquad \mu_X(B) = \int_B \rho(x)\,dx \quad (3)$$

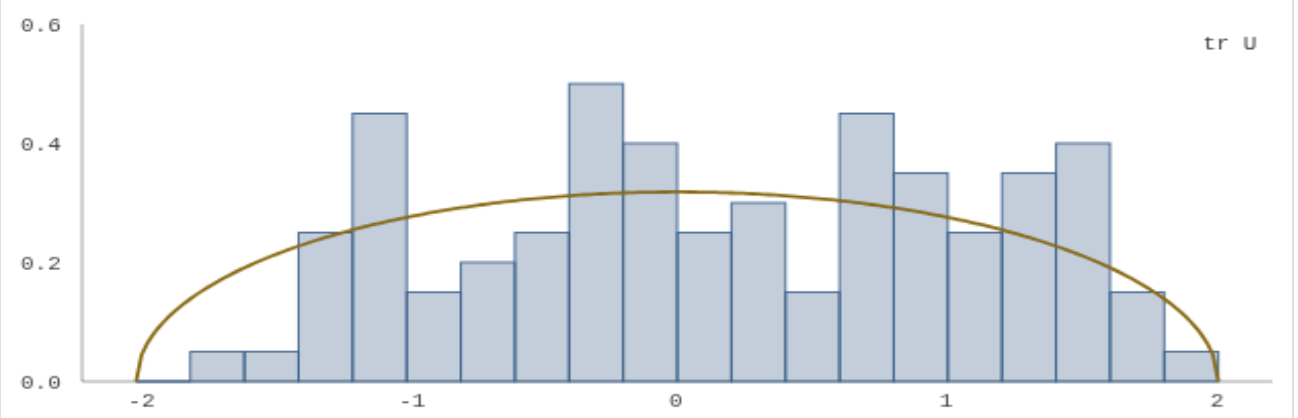
F_X is the **distribution function**, and ρ the **density**. **Example:** the number of heads in n tosses of a coin with probability of heads p follows the **binomial distribution** $\binom{n}{k}p^k(1-p)^{n-k}$, and keeping $np = \lambda$ while $n \to \infty$ gives the **Poisson distribution** $e^{-\lambda}\lambda^k/k!$.

A group of matrices that is closed and bounded, such as $SU(2)$, is called a **compact** group. A compact group G has exactly one probability measure unchanged by multiplication from either left or right, that is, satisfying $\mu(gA) = \mu(Ag) = \mu(A)$ ($gA = \{ga : a \in A\}$) for every $g \in G$ and subset A; it is called the **Haar measure** (for the circle, $d\theta/2\pi$) Known . The dU_ℓ of (1) is this. **Example:** elements of $SU(2)$ are $U = a_0 + i(a_1\sigma_1 + a_2\sigma_2 + a_3\sigma_3)$ (σ_i are the Pauli matrices, $|a| = 1$), and $SU(2)$ is the sphere S^3 (1-02). The Haar measure is the uniform distribution on S^3, and $\operatorname{tr} U = 2a_0$. The cross-section at height a_0 is a sphere of radius $r = \sqrt{1-a_0^2}$ (area $\propto r^2$), and the width of the band is da_0/r, so the density of a_0 is proportional to r, and $\operatorname{tr} U$ has the **semicircle** density Known .

$$\rho(x) = \frac{1}{2\pi}\sqrt{4 - x^2} \qquad (-2 \le x \le 2) \quad (4)$$

Making a histogram (20 bins) of $\operatorname{tr} U$ from 10^6 uniform points on S^3, the difference between the fraction falling in each bin and the probability of each bin given by (4) (0.019–0.064) was within 6.4×10^{-4} Computed . Figure 1 shows how it goes as the points increase. The same semicircle also appears as the distribution of eigenvalues of a large symmetric matrix with independent Gaussian entries, in the limit where it is suitably normalised and the size of the matrix goes to infinity (17-04).

Figure 1 The histogram of $\operatorname{tr} U$ (bars, 20 bins, area 1) made from the first n of the uniform points on S^3 (U following the Haar measure of $SU(2)$), and the semicircle (4) (curve). Increasing n from 10 to 10^6 with the slider or "Play" brings the bars towards the curve



n = 100, largest difference between bin fraction and bin probability 0.038 (1/√n = 0.10)

Computed The histogram of Figure 1 embeds the counts per bin of the first n points of a sample made on this machine (the same 10^6 points as in the text above). The curve and the bin probabilities are computed from formula (4). If JavaScript is disabled, it can be read from the numbers in the text above.

04

Expectation, variance, characteristic function — writing the mean as an integral

The **expectation** is the total of "value × weight", and for continuous cases an integral. The Lebesgue integral slices the values instead, and adds up c times "the weight of the set where the value is near c". As a metaphor, it is counting a pile of coins by denomination; to be exact, it is the limit of the sums $\sum_i c_i P(A_i)$ for functions taking only finitely many values.

$$E[X] = \int_\Omega X \, dP = \int_{\mathbb{R}} x \, d\mu_X(x), \qquad \operatorname{Var}(X) = E\big[(X - E[X])^2\big] \quad (5)$$

Taking the expectation of $\mathbf{1}_{\{|Y| \ge t\}} \le Y^2/t^2$ ($\mathbf{1}_A$ is 1 on A and 0 outside) as $Y = X - E[X]$ gives **Chebyshev's inequality** Known .

$$P\big(|X - E[X]| \ge t\big) \le \frac{\operatorname{Var}(X)}{t^2} \quad (6)$$

The **characteristic function** $\varphi_X(t) = E[e^{itX}]$ determines the distribution uniquely (Durrett § 3.3). **Example:** for a die, $E = 7/2$, $\operatorname{Var} = 35/12$. The $E[(-1)^N] = e^{-2\lambda}$ of the Poisson distribution is used in the counting of centre vortices (17-02). For (4), $E[\operatorname{tr} U] = 0$, $E[(\operatorname{tr} U)^2] = 1$.

05

Independence — how it differs from being uncorrelated

X and Y being **independent** means that the joint probability splits into a product.

$$P(X \in A,\ Y \in B) = P(X \in A)\,P(Y \in B) \qquad (\text{for all } A, B) \tag{7}$$

Then the distribution of (X, Y) is the measure that gives the weight $P(X \in A)P(Y \in B)$ to the rectangle $A \times B$ in the plane (the **product measure**), and $E[XY] = E[X]E[Y]$, $\operatorname{Var}(X+Y) = \operatorname{Var} X + \operatorname{Var} Y$, $\varphi_{X+Y} = \varphi_X \varphi_Y$ Known . If the **covariance** $E[XY] - E[X]E[Y]$ is 0 they are **uncorrelated**; independent implies uncorrelated, but the converse is false. **Example:** if X takes $\{-1, 0, 1\}$ with equal probability and $Y = X^2$, the covariance is 0 but $P(X = 0, Y = 0) = 1/3 \ \neq 1/9$.

In field theory we would like the fields at two distant points to be "almost independent". The problem statement gives **clustering** as a consequence of the mass gap Δ (4-11) Known . It is the property that the size of the correlation of local quantities $O(\underset{\sim}{x})$, $O(\underset{\sim}{y})$ with expectation 0 placed at points $\underset{\sim}{x}, \underset{\sim}{y}$ of space $\mathbb{R}^3$ (since the expectations are 0, the covariance above is just the expectation of the product) is, if $0 < C < \Delta$, at most $e^{-C|\underset{\sim}{x}-\underset{\sim}{y}|}$ where $|\underset{\sim}{x} - \underset{\sim}{y}|$ is large enough. To be exact, O are operators of quantum theory, and the correlation is the expectation in the ground state (the vacuum) (an example is in § 07).

06

The Gaussian distribution — covariance and counting pairings

A Gaussian distribution with mean 0 in n dimensions is fixed by the **covariance matrix** $C_{ij} = E[x_i x_j]$ alone (a positive definite symmetric matrix, 1-01).

$$\rho(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det C}} \exp\Big(-\frac{1}{2}\, x \cdot C^{-1} x\Big), \qquad E\big[e^{i\,t\cdot x}\big] = e^{-\frac{1}{2}\,t\cdot C t} \tag{8}$$

There are three properties Known . (i) The image under a linear map $y = Mx$ is also Gaussian, with covariance MCM^{T}. (ii) If jointly Gaussian, uncorrelated implies independent (if C is diagonal the density splits into a product). (iii) Higher moments are the sum over **all ways of pairing up** the variables of products of C (Isserlis 1918; in physics a formula of the same form is used as Wick's theorem (1950)).

$$E[x_1 x_2 x_3 x_4] = C_{12}C_{34} + C_{13}C_{24} + C_{14}C_{23} \tag{9}$$

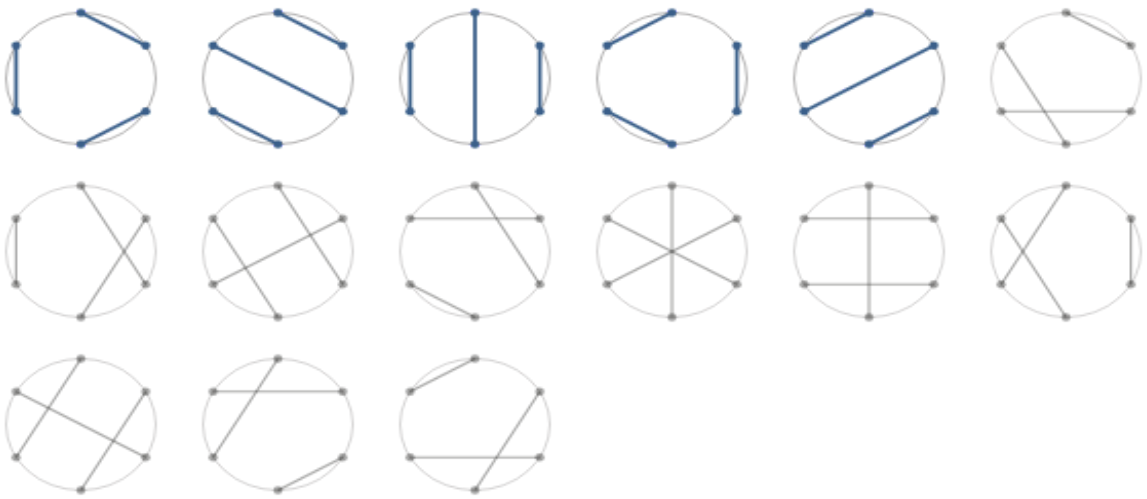
For random C of 4×4, against the right-hand side -0.0769 of (9), 4×10^6 samples gave -0.0774 ± 0.0005 Computed . In one variable it is $E[x^{2k}] = (2k-1)!!\,\sigma^{2k}$. $(2k-1)!! = 1 \cdot 3 \cdot 5 \cdots (2k-1)$ (the **double factorial**, multiplying only the odd numbers) is the number of ways of pairing up $2k$ points (the partner of point 1 can be chosen in $2k-1$ ways, and the same is repeated for the rest). On the other hand, the moments of the semicircle (4) are the number of ways of joining $2k$ points on a circle by chords that **do**

not cross, the **Catalan numbers** $\mathrm{Cat}_k = \binom{2k}{k}/(k+1)$ Paper (the detour computes both). Counting the non-crossing pairings and numerically integrating (4) both agreed with Cat_k for $k \le 6$ (the integral to 6 digits) Computed . Choosing a row of Table 1 lines up those ways of pairing for k in Figure 2.

Table 1 All pairings (Gaussian) and non-crossing pairings (semicircle). Choosing a row changes Figure 2 Computed

k	$(2k-1)!!$	Cat_k	$E[(\mathrm{tr}\,U)^{2k}]$ **(sample 10^6)**
1	1	1	1.001
2	3	2	2.001
3	**15**	**5**	**5.001**
4	105	14	14.00
5	945	42	42.01
6	10395	132	132.05

Figure 2 Ways of pairing $2k$ points on a circle by chords. Choose k by a row of Table 1 (or the slider). The coloured ones are the pairings whose chords do not cross (Cat_k ways), the faint ones the pairings that cross. The three for $k = 2$ correspond to the three terms $C_{12}C_{34}$, $C_{14}C_{23}$, $C_{13}C_{24}$ of (9) (number the points 1–4 clockwise)



k = 3, pairings 15 ((2k−1)!!), non-crossing 5 (Cat_3)

Computed The pairings of Figure 2 are listed in the browser, and their numbers have been checked to agree with the counts of Table 1 (computed on this machine). If JavaScript is disabled, they can be read from Table 1.

(9) is the prototype of Feynman diagrams (4-03), and non-crossing pairings lead to the planar diagrams of large N (17-04).

Detour: counting the Catalan numbers in two ways

A subsection you may skip.

The number of non-crossing pairings. Number the points on the circle 1 to $2k$ clockwise, and let c_k ($c_0 = 1$) be the number of non-crossing pairings. If the partner of point 1 is point p, the points between 1 and p cannot cross the chord $\{1, p\}$, so they can only pair among themselves. There is an even number of points between, so $p = 2j$, and the $j-1$ pairings inside and the $k-j$ pairings outside can be chosen separately. Written with the coefficients lined up in one power series (the **generating function**) $c(x) = \sum_{k\ge 0} c_k x^k$,

$$c_k = \sum_{j=1}^{k} c_{j-1} c_{k-j} \quad (k \ge 1) \qquad \Longleftrightarrow \qquad c(x) = 1 + x\, c(x)^2$$

Solving the quadratic equation, taking the root that is finite at $x \to 0$, and expanding $\sqrt{1-4x}$ by the binomial series gives the coefficients.

$$c(x) = \frac{1 - \sqrt{1-4x}}{2x}, \qquad \sqrt{1-4x} = 1 - \sum_{n\ge 1} \frac{2}{n}\binom{2n-2}{n-1} x^n \qquad \Longrightarrow \qquad c_k = \frac{1}{k+1}\binom{2k}{k}$$

The moments of the semicircle. Putting $x = 2\sin\theta$ in (4) gives $\sqrt{4-x^2} = 2\cos\theta$, $dx = 2\cos\theta\, d\theta$. By integration by parts $I_n = \int_{-\pi/2}^{\pi/2} \sin^n\theta\, d\theta$ satisfies $I_n = \frac{n-1}{n} I_{n-2}$, $I_0 = \pi$, so $I_{2k} = \pi\,(2k-1)!!/(2k)!!$ ($(2k)!! = 2\cdot 4 \cdots 2k = 2^k k!$). Using $\cos^2\theta = 1 - \sin^2\theta$,

$$\int_{-2}^{2} x^{2k} \rho(x)\, dx = \frac{2^{2k+2}}{2\pi}\left(I_{2k} - I_{2k+2}\right) = \frac{2^{2k+1}}{\pi} \cdot \frac{I_{2k}}{2k+2} = \frac{2^{2k}\,(2k-1)!!}{(k+1)\,(2k)!!} = \frac{1}{k+1}\binom{2k}{k}$$

The last equality follows from $(2k-1)!! = (2k)!/(2^k k!)$. The two computations arrived separately at the same Cat_k Paper . Why the semicircle counts non-crossing pairings becomes visible in the counting of planar diagrams of large N (17-04).

07

Gaussian fields on a lattice — correlation length and mass

To N variables $x_0, \dots, x_{N-1}$ ($x_N = x_0$) arranged on a ring, attach a weight under which neighbours prefer close values and the whole is pulled back towards 0.

$$\rho(x) \propto \exp\Big(-\frac{1}{2}\sum_j \left[(x_{j+1} - x_j)^2 + m^2 x_j^2\right]\Big) = \exp\Big(-\frac{1}{2}\, x \cdot (-\Delta_{\mathrm{lat}} + m^2)\, x\Big) \qquad (10)$$

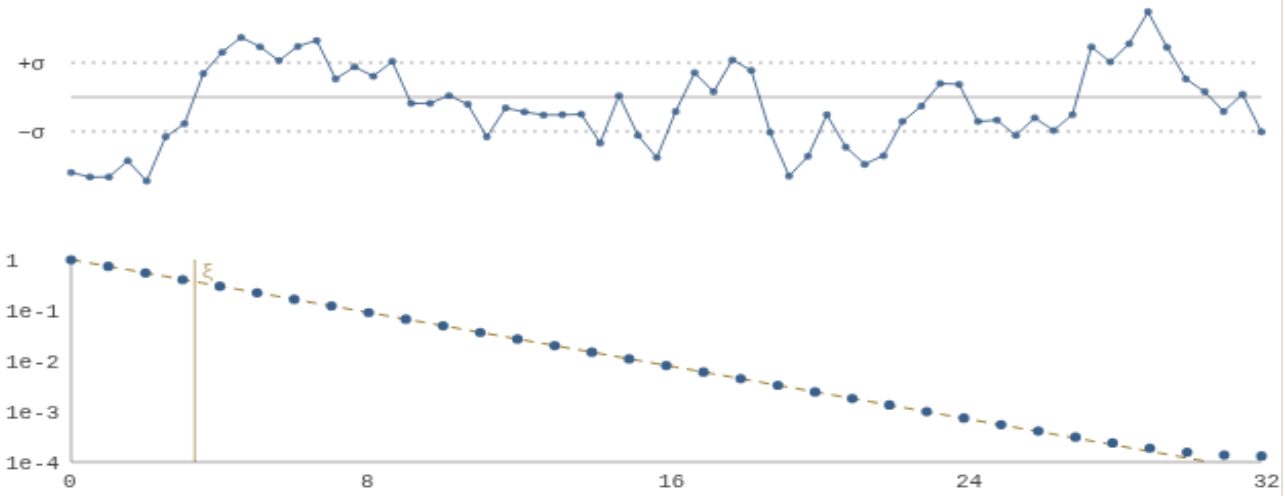
$(\Delta_{\mathrm{lat}} x)_j = x_{j+1} - 2x_j + x_{j-1}$ is the **lattice Laplacian**. It is different from the mass gap Δ of § 05, so we distinguish it with the subscript lat. Comparing with (8), (10) is the Gaussian distribution with covariance $C = (-\Delta_{\mathrm{lat}} + m^2)^{-1}$. The problem statement calls the Gaussian measure on the continuum $\mathbb{R}^d$ with a covariance of the same form (with the lattice Laplacian replaced by the continuum one) 'the free, mass-m0 field'. (10) is its analogue on a one-dimensional lattice (a ring), and on a d-

dimensional lattice too the covariance has the same form $(-\Delta_{\text{lat}} + m^2)^{-1}$ Known . If the ring is infinitely long, $C(0,x) = e^{-|x|/\xi}/(2\sinh(1/\xi))$ exactly Known .

$$\frac{C(0,x)}{C(0,0)} \approx e^{-|x|/\xi}, \qquad \cosh\frac{1}{\xi} = 1 + \frac{m^2}{2}, \qquad \xi \approx \frac{1}{m} \quad (m \text{ small}) \tag{11}$$

ξ is the **correlation length**, and the mass m is roughly its inverse (both measured in units of the lattice spacing). This rate of decrease corresponds to the "gap" of quantum theory (2-06). (10) is the discretised imaginary-time path integral of an oscillator, and m corresponds to the $a\omega$ of 2-07 (there $\omega = 1$, so it is the step a itself). The general theory of correlation length is in 2-08.

Figure 3 A Gaussian field on a ring ($N = 64$). Moving m changes the smoothness of the sample above (random numbers fixed; the dotted line is $\pm\sigma$) and the $C(0,x)/C(0,0)$ below (log scale, dashed line $e^{-x/\xi}$)



m = 0.302, ξ = 3.32, 1/m = 3.31, σ = 1.279, C(0,8)/C(0,0) = 0.0901

Table 2 Correlations on a ring ($N = 64$) Computed

m	ξ ((11))	$C(0,8)/C(0,0)$	ξ read from the slope
0.05	20.00	0.703	24.1
0.1	10.00	0.452	10.2
0.3	3.35	0.0915	3.35
1	1.04	4.5×10^{-4}	1.04
2	0.57	7.5×10^{-7}	0.57

Computed The numbers of Figure 3 and Table 2 were computed on this machine (the figure is drawn from the exact correlations and fixed random numbers; the slope from $x = 4$–12). If JavaScript is disabled, they can be read from Table 2.

For $m = 0.05$ the ξ read from the slope is larger than 20 because the ring is only about 3 times the correlation length, and the contribution $e^{-(N-x)/\xi}$ from going the other way round matters (a finite-box effect, 5-09). (10) is Gaussian, so by (9) all correlations are fixed by C, and there is no interaction (a **free field**).

08

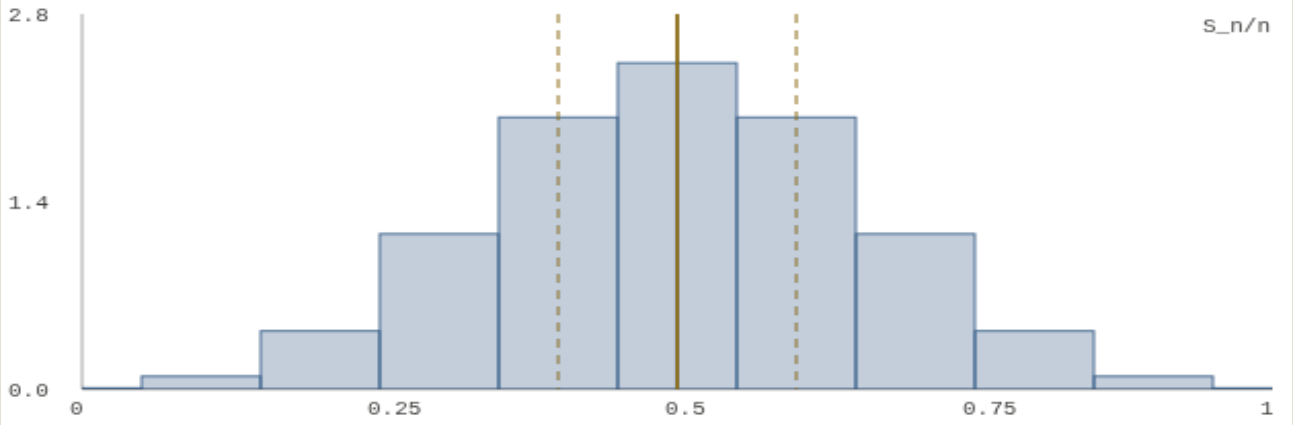
The law of large numbers — the mean settles down

Let S_n be the sum of independent $X_1, X_2, \ldots$ following the same distribution (mean μ, variance σ^2). From § 05, $\mathrm{Var}(S_n/n) = \sigma^2/n$, and putting this into (6) gives the **weak law of large numbers** Known .

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2} \longrightarrow 0 \qquad (n \to \infty) \quad (12)$$

Moreover, with $E|X_1| < \infty$ alone, $S_n/n \to \mu$ holds almost surely (the **strong law**, Durrett Theorem 2.4.1) Known . In "As is" in Figure 4, you can see the distribution of S_n/n shrinking to p and the left-hand side of (12) ($\varepsilon = 0.1$).

Figure 4 The distribution of the number of heads in n tosses of a coin with probability of heads p (bars, area 1). Increasing n, in "As is" it shrinks to p (vertical line) and the probability outside the dashed lines decreases; in "Magnify by $\sqrt{n}$" it approaches the standard Gaussian (curve)



n = 10, p = 0.50, P(|S_n/n − p| ≥ 0.1) = 0.754, right-hand side of (12) 2.50

Computed The numbers of Figure 4 and Table 3 were computed on this machine (the figure computes on the spot from the binomial distribution and an approximation formula for Φ, agreeing with the Python values to 4 digits). If JavaScript is disabled, they can be read from the next paragraph and Table 3.

For $p = 1/2$, $\varepsilon = 0.1$, the left-hand side of (12) is 0.754, 0.0569, 2.7×10^{-10} at $n = 10, 100, 1000$, and the right-hand side $25/n$ is 2.5, 0.25, 0.025 Computed . The right-hand side is a crude upper bound, but enough for the limit. The practical reading is that **the error of a sample mean is about $\sigma/\sqrt{N}$**, so ten times the precision needs a hundred times the samples. Estimating $E[(\mathrm{tr}\, U)^2] = 1$ by sample means gave 0.916 ± 0.092 for $N = 10^2$ and 1.0015 ± 0.0010 for 10^6 (a different sample from Table 1; the error bar is the sample standard deviation $/\sqrt{N}$) Computed . In lattice Monte Carlo, successively generated configurations are not independent, so the effective N decreases (5-07).

09

The central limit theorem — the fluctuations become Gaussian

The deviation $S_n/n - \mu$ is of size $1/\sqrt{n}$. Magnifying it by $\sqrt{n}$ makes the **Gaussian distribution** appear, whatever the original distribution (the **central limit theorem**, Durrett Theorem 3.4.1) Known .

$$P\Big(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le a\Big) \longrightarrow \Phi(a) = \int_{-\infty}^{a} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx \qquad \text{(for all } a\text{)} \quad (13)$$

The reason is visible through characteristic functions. If $\mu = 0$, then $\varphi(s) = 1 - \sigma^2 s^2/2 + o(s^2)$, and by independence $\varphi(t/\sigma\sqrt{n})^n = (1 - t^2/2n + o(1/n))^n \to e^{-t^2/2}$, which is the characteristic function of the one-dimensional standard Gaussian of (8). The continuity theorem, which derives convergence of distributions from convergence of characteristic functions (Durrett Theorem 3.3.17), gives (13) (the classical proof under general conditions is Lindeberg 1922). The rate of approach is given by the **Berry–Esseen inequality** ($\rho = E|X - \mu|^3$).

$$D_n = \sup_a \Big| P\Big(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le a\Big) - \Phi(a) \Big| \le \frac{C_0\, \rho}{\sigma^3 \sqrt{n}} \quad (14)$$

Durrett proves it with $C_0 = 3$, and Shevtsova (2011) has shown $C_0 < 0.4748$ Known . Both sides can be compared in the magnified view of Figure 4 and in Table 3.

Table 3 D_n of the binomial distribution (exact) and the right-hand side of (14) ($C_0 = 0.4748$) Computed

p	n	D_n	Right-hand side	$D_n\sqrt{n}$
0.5	10	0.1230	0.1501	0.389
0.5	1000	0.0126	0.0150	0.399
0.1	100	0.0832	0.1298	0.832
0.02	100	0.1767	0.3258	1.767

For $p = 1/2$, $D_n\sqrt{n}$ settles at $1/\sqrt{2\pi} \approx 0.399$. If n is even, the distribution function has a step of height about $\sqrt{2/\pi n}$ at $a = 0$, and the smooth Φ passes through its middle, so half the step always remains (Durrett § 3.4.4). If n is odd there is no step in the centre, and between the two central steps ($a = \pm 1/\sqrt{n}$) the distribution function stays flat at $1/2$, leaving a deviation of the same size at its ends ($D_n\sqrt{n} = 0.393$ at $n = 11$, 0.399 at $n = 199$) Computed . Either way it is not faster than $1/\sqrt{n}$. For $p = 0.02$, $n = 100$ we have $np = 2$: the distribution is close to an asymmetric Poisson distribution, and still far from Gaussian. Limit theorems do not say "when it is close", so an estimate of the rate is needed.

```
10
```

Detour: what does not become Gaussian — heavy tails, triviality

A section you may skip.

(13) assumes that the variance is finite. The **Cauchy distribution** (density $1/\pi(1+x^2)$, characteristic function $e^{-|t|}$) has $E|X| = \infty$, and the mean of n of them follows the same distribution as the original (Durrett Exercise 3.3.6) Known .

In field theory, conversely, "collapsing to a Gaussian" becomes the problem. For the ϕ^4 model on a four-dimensional lattice, it has been shown that the limit of refining the lattice near the critical point is Gaussian and the interaction vanishes (**triviality**) (Aizenman–Duminil-Copin 2021, 7-08) Known . As a metaphor it is the central limit theorem's way of collapsing; to be exact, it is a theorem about strongly correlated fields, and the tools of proof are different.

The "non-trivial" Yang–Mills theory of the problem statement is, roughly, a theory that is not Gaussian in the limit. As the problem statement (§ 6.2) says of scalar fields, the free field is obtained from a Gaussian measure, and 'one requires non-Gaussian dμ to obtain nonlinear fields'. The lattice (1) is not Gaussian to begin with, but there is no result showing that it survives in the limit as a non-Gaussian measure with clustering, and the problem statement itself writes that at present not a single non-trivial relativistic field theory satisfying the axioms in four dimensions is known (as written in 2000) Known . The construction published in 2026-11 made a non-trivial theory in four dimensions, at the level not of a measure but of correlation functions (the committee's assessment is that it is proved in Lean; 6-01). This site does not show the non-triviality of a measure (0-03, 7-10).

What this chapter can say

Said	Not said
Known The definitions, (6), the proof of (12), the derivation of (4). The other theorems are cited from the literature.	Proofs of measure theory (the extension theorem, the convergence theorems, the continuity theorem).
Paper That the moments of the semicircle and the number of ways of joining $2k$ points on a circle by non-crossing chords are both Cat_k (the two computations of the detour of § 06). For $k \le 6$ also checked by counting and by numerical integration Computed .	Why the two agree (the planar diagrams of large N, treated in 17-04).
Computed Tables 1–3, Figures 1–4, the numbers of § 03, § 06, § 08, § 09.	The numbers are checks with finite k, n and samples, and are not a substitute for proofs of the limits.

Said	Not said
Known From the problem statement: that the free field is a Gaussian measure, clustering, that a non-trivial theory needs a non-Gaussian measure. The triviality of four-dimensional ϕ^4.	Measures in infinite dimensions (1-09). The existence of the measure of four-dimensional Yang–Mills (unsolved; the construction published in 2026-11 did not make a measure either; 6-04). This site does not show the non-triviality or clustering of the measure either.

Sources and reproduction

Item	Kind	Source / tools
The theorems and remarks of § 02, § 04, § 08– § 10	Known	R. Durrett, *Probability: Theory and Examples*, 5th ed., Cambridge Univ. Press (2019) (Theorems 2.4.1, 3.4.1, 3.3.17, 3.4.17, the remark of § 3.4.4, Exercise 3.3.6). Text checked in the author's public version.
Original papers (axiomatisation, Vitali, Haar, (9), (13), (14))	Known	A. Kolmogoroff, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer (1933). G. Vitali, Bologna (1905). A. Haar, Ann. of Math. 34 (1933) 147. L. Isserlis, Biometrika 12 (1918) 134–139. G. C. Wick, Phys. Rev. 80 (1950) 268–272. J. W. Lindeberg, Math. Z. 15 (1922) 211–225. A. C. Berry, Trans. AMS 49 (1941) 122–136. C.-G. Esseen, "On the Liapounoff limit of error in the theory of probability", Ark. Mat. Astron. Fys. A 28, No. 9 (1942) 1–19. All bibliographic data only.
$C_0 < 0.4748$, the triviality of § 10	Known	I. Shevtsova, arXiv:1111.6554 (2011). M. Aizenman, H. Duminil-Copin, Ann. of Math. 194 (2021) 163–235, arXiv:1912.07973 (abstract checked). Same authors, Corrigendum, Ann. of Math. 199 (2024) 479 (bibliographic data only).
Quotations from the problem statement (§ 01, § 05, § 07, § 10)	Known	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory" (the problem description of the Clay Mathematics Institute). Included in J. Carlson et al. (eds.), *The Millennium Prize Problems*, Clay Math. Inst. / AMS (2006) 129–152. Text checked in Clay's public version (formula (2) of § 5, § 6.2).
The Catalan numbers of § 06 (detour)	Paper	Computations within this chapter (the recursion by integration by parts and the binomial series). The semicircle of Gaussian matrices is in 17-04 and its table of sources.
Tables, figures, numbers	Computed	Python 3.12 on this machine (NumPy, SciPy): moments.py (Table 1, § 03, § 08; the numerical integration of the semicircle is $k \le 6$), pairings.py (counting for $k \le 6$, (9)), fig_semicircle.py (the counts per bin for Figure 1), lattice_gauss.py (Table 2, Figure 3), clt.py (Table 3, Figure 4), clt_odd.py (odd n in § 09). The functions of the figures were checked against the tables with fig_check.js and fig_check2.js (node).

Next chapter: 1-09 Probability in infinite dimensions — probability on fields themselves, and the Gaussian measure as the starting point.

Probability in infinite dimensions — there is no uniform weight, so we start from Gaussian measures

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

What this chapter shows — That in infinite dimensions there is no weight unchanged by translation (Lebesgue measure), and the starting point becomes the Gaussian measure. That samples fall, with probability 1, outside the set of directions in which one may shift (the Cameron–Martin space). That Brownian motion, white noise and the free field stop even being functions as the dimension grows.

Prerequisite chapters — 1-08. ℓ^2, Fourier coefficients on the torus and distributions are defined on the spot as far as needed (properly in 1-10 and 1-11).

Said first — The big theorems of measure theory are only cited. All the measures made in this chapter are Gaussian, and measures with interaction are touched on only by citation in § 10. Constructing four-dimensional Yang–Mills theory as a measure is unsolved, and this site does not show it either. What the construction published in 2026-11 (6-00) made is correlation functions and a reconstructed field theory, not a measure (6-04).

Order of this chapter

1. Why infinite dimensions — field variables are infinitely many
2. There is no Lebesgue measure — packing balls in orthogonal directions
3. Gaussians in high dimensions live on a shell — reducing the variance
4. Gaussian measures in infinite dimensions — every shadow is Gaussian
5. The Cameron–Martin theorem — directions in which one may shift (Figure 1)
6. Brownian motion — the path settles, the energy diverges (Figure 2)
7. White noise — no value at a point
8. The free field and dimension — the variance at a point diverges (Figure 3)
9. Products of fields — distributions cannot be multiplied
10. Detour: mutually singular measures — scale, mass, interaction

01

Why infinite dimensions — field variables are infinitely many

The lattice probability of 1-08 could be written as "density × volume element" because the variables were finitely many. Refining the lattice and enlarging the box makes the variables infinitely many. Physics books write the probability of a free scalar field of mass m (Euclideanised, 4-02) like this.

$$d\mu(\varphi) \overset{\text{formally}}{=} \frac{1}{Z}\exp\Big(-\frac{1}{2}\int_{\mathbb{R}^d}\big(|\nabla\varphi|^2+m^2\varphi^2\big)\,dx\Big)\prod_{x\in\mathbb{R}^d} d\varphi(x) \quad (1)$$

Neither of the two factors on the right-hand side has a meaning on its own. The measure corresponding to $\prod_x d\varphi(x)$ does not exist (§ 02), and the integral in the exponent is $+\infty$ for almost every sample (§ 05, § 06). Even so, μ as a product can be constructed as a Gaussian measure (§ 04). The problem statement too describes the measures produced by constructive field theory as 'non-Gaussian, Euclidean-invariant measures on spaces of generalized functionals' Known . That means measures on spaces of distributions (§ 07) that are not Gaussian and are unchanged by rotations and translations. This chapter treats the Gaussian case that is their starting point.
Example: dividing the interval into N equal parts with $d=1$, (1) is the Gaussian distribution in N variables of 1-08 § 07, and where $N\to\infty$ leads is the subject of this chapter.

02

There is no Lebesgue measure — packing balls in orthogonal directions

Length and area (Lebesgue measure) are unchanged by translation and give finite weight to bounded sets. In infinite dimensions these two cannot both hold. We state this with the space $\ell^2=\{x=(x_1,x_2,\dots):\sum_k x_k^2<\infty\}$ of sequences with finite sum of squares, the length $\|x\|=(\sum_k x_k^2)^{1/2}$, and the sequence e_n whose n-th component alone is 1.

Theorem Known If a measure on ℓ^2 is unchanged by translation and gives finite weight to every open ball, it is 0.

Proof. If $n\ /=m$ then $\|e_n-e_m\|=\sqrt{2}>1$, so the balls $B(e_n,\frac{1}{2})$ are mutually disjoint, and all lie in $B(0,\frac{3}{2})$.

$$\sum_{n=1}^{\infty}\mu\big(B(e_n,\tfrac{1}{2})\big)\ \le\ \mu\big(B(0,\tfrac{3}{2})\big)\ <\ \infty \quad (2)$$

Each term on the left is the same value c by translation, so $c = 0$. ℓ^2 can be covered by countably many balls of radius $\frac{1}{2}$ centred at sequences with rational components that are 0 except for finitely many, so the total weight is 0 too. ∎ If $\mathbb{R}^n$, only up to 3^n balls fit (the ratio of volumes), and there is no contradiction.

03

Gaussians in high dimensions live on a shell — the variance must decrease

The density of the standard Gaussian $N(0, I_n)$ on $\mathbb{R}^n$ (1-08 (8)) is largest at the origin, but most of the weight lies on a thin shell far away. $|x|^2 = \sum_{k \le n} x_k^2$ is a sum of n independent terms, so by the law of large numbers (1-08 (12)), $|x|^2/n \to 1$.

$$E|x|^2 = n, \qquad \operatorname{Var}|x|^2 = 2n, \qquad |x| = \sqrt{n} + O(1) \quad (3)$$

$E|x|/\sqrt{n} = 0.975,\ 0.9975,\ 0.99975$ at $n = 10,\ 100,\ 1000$; the standard deviation of $|x|$ is $0.698,\ 0.706,\ 0.707$, and the thickness of the shell settles at $1/\sqrt{2}$ regardless of n Computed . Letting $n \to \infty$ with variance 1 gives $|x| \to \infty$, and the limit is not in ℓ^2. So we reduce the variance σ_k^2. For a sequence $\xi_1, \xi_2, \dots$ of independent $N(0,1)$,

$$x = \sum_{k=1}^{\infty} \sigma_k \xi_k e_k, \qquad E\|x\|^2 = \sum_{k=1}^{\infty} \sigma_k^2 \quad (4)$$

the finiteness of the sum on the right is necessary and sufficient for x to be a Gaussian random variable with values in ℓ^2 (Hairer Exercise 4.7) Known . **Example:** if $\sigma_k = 1/k$, then $E\|x\|^2 = \pi^2/6$. A sequence with variance 1 can also be built, on the space $\mathbb{R}^\infty$ of all sequences. What is used is **Kolmogorov's extension theorem** (the theorem that if the distributions of finitely many at a time are given consistently, the distribution of the infinite sequence is fixed uniquely; Durrett Theorem 2.1.21) Known , and this is the seed of the white noise of § 07.

04

Gaussian measures in infinite dimensions — every one-dimensional shadow is Gaussian

Since densities cannot be used, Gaussian measures are defined by their "shadows". A probability measure μ on a space E (ℓ^2 or a space of continuous functions) is a **Gaussian measure** if, for every continuous linear map $\ell : E \to \mathbb{R}$, $\ell(x)$ follows a one-dimensional Gaussian distribution (Hairer Definition 4.4) Known . With mean 0, the covariance $C(\ell, \ell') = E[\ell(x)\ell'(x)]$ determines μ (the form of 1-08 (8)).

$$\int_E e^{\,i\ell(x)}\,d\mu(x) = \exp\Big(-\frac{1}{2}\,C(\ell,\ell)\Big) \quad (5)$$

On ℓ^2 it can be written with $C(\ell,\ell') = \langle \ell, C\ell'\rangle$ and the **covariance operator** C; for (4), $Ce_k = \sigma_k^2 e_k$. If a symmetric positive C has finite trace (sum of eigenvalues), there is a Gaussian measure with it as covariance (Hairer Proposition 4.17) Known . **Example:** $\ell(x) = \sum_k a_k x_k$ is Gaussian with variance $\sum_k a_k^2\sigma_k^2$. The problem statement gives the free field as the Gaussian measure with mean 0 and covariance $(-\Delta + m_0^2)^{-1}$ on the space of distributions $\mathcal{S}'(\mathbb{R}^d)$ Known . The two factors of (1) have been replaced by a single piece of information, the covariance.

05

The Cameron–Martin theorem — the set of directions in which one may shift has probability 0

Two probability measures are **equivalent** if they have the same null sets (sets of weight 0). Conversely, when a set of probability 1 for one has probability 0 for the other, the two are said to be **mutually singular**. A finite-dimensional Gaussian measure acquires a density factor when shifted (in one dimension $N(h,1)$ is $e^{hx-h^2/2}$ times $N(0,1)$). In general a function D with $\mu_h(A) = \int_A D(x)\,d\mu(x)$ for every set A is called the **relative density** and written $d\mu_h/d\mu$. For μ_h, obtained by shifting μ by h in the first n coordinates of (4), we have the following.

$$\frac{d\mu_h}{d\mu}(x) = \exp\Big(\sum_{k\le n}\frac{h_k x_k}{\sigma_k^2} - \frac{1}{2}\|h\|_H^2\Big), \qquad \|h\|_H^2 = \sum_{k\le n}\frac{h_k^2}{\sigma_k^2} \quad (6)$$

$\|h\|_H$ is "the length of the shift measured by the fluctuation in that direction". The degree to which the two can be told apart (the total variation distance) reduces to a one-dimensional problem in the direction of $C^{-1}h$, and is exactly the following formula (Φ is 1-08 (13)).

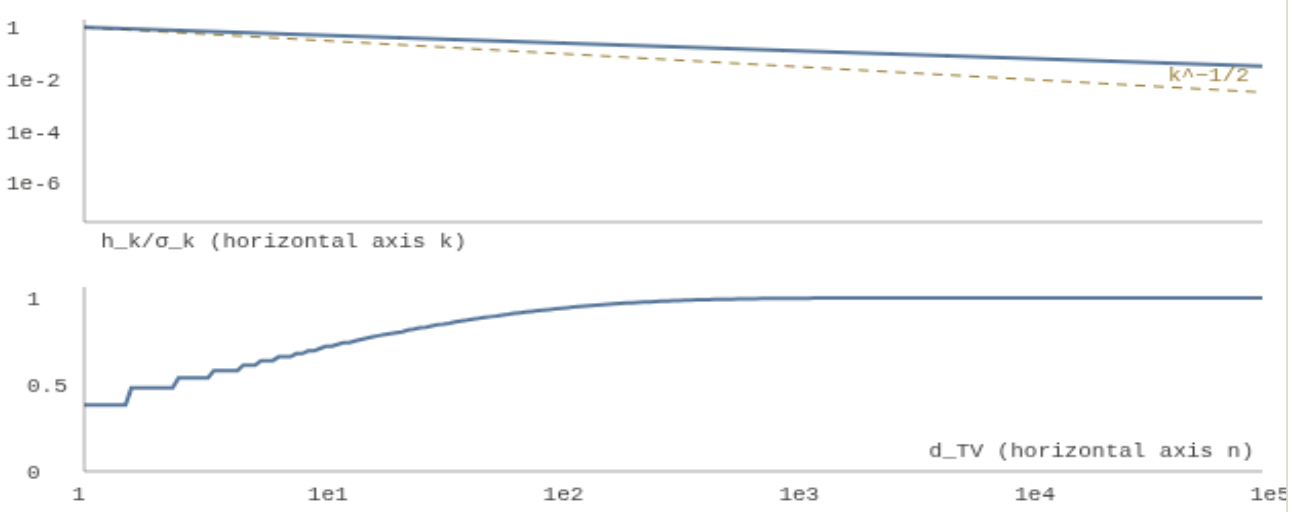
$$d_{\mathrm{TV}}(\mu,\mu_h) = \sup_A\big|\mu(A)-\mu_h(A)\big| = 2\,\Phi\big(\tfrac{1}{2}\|h\|_H\big) - 1 \quad (7)$$

(7) was checked by numerical integration and with samples in 40 dimensions Computed . As $n\to\infty$, if $\|h\|_H$ is bounded then $d_{\mathrm{TV}} < 1$; if it diverges then $d_{\mathrm{TV}}\to 1$, and in the latter case the two measures become mutually singular.

The Cameron–Martin theorem Known μ_h is equivalent to μ only when $\|h\|_H < \infty$, and otherwise they are mutually singular (Hairer Theorem 4.44 and its proof; the original paper is Cameron–Martin 1944). The set H of these h is the **Cameron–Martin space**. Yet for samples of (4), $\|x\|_H^2 = \sum_k \xi_k^2 = +\infty$ (the law of large numbers), and in infinite

dimensions $\mu(H) = 0$ (Hairer Proposition 4.45) Known . Samples fall, with probability 1, outside the set H of directions in which one may shift.

Figure 1 Shifting the measure of $\sigma_k = 1/k$ by $h_k = k^{-a}$. Moving a changes whether the ratio $h_k/\sigma_k = k^{1-a}$ above (log–log) decreases faster than the dashed line $k^{-1/2}$, and whether d_{TV} below ((7) in the first n coordinates) sticks to 1 or stays at the dashed limit



a = 1.30, at n = 10^5 ‖h‖² = 248.0、d_TV = 1.0000, limit ∞、1

Table 1 $\|h\|_H^2 = \sum_{k \le n} k^{2-2a}$ and (7) for $\sigma_k = 1/k$, $h_k = k^{-a}$ Computed

a	$\|h\|_H^2$ $(n = 10^5)$	limit	d_{TV} $(n = 10^5)$	limit
1	10^5	∞	1.0000	1
1.25	631	∞	1.0000	1
1.5	12.09	∞	0.9179	1
1.75	2.606	2.612	0.5804	0.5810
2	1.645	$\pi^2/6$	0.4787	0.4787

Computed The numbers of Figure 1 and Table 1 were computed on this machine (the figure is drawn on the spot from (7) and partial sums, agreeing with the table to 4 digits). If JavaScript is disabled, they can be read from Table 1.

At the borderline $a = 3/2$, $\|h\|_H^2 \approx \ln n + 0.577$ grows only logarithmically, and even at $n = 10^{20}$, $d_{\mathrm{TV}} = 0.9994$ Computed .

06

Brownian motion — the path settles, the energy diverges

The Gaussian measure with values in continuous functions on $[0, 1]$, mean 0 and covariance $E[B(s)B(t)] = \min(s, t)$ is the **Wiener measure**, and its sample B is **Brownian motion** (said to begin with Wiener 1923) Known . The paths are Hölder continuous with any exponent below $1/2$ (if $\gamma < 1/2$, then with a constant C depending on the path, $|B(t) - B(s)| \le C|t - s|^\gamma$ for all s, t), and with probability 1 they are

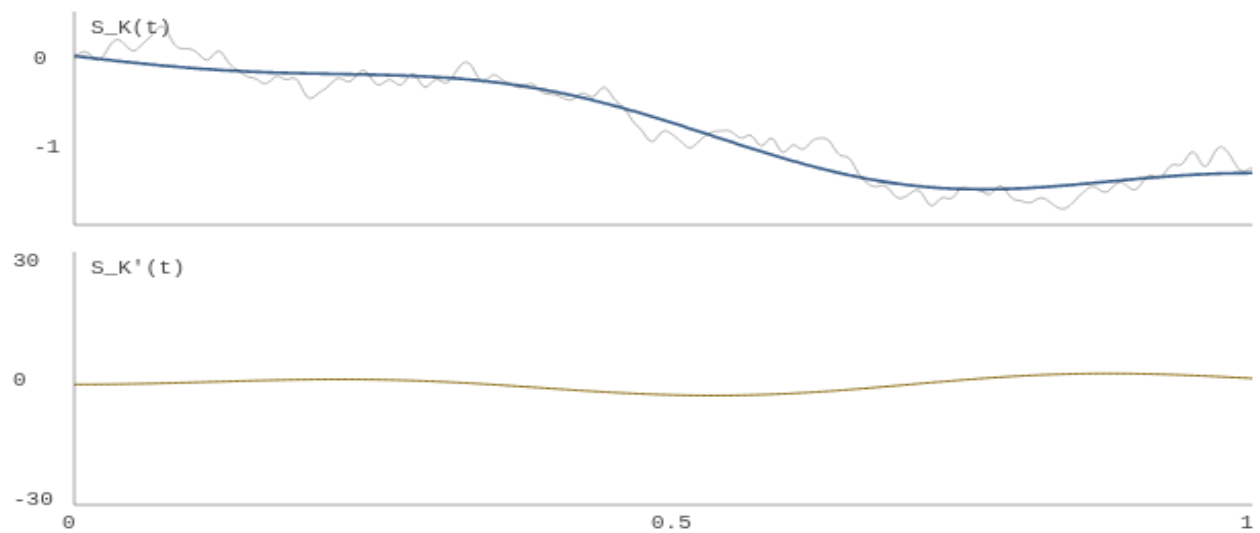
differentiable at no point (Durrett Theorems 7.1.5, 7.1.6) Known . Durrett gives the former as a result of Wiener 1923 and the latter as a discovery of Paley–Wiener–Zygmund 1933.

To put it in the form (4), we look for the eigenfunctions of the covariance operator $(Cf)(s) = \int_0^1 \min(s,t) f(t)\,dt$. $g = Cf$ satisfies $g'' = -f$, $g(0) = 0$, $g'(1) = 0$, so the eigenvalue problem becomes the ordinary differential equation $-g'' = \lambda^{-1} g$, and from $\cos\omega = 0$ the following comes out.

$$B(t) = \sum_{k=1}^{\infty} \xi_k \frac{\sqrt{2}\,\sin(\omega_k t)}{\omega_k}, \qquad \omega_k = \left(k - \frac{1}{2}\right)\pi \tag{8}$$

$\sigma_k = 1/\omega_k \approx 1/(\pi k)$ decreases at the same rate as the example of § 05, and the covariance with 4000 terms agrees with $\min(s,t)$ to within 5.1×10^{-5} Computed . The Cameron–Martin space is the set of functions with $h(0) = 0$, $\|h\|_H^2 = \int_0^1 \dot{h}^2\,dt < \infty$ (Hairer Exercise 4.29) Known , and this $\|h\|_H^2$ is the content of the exponent $m = 0$ in the $d = 1$ of (1). For the partial sum S_K of K terms, $\int_0^1 \dot{S}_K^2\,dt = \sum_{k \le K} \xi_k^2 \approx K$: the path of Figure 2 settles, yet this "energy" keeps growing.

Figure 2 Partial sums of (8). Moving K (the number of terms), S_K above settles (the faint line is $K = 128$), and $\dot{S}_K$ below (an approximation of white noise, § 07) grows in amplitude roughly as $\sqrt{K}$



K = 4 ∫S_K'² = Σξ_k² = 4.62 (÷K 1.154) max|S_K − S_128| = 0.442

Computed The numbers of Figure 2 were computed on this machine (partial sums with 128 fixed random numbers ξ_k). At $K = 1, 16, 64, 128$, $\sum \xi_k^2 = 2.13,\ 13.4,\ 56.5,\ 132.8$, $\max_t |S_K - S_{128}| = 0.589,\ 0.202,\ 0.090,\ 0$. If JavaScript is disabled, it can be read from these numbers and Table 2.

Seen on a lattice — quadratic variation

The same thing is seen when cut into a lattice. The sum of squares of the increments over N equal parts (the **quadratic variation**) approaches 1 (Durrett Exercise 7.1.6)

Known , while for smooth functions it approaches 0, so the discretised energy $\sum(\Delta B)^2/\Delta t$ diverges in proportion to N (Table 2). The $e^{-\frac{1}{2}\int \dot{x}^2 dt}\prod_t dx(t)$ of the path integral (2-07) is a formal way of writing this measure. To be exact, neither the exponent nor $\prod_t dx(t)$ has a meaning on its own (§ 01); as a pair they denote the Wiener measure.

Table 2 One path made with 2^{20} points, viewed divided into N equal parts Computed

N	$\sum(\Delta B)^2$ (fluctuation $\sqrt{2/N}$)	$\sum(\Delta B)^2/\Delta t$	$\sum(\Delta h)^2$ of $h = t^2$
16	0.689 (0.35)	11.0	0.083
4096	0.979 (0.02)	4010	3.3×10^{-4}
1048576	0.9999 (0.001)	1.05×10^{6}	1.3×10^{-6}

07

White noise — no value at a point

The formal derivative $\xi = \dot{B}$ is **white noise**. It has no pointwise values (the lower panel of Figure 2), so it is defined paired with functions. White noise is a Gaussian random variable $\xi(f)$ for each square-integrable function $f \in L^2[0,1]$, linear in f and satisfying the following (Hairer Example 4.63) Known .

$$E\big[\xi(f)\,\xi(g)\big] = \int_0^1 f(t)\,g(t)\,dt \qquad \big(\text{formally, } E[\xi(s)\xi(t)] = \delta(s-t)\big) \tag{9}$$

Example: if $\mathbf{1}_{[0,t]}$ is the function equal to 1 on $[0,t]$ and 0 outside, the covariance of $\xi(\mathbf{1}_{[0,s]})$ and $\xi(\mathbf{1}_{[0,t]})$ is $\int_0^1 \mathbf{1}_{[0,s]}(u)\,\mathbf{1}_{[0,t]}(u)\,du = \min(s,t)$, and this is Brownian motion (Hairer Exercise 4.64). The variance of the mean $\xi(\mathbf{1}_{[t,t+\varepsilon]})/\varepsilon$ over an interval of width ε is $1/\varepsilon$; for a box of side ε in d dimensions it is ε^{-d}, and it diverges when shrunk to a point.

White noise is the Gaussian whose covariance is the identity map (all $\sigma_k = 1$), and as in § 03 it is not in L^2. To measure, dimension by dimension, how far it misses, we now move the stage from $[0,1]$ to the d-dimensional **torus** $\mathbb{T}^d = [0,2\pi)^d$ (a box with opposite ends joined) (the integral of (9) is replaced by the integral over $\mathbb{T}^d$). A function on $\mathbb{T}^d$ is written $f(x) = \sum_{k\in\mathbb{Z}d} c_k\, e^{ik\cdot x}$, and c_k are called the **Fourier coefficients**. $-\Delta = -\sum_i \partial_i^2$ multiplies $e^{ik\cdot x}$ by $|k|^2$, so on coefficients it is the operation turning c_k into $|k|^2 c_k$. The Fourier coefficients of white noise are independent except for the pairs related by $c_{-k} = c_k$, and $E|c_k|^2$ is the same constant for all.

Let H^{-s} be the space of formal series $\sum_k c_k e^{ik\cdot x}$ given the length $\big(\sum_k (1+|k|^2)^{-s}|c_k|^2\big)^{1/2}$ (the larger s, the more lightly high frequencies are counted). For white noise the mean of the square of this length is a constant times $\sum_k (1+|k|^2)^{-s}$, and the shell of

radius R holds about R^{d-1} values of k, so it is finite when $s > d/2$. In fact, exactly then can white noise be realised in H^{-s} (Hairer Exercise 4.65) Known . Things that only give values when paired with functions are **distributions**; the general theory is in 1-11, and stochastic quantisation, which shakes a field with white noise, is in 7-09.

08

The free field and dimension — the variance at a point diverges

The free field of (1), considered on the torus, is white noise smoothed by multiplying its Fourier coefficients c_k by $(|k|^2 + m^2)^{-1/2}$. Writing this operation $(-\Delta + m^2)^{-1/2}$, the covariance is $(-\Delta + m^2)^{-1}$ (the form of the problem statement in § 04), and the variance of the coefficients is proportional to $1/(|k|^2 + m^2)$. By the count of § 07, the condition to lie in H^{-s} becomes $s > d/2 - 1$: one order smoother than white noise, but for $d \geq 2$ it is not a function. The free field on a bounded region can be made as samples in L^2 if $d = 1$, and cannot for $d = 2$ (the remark after Sheffield Proposition 2.7) Known .

On the lattice $a\mathbb{Z}^d$ with lattice spacing a, we replace $-\Delta$ by differences: the $(\Delta x)_j = x_{j+1} - 2x_j + x_{j-1}$ of 1-08 (10) added over coordinate directions and divided by a^2. $e^{ik\cdot x}$ are eigenfunctions of this difference Laplacian, and from $e^{ik_j a} - 2 + e^{-ik_j a} = -4\sin^2 \frac{k_j a}{2}$ the eigenvalues are $-\sum_i \frac{4}{a^2}\sin^2\frac{k_i a}{2}$. Hence the variance at a point is the following integral (k runs over the box that the lattice can distinguish).

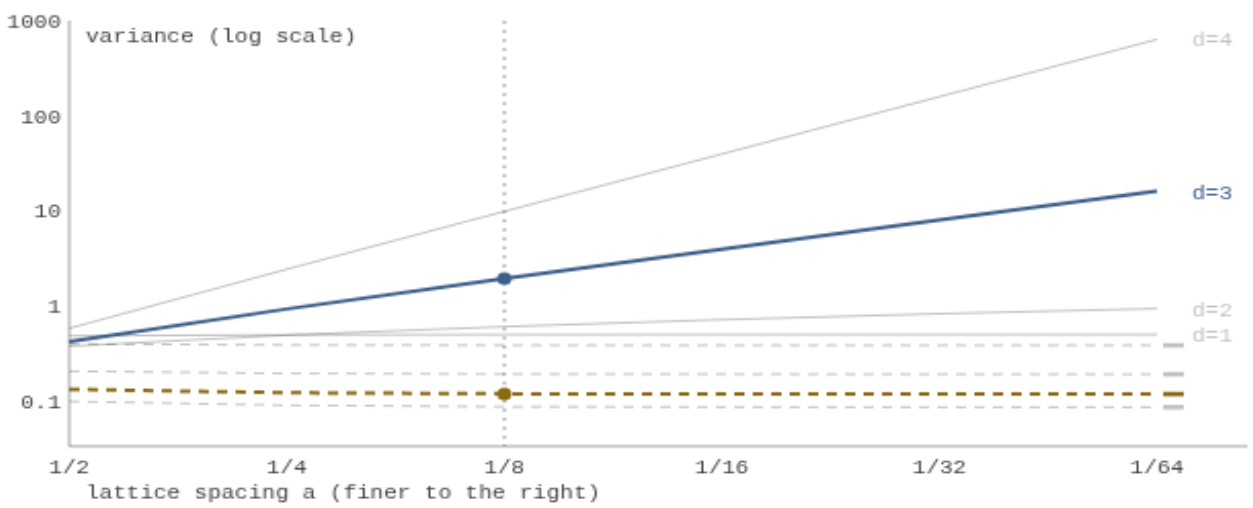
$$G_a(0) = \int_{[-\pi/a,\,\pi/a]^d} \frac{d^d k}{(2\pi)^d}\, \frac{1}{m^2 + \sum_{i=1}^{d} \frac{4}{a^2}\sin^2\frac{k_i a}{2}} \qquad (10)$$

For $a \to 0$ the contribution of large $|k|$ is about $\int^{1/a} k^{d-3}\,dk$, and it diverges as $d \geq 2$. Instead of the value at a point, look at the value $\varphi(\rho_\ell)$ paired with a Gaussian kernel $\rho_\ell(x) = (2\pi\ell^2)^{-d/2} e^{-|x|^2/(2\ell^2)}$ of standard deviation ℓ (formally $\int \varphi(x)\rho_\ell(x)\,dx$, the **smoothed value**); its variance is

$$\operatorname{Var}\varphi(\rho_\ell) = \int_{\mathbb{R}^d} \frac{d^d k}{(2\pi)^d}\, \frac{e^{-\ell^2|k|^2}}{|k|^2 + m^2}$$

(on the lattice, k is restricted to the same box as in (10) and the denominator replaced by that of (10)). The factor $e^{-\ell^2|k|^2}$ suppresses large $|k|$, so it is finite for every d.

Figure 3 Refining the lattice (a from $1/2$ to $1/64$ with the slider), the variance at a point $G_a(0)$ (solid line) keeps growing as $d \geq 2$, while the variance of the smoothed value (dashed) settles at the value marked on the right edge. Pressing a dimension button darkens the two lines for that d (vertical axis logarithmic, $m = 1$, $\ell = 1/4$)



a = 1/8, d = 3, variance at a point G_a(0) = 1.941 (a·G = 0.2426) , variance of the smoothed value 0.1193 (as a→0, 0.1183)

Table 3 The variance at a point $G_a(0)$ of the lattice free field ($m = 1$), and the variance of the value $\varphi(\rho\ell)$ smoothed by a Gaussian kernel of standard deviation $\ell = 1/4$ Computed . This is the authoritative table in this book for the dependence of the variance at a point on d (2-07 § 09 and 2-02 cite it)

a	$d = 1$	$d = 2$	$d = 3$	$d = 4$
1	0.4472	0.2540	0.1705	0.1271
1/2	0.4851	0.3769	0.4223	0.5816
1/4	0.4961	0.4932	0.9289	2.431
1/8	0.4990	0.6057	1.941	9.859
1/16	0.4998	0.7168	3.963	39.60
1/32	0.4999	0.8273	8.007	158.6
1/64	0.5000	0.9377	16.10	634.5
Growth ($a \to 0$)	→ 1/2	$\approx \frac{1}{2\pi} \log \frac{1}{a} + 0.276$	$\approx 0.2527/a$	$\approx 0.1549/a^2$
Smoothed value ($a \to 0$)	0.3852	0.1912	0.1183	0.0861

(10) has been checked against exact values for $d = 1$ and against direct sums on lattices for $d = 2, 3$. $d = 1$ converges to $1/(2m)$; for $d = 2$ the difference from $\frac{1}{2\pi} \ln \frac{1}{a}$ settles at 0.276; for $d = 3$, $a\,G_a(0)$ approaches about 0.253 (0.2515 at $a = 1/64$); and for $d = 4$, $a^2 G_a(0)$ approaches 0.155 Computed . These two destinations are the values at the origin of the lattice Green's functions for $\mathbb{Z}^3$, $\mathbb{Z}^4$, and come out as 0.2527 and 0.1549 from a different one-dimensional integral ($d = 3$ is half of Watson's integral) Computed . The variance of the smoothed value converges to a finite value for every d, and agrees with the limit to the third decimal place at $a = 1/32$. The value of the field is "infinite at a point, finite once smoothed". A figure setting this quantity, whose borderline is 2, beside quantities that switch at four is the dimension slider of 1-12 § 10.

Computed The numbers of Figure 3 and Table 3 were computed on this machine (the values for $a = 2^{-p}$, $p = 1, \ldots, 6$ are embedded in the figure as a table; Table 3 adds to them the row for $a = 1$ and the row of growth). If JavaScript is disabled, they can be read from Table 3.

09

Products of fields — without values at points, multiplication is in trouble

As in § 08, for $d \geq 2$ the mean of $\varphi(x)^2$ is infinite. So terms like φ^4 need treatment by subtraction (renormalisation, 4-04, 7-03), and products of distributions cannot in general be defined (1-11). The four-dimensional Yang–Mills field A is expected to be, at short distances, as rough as the four-dimensional free field (asymptotic freedom, 4-06) Physics . To be exact, the expectation is that correlation functions at short distances approach those of the free field up to logarithmic corrections. The problem statement explains asymptotic freedom as 'at short distances the field displays quantum behavior very similar to its classical behavior' (§ 1), and requires that the correlation functions of the theory to be constructed 'should agree at short distances with the predictions of asymptotic freedom' (§ 4) Known . The $A \wedge A$ of the field strength $F = dA + A \wedge A$ (1-07) is a product of such distributions (7-10).

10

Detour: mutually singular measures — scale, mass, interaction

A section you may skip.

Two Gaussian measures are either equivalent (§ 05) or mutually singular, with nothing in between (Feldman 1958, Hájek 1958; in textbooks it is said to be found in Bogachev Theorem 2.7.2 and elsewhere) Known . In finite dimensions, Gaussian measures with positive definite covariance are always equivalent. In infinite dimensions, as in the following examples, a slight difference makes them mutually singular.

Scale. Multiplying the covariance by c^2 makes $\frac{1}{N}\sum_{k\leq N} x_k^2/\sigma_k^2$ converge not to 1 but to c^2 (the law of large numbers). c can be read from a single sample, so if $c \neq \pm 1$ they are mutually singular (Hairer Proposition 4.43) Known .

Mass. Comparing free fields of masses m_1, m_2 (both positive) on the torus, the ratio of variances of the Fourier coefficients is $1+\varepsilon_k$, $\varepsilon_k = (m_1^2 - m_2^2)/(|k|^2 + m_2^2)$. To compare two product measures of independent coordinates we use Kakutani's criterion. Let p_k, q_k be the two densities of the k-th coordinate (here both positive Gaussian densities). If the infinite product of the overlaps $\int \sqrt{p_k(x) q_k(x)}\, dx$ for each dimension is positive the two are equivalent, and if 0 they are mutually singular (Kakutani 1948) Known . For a real-valued field the real and imaginary parts of the Fourier coefficients are independent coordinates, but the ratio of variances is $1+\varepsilon_k$ for both, so the condition below does not change. The overlap of two Gaussians with ratio of variances $r^2 = 1+\varepsilon$ is $\sqrt{2r/(1+r^2)} = 1 - \varepsilon^2/16 + O(\varepsilon^3)$, and $\varepsilon_k \to 0$, so the product being positive is equivalent to $\sum_k \varepsilon_k^2 < \infty$.

$$\sum_{k\in\mathbb{Z}^d} \varepsilon_k^2 = (m_1^2 - m_2^2)^2 \sum_{k\in\mathbb{Z}^d} \frac{1}{(|k|^2 + m_2^2)^2} < \infty \quad \Longleftrightarrow \quad d \leq 3 \qquad (11)$$

Looking at the partial sums $\sum(|k|^2+1)^{-2}$ of $|k| \le R$ for $m_2 = 1$ with $R = 16, 32, 64, 128$: for $d = 3$ they are 9.21, 9.61, 9.80, 9.90, the increments for each doubling of R halve as 0.39, 0.20, 0.10, and they converge near 10.0 (the remainder is about $4\pi/R$; a separate integral gives 9.998 for the whole sum). For $d = 4$ the increments do not halve, and grow by $2\pi^2 \ln 2 \approx 13.7$ for each doubling of R Computed . Hence, if $m_1 \neq m_2$, the measures of two free fields of different masses on the torus are equivalent if $d \le 3$ and mutually singular if $d \ge 4$ Known (Schröfl–Floerchinger 2025 § 4.1; the cases of Dirichlet, Neumann and periodic boundaries on bounded regions). (11) is the same condition as the Feldman–Hájek criterion (with C_i the covariance of mass m_i, $C_1^{-1/2} C_2 C_1^{-1/2} - I$ is a Hilbert–Schmidt operator; here, the sum of squares of the eigenvalues ε_k is finite).

Interaction. The construction that multiplies a Gaussian measure by a density e^{-V} breaks down when V cannot be defined for almost every sample. Measures made as limits in general have no density (Sheffield § 3.3), and the φ^4 measure on the three-dimensional torus is mutually singular with the free field (Barashkov–Gubinelli 2020) Known . Constructive field theory makes measures as limits of cut-off measures (7-01, 7-05).

What this chapter can say

Said	Not said
Known The theorem of § 02 (proved in the text). The other theorems are cited from the literature.	Proofs of the theorems of Kolmogorov, Cameron–Martin, Kakutani and Feldman–Hájek; the general theory of constructing measures on spaces of distributions.
Known The criterion by (11) for free fields of different masses on the torus ($m_1 \neq m_2$) (the same conclusion as Schröfl–Floerchinger 2025 § 4.1).	No machine checking. The whole of $\mathbb{R}^d$ is not treated.
Computed Tables 1–3, Figures 1–3, the numbers and checks in the text.	The limiting values (0.253, 0.155, 10.0 and so on) are readings from finite a, R and numerical integration, without rigorous error estimates.
Physics The expectation that correlation functions of the four-dimensional Yang–Mills field approach those of the free field at short distances (up to logarithmic corrections).	Not a theorem. The existence of interacting measures in four dimensions (including Yang–Mills) is unsolved, and this site does not show it either (the construction published in 2026-11 did not make a measure either; 6-04).

Sources and reproduction

Item	Kind	Source / tools
Definitions and theorems of § 03– § 10 (numbers as in the text)	Known	M. Hairer, "An Introduction to Stochastic PDEs", arXiv:0907.4178 (v2, 2023) § 4. Text checked.
Theorem 2.1.21 (Kolmogorov's extension theorem), Theorems 7.1.5, 7.1.6 and the surrounding attributions (Wiener 1923, Paley–Wiener–Zygmund 1933), Exercise 7.1.6	Known	R. Durrett, *Probability: Theory and Examples*, 5th ed., Cambridge Univ. Press (2019). Text checked in the author's public version.
Proposition 2.7 and the remark after it, § 3.3	Known	S. Sheffield, "Gaussian free fields for mathematicians", Probab. Theory Relat. Fields 139 (2007) 521–541, arXiv:math/0312099. Text checked.
Original papers (Wiener measure, non-differentiability, Cameron–Martin, Kakutani, the dichotomy of Gaussian measures)	Known	N. Wiener, J. Math. Phys. 2 (1923) 131–174. R. E. A. C. Paley, N. Wiener, A. Zygmund, Math. Z. 37 (1933) 647–668 (bibliographic data from Durrett's bibliography). R. H. Cameron, W. T. Martin, Ann. of Math. 45 (1944) 386. S. Kakutani, Ann. of Math. 49 (1948) 214–224. J. Feldman, Pacific J. Math. 8 (1958) 699–708 (text checked: the alternative of equivalence or orthogonality for measures of Gaussian processes). J. Hájek, Czechoslovak Math. J. 8 (1958) 610–618 (in Russian; bibliographic data checked in DML-CZ, text bibliographic data only). The others bibliographic data only (Kakutani's statement checked in secondary sources).
The dichotomy of Gaussian measures (textbooks)	Known	V. I. Bogachev, *Gaussian Measures*, AMS (1998) Theorem 2.7.2. G. Da Prato, J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge Univ. Press (2014) Theorem 2.25. All bibliographic data only (mentions checked in secondary sources).
The criterion for free fields of different masses in § 10 (bounded region; equivalent for $d < 4$, singular for $d \geq 4$)	Known	M. Schröfl, S. Floerchinger, "Relative entropy and mutual information in Gaussian statistical field theory", Ann. Henri Poincaré 26 (2025) 3233–3319, arXiv:2307.15548. Text checked (§ 4.1, Theorem 3; periodic boundaries are the cube case of the same section).
The values at the origin of the lattice Green's functions for $\mathbb{Z}^3$ (§ 08)	Known, Computed	G. N. Watson, "Three triple integrals", Quart. J. Math. Oxford 10 (1939) 266–276. Bibliographic data only. The value 0.2527 was computed on this machine from a different one-dimensional integral.
The singularity of φ^4 on the three-dimensional torus	Known	N. Barashkov, M. Gubinelli, "The Φ^4_3 measure via Girsanov's theorem", arXiv:2004.01513 (2020). Abstract checked.

Item	Kind	Source / tools
Quotations from the problem statement in §01, §04, §09	Known	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory" (the problem description of the Clay Mathematics Institute), §1, §2, §4, §6. Text checked.
Tables, figures, numbers in the text	Computed	Python 3.12 on this machine (NumPy 1.26, SciPy 1.11): dims.py, cameron_martin.py, brownian.py, gff.py, watson.py, fig3.py. The functions of the figures were checked against the tables with fig_check.js (node).

Next chapter: 1-10 Hilbert spaces and self-adjoint operators — putting the ℓ^2, covariance operators and eigenfunction expansions of this chapter in order in the language of spectra.

Hilbert spaces and self-adjoint operators — measuring "the lowest" and "the next" in infinite dimensions

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

What this chapter shows — Spaces in which functions are treated as vectors whose length can be measured, and unbounded operators such as differentiation. That the spectrum is wider than the set of eigenvalues, and that boundary conditions are part of the operator. That the variational principle measuring the lowest value and the gap is the same thing as the best constant of the Poincaré inequality.

Prerequisite chapters — 1-01 (inner products, Hermitian matrices, diagonalisation). Sections 02, 05, 08 and 09 touch on 1-08 and 1-09 (sets of length 0, Brownian motion, Gaussian distributions).

Said first — We introduce the entrance to functional analysis, and omit the details of domains and the proofs of theorems. The Hamiltonians of quantum field theory are much harder to handle (4-09), and this chapter shows nothing about the Yang–Mills mass gap.

Order of this chapter

1. Why infinite dimensions — the lower eigenvalues stay, the upper ones run away (Figure 1, Table 1)
2. Hilbert spaces — limits stay inside the space
3. Operators and self-adjointness — boundary conditions are part of the operator
4. The spectrum — some values have no eigenvalue
5. The spectral theorem — the gap decides the rate of decay
6. The variational principle — upper bounds are easy, lower bounds need other information (Table 2, box "The asymmetry of upper and lower bounds")
7. The Poincaré inequality — the best constant is the inverse of the gap

8. The weighted Poincaré inequality — $1/\kappa$ if convex, large if there are two valleys (Figure 2, Table 3)
9. With many variables — products, Gaussians, the free field, and two gaps
10. Detour: counting eigenvalues — Weyl's law and the shape of a drum

---

01

Why infinite dimensions — the lower eigenvalues stay, the upper ones run away

In finite dimensions the smallest eigenvalue and the gap to the next can always be computed (1-01). What is asked is whether the gap stays positive in the limit of more and more degrees of freedom. As an example, discretise on a lattice of N equal parts the rule assigning $-u''$ to a function u on $[0,1]$ vanishing at both ends (the same form as the Hessian matrix of 2-01 § 05).

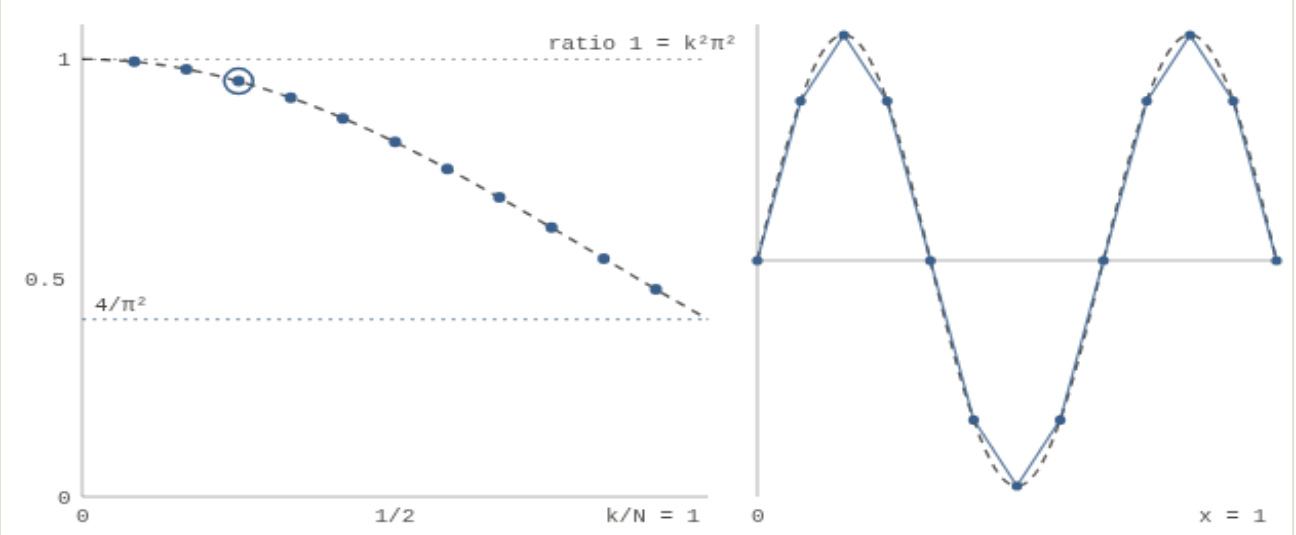
$$(A_N u)_j = N^2\,(-u_{j-1} + 2u_j - u_{j+1}), \quad u_0 = u_N = 0; \qquad \lambda_k(N) = 4N^2 \sin^2 \frac{k\pi}{2N} \quad (1 \le k$$

The eigenvectors are $u_j = \sin(k\pi j/N)$ (this can be checked by substitution). Fixing k and letting $N \to \infty$ gives $\lambda_k(N) \to k^2\pi^2$. This is an eigenvalue of $-u'' = \lambda u,\ u(0) = u(1) = 0$. The top eigenvalue grows to near $4N^2$ and has no limit.

Computed **Table 1** Eigenvalues of the matrix of (1) (the difference between a tridiagonal eigenvalue computation and (1) is at most 10^{-15} relative).

N	λ_1	λ_2	top
10	9.78870	38.19660	390.2
40	9.86453	39.39731	6390.1
160	9.86929	39.47334	102390.1
continuum	π^2 = 9.86960	$4\pi^2$ = 39.47842	(none)

Figure 1 Left: the ratios of eigenvalues of (1) λ_k(N)/k²π² (dots, k = 1 ... N − 1) plotted against t = k/N. The dots lie on the curve (sin(πt/2)/(πt/2))² (dashed) regardless of N, and increasing N brings the dot for number k (circle) towards the ratio 1. The ratio 4/π² at the right end means that the top end levels off at 4N². Right: the k-th eigenvector (dots) and sin kπx (dashed)



Computed The numbers in the figure are computed on the spot from (1) (checked against Table 1).

The ratio is fixed by $t = k/N$ alone: $\lambda_k(N)/k^2\pi^2 = \big(\sin(\pi t/2)/(\pi t/2)\big)^2$ (the dashed line on the left of Figure 1). At $k = N/2$, $8/\pi^2 \approx 0.81$; at the right end $t \to 1$, $4/\pi^2 \approx 0.41$. Fixing k and increasing N gives $t \to 0$, and the ratio approaches 1. An eigenvector with k near N is, at the lattice points, exactly the values of $\sin k\pi x$, but its oscillation is as fine as the spacing of the points, so joining the points makes it look like a different shape. In fact $\sin(k\pi j/N) = -(-1)^j \sin((N-k)\pi j/N)$: a slow oscillation multiplied by a sign that alternates point by point (visible on the right of Figure 1 by bringing k near $N - 1$). What remains in the limit is the lower eigenvalues and the gap, and the limiting object is an operator unbounded above. In the continuum limit of lattice gauge theory (5-05) too, what is read from the lattice is the lower end of the spectrum (the lightest state) Physics . In four dimensions, that the scaling limit of lattice gauge theory exists was shown by the construction published in 2026-11 (the committee's assessment is that it is proved in Lean; 6-00). The lower end of the spectrum of that limiting theory, that is, the mass gap, has not been shown (6-04).

02

Hilbert spaces — limits stay inside

To treat functions as "vectors with infinitely many components" we need an inner product (1-01 § 05) and **completeness**. Complete means that sequences whose terms come arbitrarily close to one another (Cauchy sequences) have limits inside the space. A complex vector space with these two is called a **Hilbert space**.

$$\ell^2 = \Big\{(c_1, c_2, \dots) : \sum_k |c_k|^2 < \infty\Big\}, \qquad L^2(0,1) = \Big\{f : \int_0^1 |f|^2 dx < \infty\Big\}, \qquad \langle f, g\rangle = \int_0^1$$

In L^2, functions that differ only on a set of length 0 are regarded as the same (1-08). Continuous functions alone are not complete: $\tanh(n(x-\frac{1}{2}))$ come close to one another, yet their destination is a step function. To take the destination of "a sequence making the ratio smaller and smaller" in section 06, the complete L^2 is needed Known .

$e_k = \sqrt{2}\sin k\pi x$ $(k \geq 1)$ is an orthonormal basis of $L^2(0,1)$, and for $c_k = \langle e_k, f\rangle$, $\|f\|^2 = \sum_k |c_k|^2$ (Parseval's identity) Known (1-11). Example. For $f = x(1-x)$, $c_k = 4\sqrt{2}/(k\pi)^3$ for odd k and 0 for even. $\sum_{k\ \text{odd}} 32/(k\pi)^6 = 1/30$ agrees with $\int_0^1 x^2(1-x)^2 dx = 1/30$. The ℓ^2 of 1-09 connects to spaces of functions in this way.

03

Operators and self-adjointness — domains and boundary conditions

An operator whose supremum of stretching factors $\|A\| = \sup_{\psi\neq 0}\|A\psi\|/\|\psi\|$ is finite is called bounded. $-d^2/dx^2$ multiplies e_k by $k^2\pi^2$, so it is not bounded, and there are functions whose second derivatives are not in L^2. So the subspace of functions it may be applied to, the **domain** $D(A)$, is specified as part of the operator. If $\langle \phi, A\psi\rangle = \langle A\phi, \psi\rangle$ on $D(A)$, it is **symmetric**. Self-adjointness is one step stronger: it says that the partners ϕ of this identity do not increase even when sought outside $D(A)$. To be exact, the set of ϕ for which there is a η with $\langle \phi, A\psi\rangle = \langle \eta, \psi\rangle$ for all $\psi \in D(A)$ is taken as the domain $D(A^\dagger)$ of the adjoint $A^\dagger$, and $A^\dagger\phi = \eta$ is defined. If symmetric, $D(A) \subset D(A^\dagger)$, and when $D(A^\dagger) = D(A)$ it is **self-adjoint**. The spectral theorem and time evolution e^{-itA} (2-05) presuppose self-adjointness Known . For the momentum $P = -i\,d/dx$, integration by parts gives the following.

$$\langle \phi, P\psi\rangle - \langle P\phi, \psi\rangle = -i\,\big[\,\phi\,\psi\,\big]_0^1 \quad (3)$$

(a) If $D = \{\psi(0) = \psi(1) = 0\}$, it is symmetric, but the partner ϕ needs no condition, and $P^\dagger e^{izx} = ze^{izx}$ for every complex number z, so it is not self-adjoint. (b) If $D_\theta = \{\psi(1) = e^{i\theta}\psi(0)\}$, the partner satisfies the same condition and it is self-adjoint, with eigenvalues $2\pi n + \theta$ $(n \in \mathbb{Z})$. The phase difference $q\Phi/\hbar$ from a magnetic flux Φ threading the ring plays the role of θ (the Aharonov–Bohm effect of 2-03). For $-d^2/dx^2$ too, with $u = 0$ at the ends (Dirichlet) the spectrum is $\{k^2\pi^2\}_{k\geq 1}$, and with $u' = 0$ (Neumann) it is $\{k^2\pi^2\}_{k\geq 0}$. **Boundary conditions are part of the operator**, and change the lowest value.

04

The spectrum — values for which no inverse can be made

In infinite dimensions there are cases where a map is one-to-one but its inverse is not bounded. So the **spectrum** is defined as follows.

$$\sigma(A) = \{\lambda \in \mathbb{C} : \ (A-\lambda)^{-1} \text{ does not exist as a bounded operator}\}, \qquad \|(A-\lambda)^{-1}\| = \frac{1}{\text{dis}}$$

The identity on the right holds when A is self-adjoint and $\lambda \notin \sigma(A)$, and then $\sigma(A)$ is a closed subset of the real line Known . Example. The operator M multiplying by x on $L^2(0,1)$ has no eigenvalues, since $xf = \lambda f$ implies $f = 0$. Even so, for $\lambda \in [0,1]$, $1/(x-\lambda)$ is not bounded, and $\sigma(M) = [0,1]$. A function of length 1 that is constant only on $[\lambda, \lambda + \varepsilon]$ has $\|(M-\lambda)f\| \le \varepsilon$: there are things arbitrarily close to eigenvectors (Weyl's criterion of 2-05 § 03). Discretising makes the eigenvalues lattice points, which fill up $[0,1]$ (the lower eigenvalues of section 01 stay apart).

Isolated eigenvalues of finite multiplicity are called the **discrete spectrum**, and the rest the **essential spectrum**. The mass gap is the claim that the vacuum eigenvalue 0 is isolated and the rest of the spectrum lies at $\Delta > 0$ or above (4-11).

05

The spectral theorem — the gap decides the rate of decay

A self-adjoint operator can be carried by a suitable unitary to "an operator multiplying by a real-valued function", like M (the spectral theorem) Known . It is the infinite-dimensional version of the diagonalisation of matrices (1-01), and to be exact it includes parts that cannot be written with eigenvalues alone. Functions g can be substituted, and if there is a basis of eigenvectors it takes the following form.

$$g(A)\,\psi = \sum_{k} g(\lambda_k)\,\langle e_k, \psi\rangle\, e_k, \qquad \left\| e^{-tA}\psi - \langle e_0,\psi\rangle e_0 \right\| \le e^{-\Delta t}\left\| \psi - \langle e_0,\psi\rangle e_0 \right\| \quad (5)$$

The right-hand side is for the case $A \ge 0$, with the eigenspace of 0 spanned by e_0 of length 1 and the rest of the spectrum in $[\Delta, \infty)$; it follows by using $e^{-t\lambda} \le e^{-\Delta t}$ on each component. **The gap is the slowest rate at which the components other than the lowest disappear** (for the reading in imaginary time see 2-05 § 05). Example 1. For the heat equation $\partial_t u = \partial_x^2 u$ on a rod insulated at both ends, A is the Neumann $-d^2/dx^2$, $\Delta = \pi^2$, and the deviation of the temperature from its mean is at most $e^{-\pi^2/2} \approx 0.0072$ times at $t = 0.5$. Example 2. The inverse of the Dirichlet $-d^2/dx^2$ is $(Gf)(x) = \int_0^1 G(x,y) f(y)\,dy$, $G(x,y) = \min(x,y) - xy$ ($-(Gf)'' = f$ and the zeros at both ends can be checked). Making $X(t) = B(t) - tB(1)$ from the Brownian motion B of 1-09 gives zeros

at both ends (the **Brownian bridge**), and its covariance $E[X(s)X(t)]$ is exactly $G(s,t)$. The eigenvalues of G are $(k\pi)^{-2}$, and the finiteness of the sum $\sum_k (k\pi)^{-2} = 1/6$ is the condition of 1-09 (4) for samples of the Brownian bridge to lie in L^2.

06

The variational principle — easy from above, hard from below

The lower end of the spectrum of a self-adjoint operator bounded below is the infimum of the **Rayleigh quotient**. The n-th too can be written in the form "take the maximum over a n-dimensional subspace, and the minimum over subspaces" (the min-max principle; for the eigenvalues below the essential spectrum, in increasing order counted with multiplicity) Known .

$$\inf \sigma(A) = \inf_{0 \neq \psi \in D(A)} \frac{\langle \psi, A\psi \rangle}{\langle \psi, \psi \rangle}, \qquad \lambda_n = \min_{\dim V = n} \ \max_{0 \neq \psi \in V} \frac{\langle \psi, A\psi \rangle}{\langle \psi, \psi \rangle} \tag{6}$$

Expanded in eigenvectors, the Rayleigh quotient is a weighted average of eigenvalues $\sum \lambda_k |c_k|^2 / \sum |c_k|^2$, and does not fall below λ_1. The eigenvalues of the matrix restricted to a subspace V are **upper bounds** for $\lambda_1, \lambda_2, \ldots$, and decrease as V is enlarged (the Rayleigh–Ritz method). Example. Putting $x(1-x)$ into the Dirichlet $-d^2/dx^2$ gives the quotient $(1/3)/(1/30) = 10 \geq \pi^2$. For estimates from below there is Temple's inequality. If R is the Rayleigh quotient of $\psi \in D(A)$, and we know a lower bound μ for the next eigenvalue ($R < \mu \leq \lambda_2$), the following holds Known .

$$\lambda_1 \ \geq \ R - \frac{\|A\psi\|^2/\|\psi\|^2 - R^2}{\mu - R} \tag{7}$$

If $\psi = x(1-x)$, then $A\psi = 2$, $\|A\psi\|^2/\|\psi\|^2 = 120$. Here $\lambda_2 = 4\pi^2$ is known exactly (section 03), so we can put $\mu = 39$, giving $\lambda_1 \geq 10 - 20/29 = 9.310$. The Ritz value for $39.50\ldots$ appearing below Table 2 is an upper bound for λ_2, and is no basis for μ. **The variational principle alone gives only upper bounds; holding it down from below needs other information.** In Temple's inequality it is the lower bound μ for the next level (the convexity $V'' \geq \kappa$ of section 08 is another example). This is, with finitely many degrees of freedom, the shape of why the proof of the mass gap is hard (2-05 § 04, 4-11).

Computed **Table 2** Rayleigh–Ritz values on $V_n = \{x(1-x)p(x) : \ p \text{ of degree at most } n-1 \}$ and the lower bound from (7) ($\mu = 39$). The matrices were built exactly with fractions.

n	Upper bound for λ_1	Lower bound for λ_1 (7)	Upper bound for λ_3
1	10.000000000	9.310344828	—
3	9.869749621	9.866602308	102.1303
5	9.869604435	9.869602524	89.1736
exact	9.869604401		88.82644

The antisymmetric functions added for even n improve only λ_2, and the upper bound comes down to $42,\ 39.50155,\ 39.47847$ ($4\pi^2 = 39.47842$). At $n = 5$, π^2 is caught in an interval of width 2×10^{-6} (the lower end depends on $\mu = 39 \le 4\pi^2$).

The asymmetry of upper and lower bounds — one from above, all from below

To hold a gap or the lowest eigenvalue **from above**, it is enough to exhibit **one** trial vector. Every value of the quotient (6) is an upper bound as it stands. To hold it **from below**, one must show that the quotient does not fall below it for **every** vector, and information outside the variational principle is needed (the μ of Temple's inequality, the convexity of section 08, the decomposition of the whole by the spectral theorem) Known . The Poincaré constant C_P (section 07) is the inverse of the gap, so the directions swap: one trial function gives a lower bound for C_P, and an upper bound needs every function.

The mass gap $\Delta > 0$ is a **lower bound** on the energy of states orthogonal to the vacuum (4-11 § 01). This book puts it in one sentence: **a proof of the mass gap is a proof of a lower bound.** Trial states and measurements of finitely many correlations give upper bounds for Δ, and say nothing about the gap being positive (4-11 § 02, 18-01 § 08). As a metaphor: looking for the lowest valley in the mountains, finding one valley lets you say "at most this low", but "there is no valley lower than this" needs the whole range to be examined. To be exact: evaluating the infimum of (6) with one ψ gives an upper bound, and holding the infimum itself from below needs every ψ.

07

The Poincaré inequality — the best constant is the inverse of the gap

The Poincaré inequality is "a function with small slope does not stray far from its mean" put in numbers. On the interval $[0, 1]$ it takes the following form, and the constant $1/\pi^2$ cannot be made smaller Known .

$$\int_0^1 |f - \bar{f}|^2 dx \ \le\ \frac{1}{\pi^2} \int_0^1 |f'|^2 dx, \qquad \bar{f} = \int_0^1 f\, dx \quad (8)$$

Proof. Writing $f-\bar f=\sum_{k\ge 1}a_k\sqrt2\cos k\pi x$ with the Neumann eigenfunctions $1,\ \sqrt2\cos k\pi x$, the left-hand side is $\sum|a_k|^2$ and the integral on the right is $\sum k^2\pi^2|a_k|^2\ge\pi^2\sum|a_k|^2$. Equality holds only when $f=a+b\cos\pi x$. On an interval of length L the constant is $(L/\pi)^2$, **proportional to the square of the size** (the version on a circle is (4) of 17-06).

This is the variational principle of section 06 used for "the next after 0" of the Neumann operator. In general, if for a self-adjoint $A\ge 0$ only constants satisfy $Af=0$ and the gap is Δ, then the best constant of $\operatorname{Var}(f)\le C_P\langle f,Af\rangle$ is $C_P=1/\Delta$, and by (5) the smoothing motion settles at rate $1/C_P$. The Dirichlet version $\int|f|^2\le\pi^{-2}\int|f'|^2$ ($f(0)=f(1)=0$) has the same number, but equality is at $\sin\pi x$, and the content is different. Example. For $f=x$ the left-hand side is $1/12$ and the right-hand side $1/\pi^2$, with ratio 0.82.

08

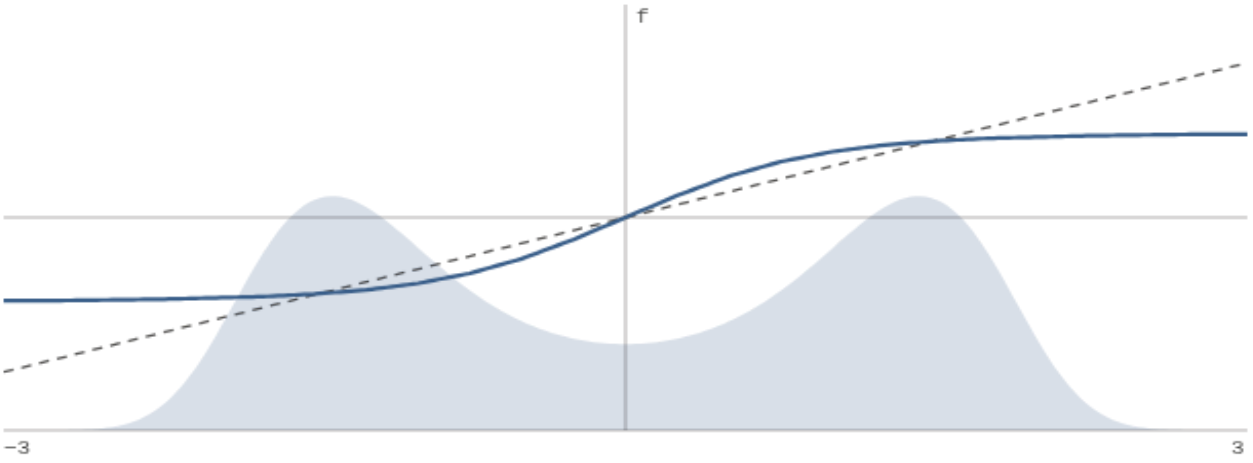
The weighted Poincaré inequality — 1/κ if convex, large if there are two valleys

What field theory needs is the form for a probability measure $\mu(dx)=e^{-V(x)}dx/Z$ (Z is the constant making the total 1). Write the inner product as $\langle f,g\rangle_\mu=\int f\,g\,d\mu$, the space of functions of finite length in this inner product as $L^2(\mu)$, and the variance as $\operatorname{Var}_\mu(f)=\int\bigl|f-\int f\,d\mu\bigr|^2d\mu$.

$$\operatorname{Var}_\mu(f)\le C_P\int|f'|^2\,d\mu;\qquad \operatorname{Var}_\mu(x)\ \le\ C_P\ \le\ \frac{1}{\kappa}\quad(V''\ge\kappa>0)\ \text{(9)}$$

The best C_P is the inverse of the gap of $Lf=-f''+V'f'$ after 0. L is self-adjoint on $L^2(\mu)$, integration by parts gives $\langle f,Lf\rangle_\mu=\int|f'|^2d\mu$, and only constants satisfy $Lf=0$ (the same form as section 07; from the side of the ground state, 2-05 § 08). The inequality on the left follows by putting in $f=x$. The one on the right is said to be a corollary of the variance inequality of Brascamp–Lieb (1976), and also follows from the logarithmic Sobolev inequality of Bakry–Émery (1985) (an inequality stronger than the Poincaré inequality) Known . In one dimension the following line suffices. Differentiating $Lf=\lambda f$, $g=f'$ gives $Lg+V''g=\lambda g$; the inner product with g gives $\int|g'|^2d\mu+\int V''g^2d\mu=\lambda\int g^2d\mu$, hence $\lambda\ge\kappa$ (assuming there is an eigenfunction and the boundary terms vanish). For the normal distribution $V=x^2/(2\sigma^2)$ all three are σ^2 (Chernoff 1981).

Figure 2 Moving c (from −2 to 4) splits the weight exp(−V_c) (shaded) into two peaks, and the f giving equality (solid line) changes from the straight line x/√Var(x) (dashed) to a step



c = 2.0, barrier c²/4 = 1.00, Var(x) = 1.6655 ≤ C_P = 2.6714 (there is no κ > 0, so there is no upper bound on the right)

Computed The numbers in the figure were computed on this machine (the best f and C_P were found with python3 and embedded; the weight e^{-V_c} is computed on the spot).

Computed **Table 3** C_P for $V_c = x^4/4 - cx^2/2$ (found by turning L into an operator of Schrödinger type and using three-point differences; the method is in the table of sources; halving the step agrees to 4 digits). At $c < 0$, $\kappa = -c$.

c	1/κ	Var(x)	C_P
−2	0.500	0.3451	0.3503
0	—	0.6760	0.7307
1	—	1.0418	1.2625
2	—	1.6655	2.6714
3	—	2.5853	7.6577
4	—	3.6707	33.580

With two valleys ($c > 0$, height of the hill between the valleys $c^2/4$), C_P grows much faster than $\mathrm{Var}(x)$, and at $c = 4$ it is 9 times as large. The best f becomes a step close to -1 in the left valley and $+1$ in the right, and its slope matters only on the hill, where there is almost no weight. **A narrowing of the weight makes C_P large.**

For how this book uses this, see 23-14 § 02 (the distance to the nearest known theorem, condition by condition).

09

With many variables — products, Gaussians, the free field, and two gaps

When there are many variables $x = (x_1, \dots, x_n)$, replace the $|f'|^2$ of (9) by the squared length of the gradient $|\nabla f|^2 = \sum_j |\partial_j f|^2$. The basics are the following two Known .

(i) The CP of a product measure of independent variables is the maximum of the CP of the factors, and does not depend on the number of variables. L becomes the sum of the L for each variable, the eigenfunctions are products $f_1(x_1)\cdots f_n(x_n)$ and the eigenvalues sums, so the value after 0 is the minimum of the gaps of the factors (assuming a basis of eigenfunctions).

(ii) The CP of the Gaussian measure with mean 0 and covariance matrix C is the largest eigenvalue of C. Taking the eigenvectors of C as orthogonal coordinates, the measure splits into a product of independent one-dimensional Gaussians whose variances are the eigenvalues of C (1-08), and $|\nabla f|^2$ does not change under a change of orthogonal coordinates. The rest uses (i) and the one-dimensional $CP = \sigma^2$ (section 08).

For the free field on the ring of 1-08 (10), $C = (-\Delta_{\text{lat}} + m^2)^{-1}$. The eigenvalues of $-\Delta_{\text{lat}}$ are at least 0, and 0 is the constant direction (the direction shifting every x_j by the same amount), so the smallest eigenvalue of C^{-1} is m^2, hence $CP = 1/m^2$. Combined with the correlation length $\xi \approx 1/m$ (1-08 (11)) this gives $CP \approx \xi^2$ (this CP is measured with the gradient in the field variables x_j. In lattice gauge theory, taking the angles of the links as variables changes the constant factor, so we always state which is used. 23-01).

Here we distinguish two gaps. For the free field, the inverse m^2 of CP is the gap of L (the rate at which e^{-tL} removes the deviation from the mean in (5)). On the other hand, the gap of the Hamiltonian when 1-08 (10) is read as an imaginary-time path integral is $1/\xi$ in units of the lattice spacing (1-08 (11); $\approx m$ when m is small). The squared relation $CP \approx \xi^2$ is a free-field computation; whether it holds in general in interacting theories is the question of 23-01, and has not been proved (the same form on a group is in 5-11).

10

Detour: counting eigenvalues — Weyl's law and the shape of a drum

A section you may skip. Kac (1966) relates that at the end of October 1910, Lorentz, invited to give the Wolfskehl lectures in Göttingen, posed at the end of his fourth lecture the problem that "the number of high-frequency standing waves in a cavity should be proportional to its volume regardless of its shape", and Weyl, who was in the audience, proved it in less than two years. For the Dirichlet $-\Delta$ on a planar region Ω it takes the following form Known .

$$N(\lambda) = \#\{k:\ \lambda_k \le \lambda\} \ \sim\ \frac{|\Omega|}{4\pi}\lambda \quad (\lambda\to\infty); \qquad d \text{ dimensions give } N(\lambda) \propto |\Omega|\,\lambda^{d/2} \quad (10)$$

Computed For the unit square (eigenvalues $\pi^2(j^2+k^2)$), $N = 79269$ at $\lambda = 10^6$. Adding to the leading term 79577.5 the correction $-\sqrt{\lambda}/\pi$ for the perimeter 4 gives 79259.2. Kac discussed that

the area and the perimeter can be "heard" from the sequence of eigenvalues (to be exact, are determined by the sequence of eigenvalues alone), and asked about the shape itself. The answer is no: Gordon, Webb and Wolpert (1992) constructed planar regions of different shapes with the same sequence of eigenvalues. Reading λ as the squared momentum Λ^2, the number of modes up to Λ is proportional to Λ^4 in four dimensions. This abundance of short-distance degrees of freedom is one of the starting points for thinking about divergences (4-04, 1-12).

What this chapter can say

Content	Status
The eigenvalues of (1), the convergence of the lower eigenvalues and the divergence of the top end	Said Known (Table 1, Figure 1 Computed)
Reading the lower end of the spectrum from the lattice in the continuum limit (end of section 01)	At the level of Physics . The scaling limit in four dimensions was made by the construction published in 2026-11 (at the level of the committee's assessment; 6-00). Its lower end (the mass gap) has not been shown (6-04)
Hilbert spaces, self-adjointness, the spectrum, the spectral theorem, Weyl's law (10)	Definitions and examples given, theorems cited Known
The variational principle (6), Temple's lower bound (7), Table 2	Said Known , Computed . $\mu = 39$ is a value we could put because we know $\lambda_2 = 4\pi^2$
The Poincaré inequality (8), $C_P = 1/\Delta$, (9), the C_P of products and of Gaussian measures, the $C_P = 1/m^2$ of the free field	Said Known . The one-dimensional line for (9) and the line for (i) of section 09 assume the existence of eigenfunctions. Table 3 and Figure 2 are Computed
The relation between C_P and the correlation length in interacting theories	Not treated (23-01; not proved). $C_P \approx \xi^2$ is a free-field computation
The spectrum of the Hamiltonian of quantum field theory, the Yang–Mills mass gap	Not treated (4-09, 4-11). Nothing is shown

Sources and reproduction

Item	Kind	Source / tools
Tables 1–3, Figures 1–2, the numbers of section 10	Computed	python3 on this machine (NumPy 1.26, SciPy 1.11). Table 3 was found by putting $f = \psi e^{V/2}$, turning L into $-d^2/dx^2 + V'^2/4 - V''/2$ (the $\psi = f\phi_0$ of 2-05 § 08 with $\phi_0 = e^{-V/2}$), and using three-point differences on $[-7, 7]$. The figures are plain JavaScript, with the readouts checked against the tables with node
(4)–(7), the spectral theorem, Weyl's criterion	Known	G. Temple, Proc. R. Soc. Lond. A 119 (1928) 276–293 / W. Ritz, J. reine angew. Math. 135 (1909) 1–61 / J. von Neumann, Math. Ann. 102 (1930) 49–131 (regarded as a representative paper on the spectral theorem in infinite dimensions) / M. Reed, B. Simon, *Methods of Modern Mathematical Physics I* (Academic Press, 1972). All bibliographic data only
(9), the C_P of products and of Gaussian measures in section 09	Known	H. J. Brascamp, E. H. Lieb, J. Funct. Anal. 22 (1976) 366–389 (bibliographic data only; that the inequality on the right of (9) follows as its corollary is not checked) / D. Bakry, M. Émery, Sém. Probab. XIX (1985) 177–206 (the summary at the start checked) / H. Chernoff, Ann. Probab. 9 (1981) 533–535 (abstract checked) / D. Bakry, I. Gentil, M. Ledoux, *Analysis and Geometry of Markov Diffusion Operators* (Springer, 2014; bibliographic data only)
Weyl's law, Lorentz's problem, drums	Known	H. Weyl, Math. Ann. 71 (1912) 441–479 (title checked, bibliographic data only) / M. Kac, Amer. Math. Monthly 73 (1966) no. 4, part II, 1–23 (text checked) / C. Gordon, D. L. Webb, S. Wolpert, Bull. AMS 27 (1992) 134–138, arXiv:math/9207215 (abstract checked)

Next chapter: 1-11 Fourier analysis and distributions. For use in quantum mechanics, see 2-05.

Fourier analysis and distributions — handling things without values at points by waves and averages

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

What this chapter shows — That splitting into frequencies turns differentiation into multiplication. Defining objects without values at points (distributions) by their values when paired with smooth functions. That the propagator diverges near by and decreases as e^{-mr} far away, that its square becomes just barely non-integrable in four dimensions, and that a product of distributions is not determined in general, but is determined if the total roughness is positive.

Prerequisite chapters — 1-10 (Hilbert spaces, orthonormal bases). § 01, § 06 and § 08 use the white noise and free field of 1-09 and the counting of pairings of 1-08 § 06 (formulas restated).

Said first — Theorems are cited with proofs omitted. Products of interacting fields are only touched at the entrance (7-09, 7-10). This chapter shows nothing about the Yang–Mills mass gap.

Order of this chapter

1. Why this chapter — measuring averages instead of values at points
2. Fourier series — smoothness shows in the decay of the coefficients (Figure 1)
3. The Fourier transform — differentiation becomes multiplication, equations become division
4. Test functions and distributions — measuring by testing
5. The Fourier transform of distributions — δ contains all frequencies equally
6. Rulers of roughness — Sobolev spaces and the scaling exponent ((9a)(9b))
7. The propagator — divergent near by, e^{-mr} far away (Figure 2, Table 1)
8. Counting singularities — four dimensions is the borderline (Table 2, Figure 3)
9. The trouble with products — the way of smoothing remains in the answer (Figure 4)

10. When they can be made — if the total roughness is positive

---

01

Why this chapter — averages instead of values at points

The free field in two or more dimensions has no value at a single point, because the variance at a point is infinite (1-09 § 08). What has meaning is the value averaged with a smooth weight f, and its variance can be written as a double average of the two-point correlation G.

$$\varphi(f) = \int_{\mathbb{R}^d} \varphi(x)\, f(x)\, d^d x, \qquad E\big[\varphi(f)^2\big] = \iint f(x)\, G(x-y)\, f(y)\, d^d x\, d^d y \quad (1)$$

G is called the **propagator**, and formally $E[\varphi(x)\varphi(y)] = G(x-y)$. G is infinite at $x = y$, but the right-hand side of (1) is finite. **Example:** the variance of the value averaged over a box of side ε is, for $d \geq 3$, proportional to ε^{2-d} by $G \approx c\, r^{2-d}$ (§ 07); in four dimensions it is 4 times larger each time the box is halved. Infinite at a point, finite when averaged.

02

Fourier series — splitting a periodic function into waves

Split a function of period 2π into a superposition of waves e^{ikx}. $e^{ikx}/\sqrt{2\pi}$ ($k \in \mathbb{Z}$) is an orthonormal basis of $L^2(-\pi, \pi)$ (the complex version of 1-10 § 02), and the coefficients are extracted by inner products Known .

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\, e^{-ikx}\, dx, \qquad f = \sum_{k \in \mathbb{Z}} c_k\, e^{ikx}, \qquad \frac{1}{2\pi} \int_{-\pi}^{\pi} |f|^2\, dx = \sum_{k \in \mathbb{Z}} |c_k|^2 \quad (2)$$

The last is Parseval's identity. Integrating by parts, differentiation is replaced by multiplying the coefficients by ik.

$$c_k[f'] = ik\, c_k[f] \qquad (f \text{ periodic and smooth}) \quad (3)$$

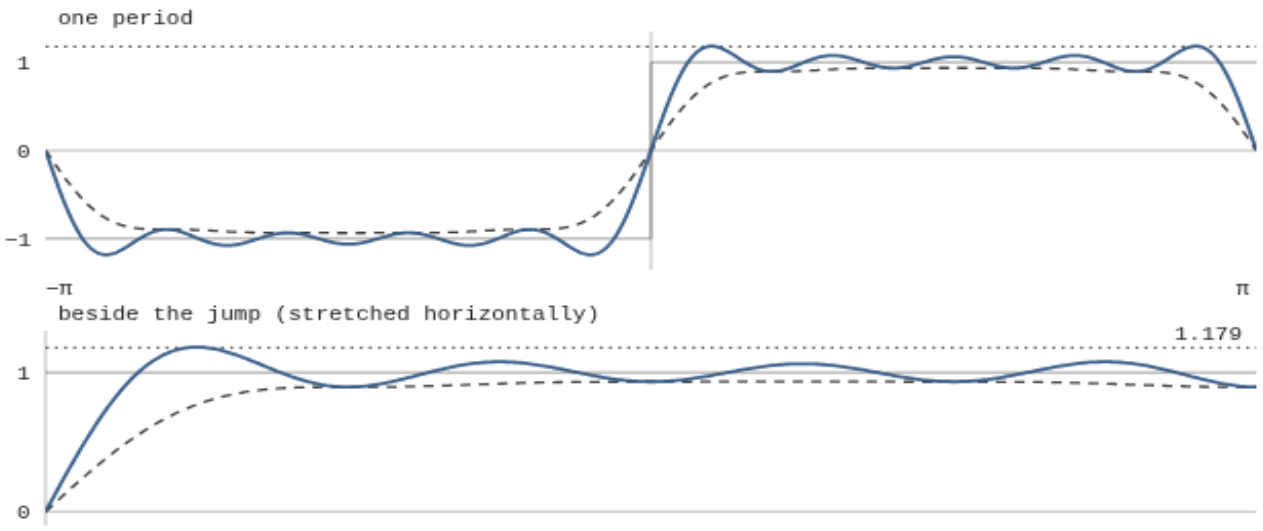
So **smoothness shows in the decay of the coefficients**. If the function is still square-integrable after n derivatives, $\sum_k k^{2n} |c_k|^2 < \infty$. With a jump, the coefficients decrease only as $1/k$.

Example: the partial sums of the square wave $f(x) = \operatorname{sign} x$ are $S_N(x) = \frac{4}{\pi} \sum_{k \text{ odd} \leq N} \frac{\sin kx}{k}$. The mean-square error goes to 0 as $\frac{8}{\pi^2} \sum_{k \text{ odd} > N} k^{-2} \approx \frac{4}{\pi^2 N}$ by Parseval, yet

the maximum overshoots to about 1.179 beside the jump, and the ratio $\frac{1}{\pi}\int_0^{\pi}\frac{\sin t}{t}dt-\frac{1}{2}$ $=0.08949\cdots$ to the jump size 2 remains regardless of N. This is the **Gibbs phenomenon**; Wilbraham (1848) is said to have studied it first, Gibbs's letter of 1899 to have made it widely known, and Bôcher (1906) to have named it Known .

Computed The maximum of S_N (the value at $x=\pi/(N+1)$) is 1.27324, 1.18233, 1.17898 at $N=1,9,999$ (limit $\frac{2}{\pi}\,\mathrm{Si}(\pi)$), and the mean-square error is 0.18943, 0.04040, 0.000405 (Parseval's formula and numerical integration agree; for $N=999$, 0.0004052846 agrees to 10 digits). It converges in mean square, and not uniformly. "Convergence" has meanings of different strength. The **Fejér mean** $\sigma_N=\frac{1}{N+1}(S_0+S_1+\cdots+S_N)$, which averages the partial sums, smooths with a positive kernel, so no overshoot occurs ($|\sigma_N|\le 1$) Known . "Smooth first, then take the limit" is the basic move of approximation in § 04.

Figure 1 Partial sums S_N (solid line) and Fejér means σ_N (dashed) of the square wave (thin line). Above, one period; below, the region beside the jump $0<x<8\pi/(N+1)$ ($0<x<\pi$ at $N<7$) stretched horizontally. Increasing N, above S_N approaches the square wave, yet below the height of the peak stays at about 1.179 (dotted line), and σ_N does not exceed 1. "Play" increases N



N = 9, maximum of partial sum 1.18233 (x = π/(N+1), limit 1.17898), overshoot / jump 0.09116, maximum of Fejér mean 0.93573, mean-square error 0.040395 (Parseval)

Computed The numbers in this figure were computed on this machine (the partial sums of the series are added on the spot; the readout agrees with the computation above to 5 digits at $N=1,3,5,9,19,49,99,199$).

03

The Fourier transform — differentiation becomes multiplication

For functions without a period the frequencies are arranged continuously, and sums are replaced by integrals. The conventions of this book are as follows. Taking the measure as $d^dk/(2\pi)^d$ matches (10) of 1-09.

$$\hat{f}(k) = \int_{\mathbb{R}^d} f(x)\, e^{-ik\cdot x}\, d^d x, \qquad f(x) = \int_{\mathbb{R}^d} \hat{f}(k)\, e^{ik\cdot x}\, \frac{d^d k}{(2\pi)^d} \quad (4)$$

The same rules as for series hold Known . Length is preserved (Plancherel), differentiation becomes multiplication, and convolution $(f * g)(x) = \int f(y) g(x-y)\, d^d y$ becomes a product.

$$\int |f|^2\, d^d x = \int |\hat{f}|^2\, \frac{d^d k}{(2\pi)^d}, \qquad \widehat{\partial_j f} = ik_j\, \hat{f}, \qquad \widehat{f * g} = \hat{f}\, \hat{g} \quad (5)$$

Then $\widehat{-\Delta f} = |k|^2 \hat{f}$, and a differential equation with constant coefficients becomes division frequency by frequency. If $m > 0$ the denominator is never 0, and the solution is determined uniquely.

$$(-\Delta + m^2)\, u = f \quad \Longleftrightarrow \quad \hat{u}(k) = \frac{\hat{f}(k)}{|k|^2 + m^2} \quad (6)$$

Example: the transform of a Gaussian of width σ is a Gaussian of width $1/\sigma$ (the uncertainty relation of 2-05). The transform of $e^{-m|x|}$ is $\frac{2m}{k^2+m^2}$: the corner at $x = 0$ shows up as the slow decrease $1/k^2$ (the one-dimensional propagator of § 07).

04

Test functions and distributions — measuring by testing

Objects without values at points are defined by their values when paired with smooth functions. The partners are called **test functions**, taken to be functions that can be differentiated any number of times and are 0 outside a bounded set (the whole set is $\mathcal{D}$. Example: $e^{-1/(1-|x|^2)}$ on $|x| < 1$, 0 outside).

A **distribution** is a linear and continuous rule assigning a number $\langle T, \phi \rangle$ to a test function ϕ (continuous meaning that $\langle T, \phi_n \rangle \to 0$ for sequences ϕ_n whose supports (the closure of the set of points where the function is non-zero) lie in one bounded set and all of whose derivatives approach 0 uniformly) Known . "Measuring by testing" in the title is a metaphor; to be exact, it means regarding this rule $\phi \mapsto \langle T, \phi \rangle$ itself as the object instead of values at points. The English word is distribution, and this is the "distributions" of the chapter title (not the "distribution" of probability). The idea of distributions is said to begin with Sobolev (1936) and to have been systematised by Schwartz Known .

Examples: (a) a function f integrable on bounded sets becomes a distribution by $\langle f, \phi \rangle = \int f \phi\, d^d x$. (b) The δ of $\langle \delta, \phi \rangle = \phi(0)$ cannot be written as the integral of any function. Differentiation is defined by promoting integration by parts to a definition.

$$\langle \partial_j T, \phi \rangle = -\langle T, \partial_j \phi \rangle \quad (7)$$

The right-hand side always makes sense, so every distribution can be differentiated any number of times. For the step function H (1 for $x > 0$, 0 for $x < 0$), $\langle H', \phi \rangle = -\int_0^\infty \phi'\, dx = \phi(0)$, that is, $H' = \delta$.

Approximation: if $\rho \geq 0$ is smooth with $\int \rho = 1$ and $\rho_\varepsilon(x) = \varepsilon^{-d}\rho(x/\varepsilon)$, then $\langle \rho_\varepsilon, \phi \rangle \to \phi(0)$, that is, $\rho_\varepsilon \to \delta$. $\sin nx$ does not converge pointwise, but integration by parts gives $\frac{1}{n}$, so as a distribution it converges to 0.

05

The Fourier transform of distributions — δ contains all frequencies equally

For transforms we use partners wider than $\mathcal{D}$. On the set $\mathcal{S}$ of functions that can be differentiated any number of times and whose values and derivatives decrease faster than any power of $|x|$ (the Schwartz space), the transform maps $\mathcal{S}$ one-to-one onto $\mathcal{S}$ Known . The transform of a continuous linear rule on $\mathcal{S}$ (a **tempered distribution**) is defined to match $\int \hat{f}\phi = \int f\hat{\phi}$.

$$\langle \hat{T}, \phi \rangle = \langle T, \hat{\phi} \rangle \qquad (\phi \in \mathcal{S}) \quad (8)$$

Example: since $\langle \hat{\delta}, \phi \rangle = \hat{\phi}(0) = \int \phi\, d^d x$, $\hat{\delta} = 1$. Something concentrated at one point contains all frequencies with the same weight (the origin of the "white" in white noise).

Conversely, the transform of the constant 1 is $(2\pi)^d\delta$: something of the same size everywhere contains only the frequency 0. The rules of (5) also extend as they are, giving $\widehat{\partial_j T} = ik_j\hat{T}$ Known . This is why in § 07 the propagator can be written $\hat{G} = 1/(|k|^2 + m^2)$.

06

Rulers of roughness — Sobolev spaces and the scaling exponent

Smoothness shows in the decay of the coefficients or of $\hat{f}$ ((3)(5)). We set up two rulers to measure this with numbers. The first is a length that puts weights on frequencies, and the set of objects for which it is finite is called the Sobolev space H^s. The larger s, the smoother the things admitted; the more negative s, the rougher the things admitted.

$$\|f\|_{H^s}^2 = \int (1 + |k|^2)^s\, |\hat{f}(k)|^2\, \frac{d^d k}{(2\pi)^d} \quad (9a)$$

The second is behaviour under rescaling. The $T(\lambda\,\cdot)$ of a distribution is defined by $\langle T(\lambda\,\cdot),\phi\rangle = \lambda^{-d}\langle T,\phi(\cdot/\lambda)\rangle$, promoting the change-of-variables formula to a definition.

$$T(\lambda\,\cdot) = \lambda^{\alpha}\,T \quad (\text{for all } \lambda>0) \quad : \text{then } \alpha \text{ is called the scaling exponent} \tag{9b}$$

"Ruler" is a metaphor; to be exact, it means the two numbers s of (9a) and α of (9b). Since $\hat{\delta}=1$, $\|\delta\|_{H^s}$ is finite only when $s<-d/2$. The scaling exponent is $-d$ for δ and $-a$ for $|x|^{-a}$.

For a random object T, when $T(\lambda\,\cdot)$ and $\lambda^{\alpha}T$ have the same probability distribution, we take the exponent to be α. For white noise ξ ($E[\xi(f)\xi(g)] = \int fg\,d^dx$, the $\mathbb{R}^d$ version of 1-09 § 07), the covariance of the values of $\xi(\lambda\,\cdot)$ paired with test functions ϕ,ψ is $\lambda^{-2d}\int\phi(x/\lambda)\psi(x/\lambda)\,d^dx = \lambda^{-d}\int\phi\psi\,d^dx$, and a Gaussian with mean 0 is fixed by its covariance, so $\xi(\lambda\,\cdot)$ has the same probability distribution as $\lambda^{-d/2}\xi$. The exponent is $-d/2$. The massless free field ($m=0$, $d\ge 3$) has covariance $G(\lambda x)=\lambda^{2-d}G(x)$ (the near-side formula of § 07 is exact for $m=0$), so its exponent is $1-d/2$. For $d=2$, G is logarithmic, so the same holds up to constants Known . Measuring white noise in spacetime, including time, by the parabolic scaling that multiplies time by λ^2 and space by λ, the exponent is $-(d+2)/2$ (Exercise 2.3 of Hairer's lectures) Known . This exponent decides in § 10 whether products can be made.

07

The propagator — a field responding to δ

The solution of the equation with δ on the right-hand side of (6) is the propagator. "Responding to δ" is a metaphor; to be exact, it means satisfying the first formula of (10) as an identity of distributions, and such a solution is called a **fundamental solution**. From $\frac{1}{|k|^2+m^2} = \int_0^\infty e^{-t(|k|^2+m^2)}dt$ and the transform of the Gaussian, there also comes a representation as the time integral of the fundamental solution $(4\pi t)^{-d/2}e^{-|x|^2/4t}$ (the **heat kernel**) of the heat equation $\partial_t u=\Delta u$.

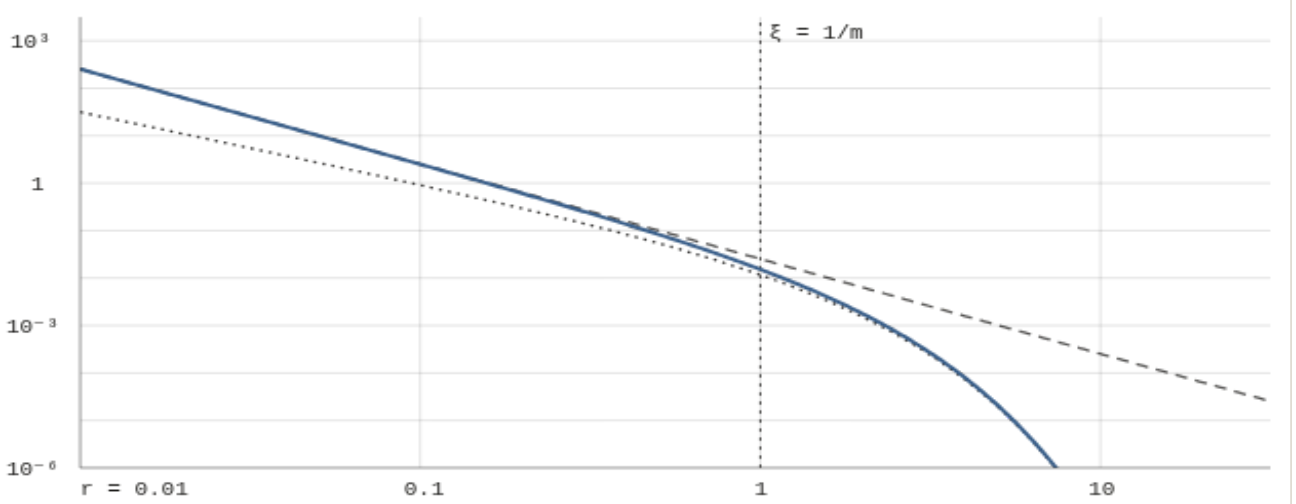
$$(-\Delta+m^2)\,G=\delta, \qquad \hat{G}(k)=\frac{1}{|k|^2+m^2}, \qquad G(x)=\int_0^\infty \frac{e^{-|x|^2/4t-m^2t}}{(4\pi t)^{d/2}}\,dt \tag{10}$$

G is the covariance of the free field, and in physics the Euclidean propagator of a particle of mass m (4-02, 4-03). With $r=|x|$, it is $e^{-mr}/2m$ for $d=1$, $e^{-mr}/4\pi r$ for $d=3$, and in general $(2\pi)^{-d/2}(m/r)^{d/2-1}K_{d/2-1}(mr)$ using the modified Bessel function $K_\nu(s)=\int_0^\infty e^{-s\cosh t}\cosh\nu t\,dt$ Known . What matters are its two faces.

$$G(r) \approx \begin{cases} \dfrac{\Gamma(d/2-1)}{4\pi^{d/2}}\, r^{2-d} & (r \to 0,\ d \geq 3) \\ \dfrac{1}{2\pi} \ln \dfrac{1}{r} & (r \to 0,\ d = 2) \end{cases} \qquad G(r) \approx \frac{1}{2m} \left(\frac{m}{2\pi r} \right)^{\frac{d-1}{2}} e^{-mr} \quad (r \to \infty) \quad (11$$

On the **near face** m disappears and it diverges as $d \geq 2$. On the **far face** it decreases as e^{-mr}. "Face" is a metaphor; to be exact, it means that as $r \to 0$ the ratio of $G(r)$ to the left formula of (11) approaches 1, and as $r \to \infty$ the ratio to the right formula approaches 1. Calling the ξ for which correlations decrease as $e^{-r/\xi}$ the correlation length (2-08), $\xi = 1/m$, and "rate of decrease of correlations = mass" is the free-field version of the definition of the mass gap (4-11).

Figure 2 A log–log graph of the propagator $G(r)$ (solid line). Moving d and m, the near side follows the left formula of (11) (dashed; for $d = 2$ with the constant $(\ln(2/m) - \gamma)/2\pi$ added, for $d = 1$ the value at the origin $1/(2m)$), the far side follows the right formula (dotted), and the bend comes at $\xi = 1/m$ (vertical line). "Play" moves m back and forth



d = 4, m = 1.00, G(0.01) = 2.532e+2 (ratio to the near-side formula 0.9997), G(10ξ) = 4.724e−8 (ratio to the far-side formula 1.036)

Computed The numbers in this figure were computed on this machine (the integral representation of K_ν by the trapezoidal rule; agreeing with Table 1 to 7 digits).

Computed **Table 1** G for $m = 1$ and its ratio to the formulas of (11) (the near side for $d = 2$ adds the constant $\ln 2 - \gamma$). The Bessel function formula and numerical integration of (10) agree to 15 digits.

d	r	G(r)	Ratio to near side	Ratio to far side
2	0.01	0.7514	1.0000	0.3805
4	0.01	253.2	0.9997	8.057
4	1	0.01525	0.6019	1.306
4	10	4.724×10^{-8}	0.0002	1.036

The lattice propagator

The propagator on a lattice of spacing a is the integrand of (10) of 1-09 multiplied by $e^{ik\cdot x}$ and integrated ((10) of 1-09 is its value at $x=0$). For $d=3$, $m=1$ at distance 1, the difference from the continuum G shrinks in proportion to a^2: 0.9% at $a=1/8$ and 0.2% at $a=1/16$ Computed (5-01).

08

Counting singularities — four dimensions is the borderline

Even if it diverges at the origin, it is a distribution by (a) of § 04 if it can be integrated near by. For powers it is a comparison with the dimension ($|S^{d-1}|$ is the area of the unit sphere).

$$\int_{\varepsilon<|x|<1} |x|^{-a}\, d^d x = |S^{d-1}| \int_\varepsilon^1 r^{d-1-a}\, dr \xrightarrow[\varepsilon\to 0]{} \begin{cases} \text{finite} & (a<d) \\ \ln(1/\varepsilon) \text{ divergence} & (a=d) \\ \varepsilon^{d-a} \text{ divergence} & (a>d) \end{cases} \tag{12}$$

$G\sim r^{2-d}$ is $a=d-2<d$, so it is a distribution in every dimension. Its n-th power is $a=n(d-2)$, integrable when $n(d-2)<d$; for the square when $d<4$, for the cube when $d<3$.

Powers of the propagator come from products of fields. The fourth moment of a Gaussian is a sum over ways of pairing (1-08 § 06). In $:\varphi^2: = \varphi^2 - E[\varphi^2]$, with the infinite $E[\varphi^2]=G(0)$ subtracted (the Wick product), only the pairings joining the φ of x with the φ of y remain (the pairings among the x and among the y cancel with what was subtracted) Known .

$$E\big[:\varphi^2:(x)\,:\varphi^2:(y)\big] = 2\,G(x-y)^2, \qquad E\big[:\varphi^2:(f)^2\big] = 2\iint f(x)\,G(x-y)^2\,f(y)\,d^dx\,d^dy \tag{13}$$

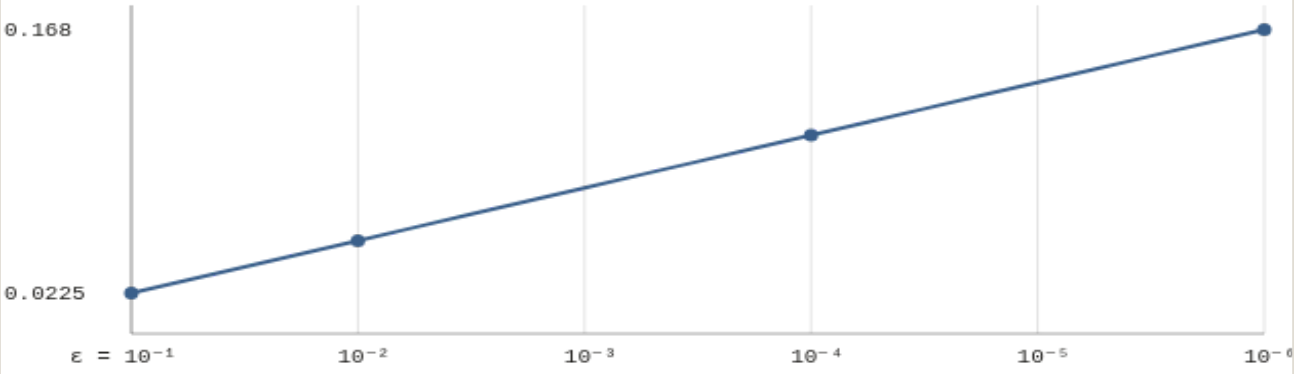
The variance of $:\varphi^2:(f)$ is finite when G^2 is integrable, $d<4$; for the cube, $6G^3$, when $d<3$. A figure setting this beside other computations that switch at four is the dimension slider of 1-12 § 10.

Computed **Table 2** $I_n(\varepsilon)=\int_{\varepsilon<|x|<1} G(x)^n\,d^dx$ ($m=1$, numerical integration in the radius). Pressing a row draws it in Figure 3.

d	n	ε = 0.1	0.01	10^{-4}	10^{-6}	ε → 0
2	3	0.01079	0.01408	0.01426	0.01426	finite
3	2	0.02719	0.03362	0.03440	0.03440	finite
3	3	0.00565	0.01866	0.04763	0.07679	diverges logarithmically

d	n	ε = 0.1	0.01	10^{-4}	10^{-6}	ε → 0
4	1	0.1851	0.1876	0.1876	0.1876	finite
4	2	0.02248	0.05143	0.1097	0.1681	diverges logarithmically
4	3	0.01404	1.598	16041	1.604×10^{8}	diverges as ε^{-2}

Figure 3 Pressing a row of Table 2 (on the keyboard, Tab and Enter) draws $I_n(\varepsilon)$ for that (d, n) against $\log_{10}(1/\varepsilon)$. If finite it levels off, if it diverges logarithmically it is a straight line, and if it diverges as a power it rises steeply (a straight line with a logarithmic vertical axis). The circles are the values in the columns of Table 2



d = 4, n = 2 (diverges logarithmically), increase from ε = 10^{-5} to 10^{-6} 0.0292, ratio 1.21

Computed The numbers in this figure were computed on this machine (a table embedding to 4 digits the same numerical integration in the radius as Table 2, done at $\log_{10}(1/\varepsilon) = 1, 1.25, \ldots, 6$).

$d = 4, n = 2$ grows by 0.02916 each time ε is divided by 10, agreeing with $\ln 10/8\pi^2$ from (11)(12) Computed . Four dimensions is the borderline where G^2 just stops being integrable, and the coupling constant being dimensionless in four dimensions (4-04) is the same counting of scales Known (1-12).

09

The trouble with products — smoothing before multiplying leaves the way of smoothing behind

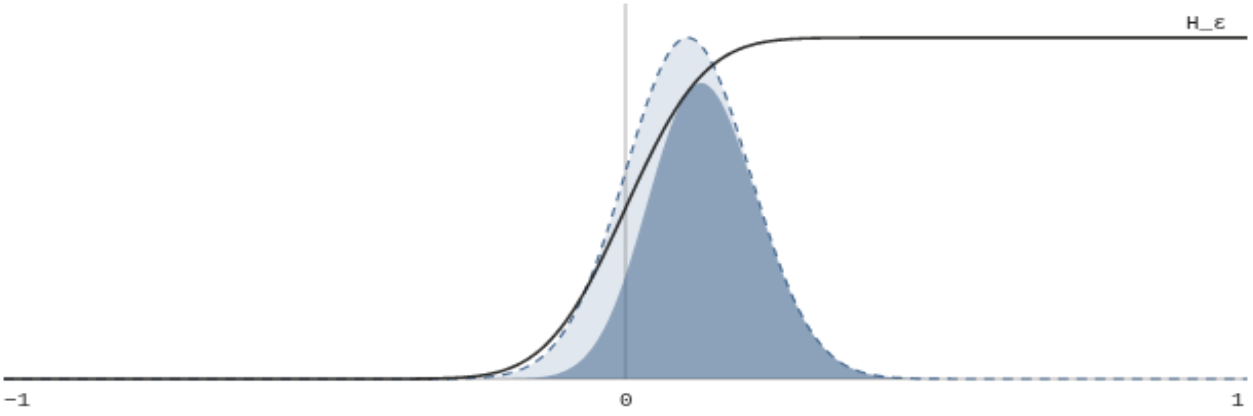
The product of a smooth g and a distribution T is determined by $\langle gT, \phi\rangle = \langle T, g\phi\rangle$. "Smoothing, multiplying and removing the smoothing" for two distributions makes three things happen. **(i) It diverges:** the area of ρ_ε^2 goes to infinity as $\varepsilon^{-d}\int \rho^2$. **(ii) The way of smoothing remains:** smooth H with ρ and δ with ψ (both probability densities), and replace by $x = \varepsilon X$.

$$\int (H * \rho_\varepsilon)\,\psi_\varepsilon\,\phi\,dx = E\big[\mathbf{1}\{Y \le X\}\,\phi(\varepsilon X)\big] \xrightarrow[\varepsilon\to 0]{} c\,\phi(0), \qquad c = P(Y \le X) \quad (Y \sim \rho,\ X$$

The limit is $c\,\delta$, but c is determined by the way of smoothing. Taking ρ to be the standard Gaussian and ψ a Gaussian centred at s gives $c = \Phi(s/\sqrt{2})$ (Φ is the

distribution function of the standard normal distribution), which can be any value between 0 and 1.

Figure 4 The smoothed H (solid line), the shape of the smoothed δ (dashed; height to scale), and their product (dark shading). Moving s changes the area of the product between 0 and 1, and shrinking ε does not change it. The area of the product of δ with itself grows as $1/\varepsilon$. "Play" shrinks ε



ε = 0.1000, s = 1.0, area of the product of H and δ = 0.76025, area of the product of δ with itself = 2.197

Computed The numbers in this figure were computed on this machine (closed formulas; checked against the numerical integration below).

Computed In numerical integration, the area of the product of H and δ is 0.078650, 0.500000, 0.760250, 0.921350 at $s = -2, 0, 1, 2$ (the same at $\varepsilon = 0.1$ and 0.01, agreeing with $\Phi(s/\sqrt{2})$ to 6 digits), and that of δ with itself is 1.038, 2.821, 2.197, 1.038 at $\varepsilon = 0.1$, and 10 times that at $\varepsilon = 0.01$.

(iii) Associativity breaks. For the principal value $\langle \mathrm{p.v.}\frac{1}{x}, \phi\rangle = \lim_{\varepsilon\to 0}\int_{|x|>\varepsilon}\frac{\phi(x)}{x}dx$, as products "smooth function $\times$ distribution", $x\cdot\delta = 0$ and $x\cdot\mathrm{p.v.}\frac{1}{x} = 1$.

$$(\delta\cdot x)\cdot\mathrm{p.v.}\frac{1}{x} = 0, \qquad \delta\cdot\left(x\cdot\mathrm{p.v.}\frac{1}{x}\right) = \delta \quad (15)$$

A product keeping all the properties of the product of functions cannot be put on the whole space of distributions. To be exact, it is said that there is no associative product on distributions (containing at least the continuous functions and δ) that agrees with the ordinary product on continuous functions and for which differentiation satisfies the Leibniz rule $(fg)' = f'g + fg'$ (Schwartz 1954) Known .

10

When they can be made — if the total roughness is positive

The condition for products to be made can be written with the scaling exponent of § 06. Let C^α be the following space. For $0 < \alpha < 1$, the set of Hölder continuous functions of order α ($|f(x) - f(y)| \le C|x-y|^\alpha$, $|x-y| \le 1$); for $\alpha > 1$, the same form of

condition imposed on derivatives; for $\alpha < 0$, the set of distributions whose scaling exponent in the sense of § 06 is locally uniformly at least α. The larger α, the smoother. If the total roughness is positive, products can be made (Proposition 4.14 of Hairer 2014) Known .

$$f \in C^{\alpha},\ \ g \in C^{\beta},\ \ \alpha + \beta > 0 \quad \Longrightarrow \quad f \cdot g \in C^{\min(\alpha,\beta)} \text{ is defined continuously } (16)$$

"Defined continuously" means that extending the product of smooth functions by limits gives a single result independent of how the limit is approached. The same proposition states that, if α is not a natural number, $\alpha + \beta > 0$ is also necessary.

Example: the step function H is bounded but has a jump, and its roughness exponent is 0 ($H(\lambda x) = H(x)$; it lies in $\alpha < 0$ for every C^{α}). The one-dimensional δ is -1. The total is not positive, so (16) cannot be used, consistent with the phenomenon of (ii) in § 09. The free field is slightly rougher than $1 - d/2$ (it lies in C^{α} for $\alpha < 1 - d/2$), so its square can be made as the determined product of (16) only for $d < 2$ Known . In § 08 the variance of $:\varphi^2:(f)$ became finite for $d = 2, 3$ because the divergence of (i) in § 09 was subtracted (renormalised) first, and the fact that it takes a form independent of the way of smoothing depends on probabilistic cancellation (the counting of Gaussian pairings).

The Yang–Mills connection A is expected to be as rough as the free field at short distances Physics , and the $A \wedge A$ of the field strength $F = dA + A \wedge A$ (1-07) is a product that does not satisfy (16) in two or more dimensions. In two and three dimensions the analysis of products of this kind is being advanced by stochastic quantisation (7-09), and in four dimensions the construction published in 2026-11 made the continuum theory as a limit of the lattice (the committee's assessment is that it is proved in Lean; 6-00). There, local products are defined as insertions renormalised on the lattice (8-04). What had been reached before that is in 7-10.

What this chapter can say

Content	Status
The rules of Fourier (2)–(6), the Gibbs overshoot, the Fejér mean	Said Known . The numbers and Figure 1 are Computed
Distributions, differentiation (7), transform (8), rulers (9a)(9b), the scaling exponents of white noise and the free field	Definitions and examples Known . The two exponents have a one-line derivation in § 06

Content	Status
The propagator (10)(11), the correlation length $= 1/m$, the condition for G^n to be integrable, the variance of the Wick product (13)	Said Known . Tables 1 and 2 and Figures 2 and 3 are Computed
The trouble with products (14)(15), the condition for products to be made (16)	The computation of (14) is shown. Figure 4 and the numerical integration are Computed . Schwartz's theorem bibliographic data only; the statement of (16) checked (the proof not followed)
The lattice error a^2 (one point at $d = 3$) / the roughness of the Yang–Mills connection / interacting fields, the mass gap	Observation Computed / expectation Physics / not treated; nothing is shown

Sources and reproduction

Item	Kind	Source / tools
Tables, figures, numbers in the text	Computed	python3 on this machine (NumPy 1.26, SciPy 1.11). The figures are plain JavaScript: the partial sums of Figure 1, the integral representation of K_ν in Figure 2 and the closed formulas of Figure 4 are computed on the spot, and Figure 3 embeds a table computed in python. The readouts were checked against the tables with node
The history of the Gibbs phenomenon	Known	E. Hewitt, R. E. Hewitt, Arch. Hist. Exact Sci. 21 (1979) 129–160 (text checked: Part II) / H. Wilbraham, Cambridge and Dublin Math. J. 3 (1848) 198–201, J. W. Gibbs, Nature 59 (1899) 606 (bibliographic data only)
The beginning of distributions and the impossibility of products	Known	S. Soboleff, Rec. Math. [Mat. Sbornik] N.S. 1(43) (1936) 39–72, L. Schwartz, *Théorie des distributions* (Hermann), L. Schwartz, C. R. Acad. Sci. Paris 239 (1954) 847–848 (all bibliographic data only). The attribution of the beginning of the idea of distributions, and the year of *Théorie des distributions* (said to be 1950–51), not checked (from secondary sources)
(16)	Known	M. Hairer, Invent. Math. 198 (2014) 269–504, arXiv:1303.5113 (text checked: § 1 and Proposition 4.14, including the second half, that $\alpha + \beta > 0$ is also necessary if $\alpha \notin \mathbb{N}$)
The scaling exponent	Known	The exponent $-(d+2)/2$ of spacetime white noise under parabolic scaling: M. Hairer, *An Introduction to Stochastic PDEs*, arXiv:0907.4178 (text checked: Exercise 2.3). The $-d/2$ of spatial white noise and the $1 - d/2$ of the massless free field are from the one-line derivation in § 06

Item	Kind	Source / tools
The transform, $\mathcal{S}$, tempered distributions, Sobolev spaces, the formulas for the propagator	Known	M. Reed, B. Simon, *Methods of Modern Mathematical Physics II* (Academic Press, 1975) and other standard textbooks (bibliographic data only)

Next chapter: 1-12 Detour: what happens only in four dimensions. For uses of the propagator see 4-03; for the axioms of fields with values in distributions, 4-09.

Detour: what happens only in four dimensions — rereading eight cases from the Yang–Mills side

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

What this chapter shows — That rotations in four dimensions split into two copies of SU(2), and the field strength splits into self-dual and anti-self-dual parts. That only in four dimensions is the action independent of scale, so that a field can collapse to a point while keeping its action. That integers appear because S^3 at infinity and SU(2) have the same shape.

Prerequisite chapters — None (it is easier if you have read 1-02 (SU(2)), 1-03 (Lie algebras), 1-06 ($dx_1 \wedge dx_2$ and the Hodge star) and 1-07 (gauge transformations and pure gauge)). For the counting of dimensions, see 2-04.

Said first — This is a detour, so you may skip it. It is a guide to known facts, and shows nothing about the mass gap or the construction of the continuum limit. The main treatment is in 3-05, 3-06 and 4-05.

Order of this chapter

1. A sketch map — the eight of the article and this book
2. Multiplying on the left and on the right
3. The field strength splits in two (Figure 1)
4. Changing the scale does not change the action (Figure 2)
5. Collapsing to a point (Table 1)
6. The sphere at infinity and SU(2)
7. Planes meet in a point (Detour: knots and Wilson loops)
8. The 24-cell is a finite subgroup of SU(2) (Table 2)
9. The area of the sphere, and integrals that begin to diverge in four dimensions (Table 3, Figure 3)
10. The dimension slider (Figure 4)

01

A sketch map — the eight of the article, and where they matter in this book

The starting point of the article "What happens only in four dimensions" was the count that, while there are n directions to move in, the number of planes one can rotate in is

$$\dim SO(n) = \binom{n}{2} = \frac{n(n-1)}{2} \quad (1)$$

In three dimensions planes correspond one-to-one with perpendicular axes, and a rotation can be described by its axis. In four dimensions what is perpendicular to each of the 6 planes is again a plane, and this 6 splitting into 3 + 3 is the starting point of many sections. The way it splits is written concretely in sections 02 and 03, as multiplication of quaternions on the left and on the right.

The table shows where the eight cases of the article appear in this book. The left column links to the headings of the article, the right column to sections of this chapter and to the chapters with the main treatment. Topics not in the article (scale and concentration, the winding number at infinity) are added in sections 04–06. The numbers on the left of the table are the numbers of the cases in the article, separate from the section numbers of this chapter. This chapter itself is a detour, so you may pick out and read only the rows that interest you.

Case in the article	Its face in this book	Section / chapter
01 The axis disappears from rotation	The rotation group splits into two copies of SU(2), the field strength into two	02, 03, 3-05
02 The regular polytopes increase (the 24-cell)	The vertices form a finite subgroup of SU(2)	08, 5-10
03 Knots come untied	Wilson loops do not see knots	07
04 A left hand becomes a right hand	A mirror swaps self-dual and anti-self-dual	03
05 Planes meet in a point / 06 Exotic $\mathbb{R}^4$	Intersection forms; Yang–Mills sees topology	07, 3-06
07 Difficulty is not monotone	The dimension at the borderline where the coupling becomes dimensionless	04, 05, 7-10
08 The ball is largest in five dimensions	Moving the dimension continuously in the same formula	09, 4-05
(not in the article)	Winding number, instanton number	06, 3-05

02

Multiplying on the left and on the right — rotations in four dimensions are two copies of SU(2)

Read a point of four dimensions as a single number. Quaternions are numbers made by adding the symbols i, j, k to the real numbers, and the rules are the following one line. Multiplication is associative but not commutative ($ij = k$, $ji = -k$).

$$x = x_1 + x_2 i + x_3 j + x_4 k, \qquad i^2 = j^2 = k^2 = ijk = -1, \qquad \bar{x} = x_1 - x_2 i - x_3 j - x_4 k,$$

The quaternions of length 1 form S^3, and under multiplication they are the same group as SU(2) (that SU(2) is S^3 is in section 05 of 1-02) Known . Multiplying by q, p of length 1 on the left and on the right gives a rotation of four dimensions, and every rotation can be written in this form. Only the pairs (q, p) and $(-q, -p)$ give the same rotation Known . Unlike $SO(3) \cong SU(2)/\{\pm 1\}$ in three dimensions (1-02), there are two copies of SU(2).

$$x \longmapsto q\, x\, \bar{p} \quad (|q| = |p| = 1), \qquad SO(4) \cong \big(SU(2) \times SU(2)\big)/\{\pm(1,1)\} \ (3)$$

Example. If $q = \cos a + i \sin a$, $p = \cos b + i \sin b$, the (x_1, x_2) plane turns by the angle $a - b$ and the (x_3, x_4) plane by the angle $a + b$. The two sliders of case 01 (the speeds of the xy plane and the zw plane) are the difference and the sum of the left and right angles. We checked with 200 random pairs that (3) is an orthogonal matrix of determinant 1 Computed .

03

The field strength splits in two — the left turns only F^+, the right only F^-

The field strength F is at each point an antisymmetric array $F_{\mu\nu} = -F_{\nu\mu}$, with 6 independent components attached one to each plane. For electromagnetism, these 6 components are the 3 of the electric field E and the 3 of the magnetic field B ((13) of 1-06). In Yang–Mills each component is an element of a Lie algebra (1-03), but in this section we take them to be real numbers. The Hodge star $*$ is the operation taking each plane to the perpendicular plane; with the totally antisymmetric symbol of $\varepsilon_{1234} = 1$,

$$(*F)_{\mu\nu} = \frac{1}{2} \sum_{\rho,\sigma} \varepsilon_{\mu\nu\rho\sigma} F_{\rho\sigma}, \qquad (*F)_{12} = F_{34}, \quad (*F)_{13} = F_{42}, \quad (*F)_{14} = F_{23} \ (4)$$

The $*$ in n dimensions sends 2-forms to $(n-2)$-forms, so only for $n = 4$ does it return to 2-forms. Since $** = 1$ (1-06), it can be split as $F^{\pm} = \frac{1}{2}(F \pm *F)$ ($|F|^2 = \sum_{\mu<\nu} F_{\mu\nu}^2$), and the density of the action and the density of topology (section 06) can be written with the same two. $F = \pm * F$ are called self-dual and anti-self-dual.

$$|F|^2 = |F^+|^2 + |F^-|^2, \qquad \frac{1}{4}\sum \varepsilon_{\mu\nu\rho\sigma} F_{\mu\nu} F_{\rho\sigma} = |F^+|^2 - |F^-|^2 \quad (5)$$

Using (4), $F^{\pm}$ has three independent components each, and arranged so that the squared length is $|F^{\pm}|^2$ they take the following form. The three axes of Figure 1 are in this order.

$$F^+ = \tfrac{1}{\sqrt{2}}(F_{12} + F_{34},\ F_{13} + F_{42},\ F_{14} + F_{23}), \qquad F^- = \tfrac{1}{\sqrt{2}}(F_{12} - F_{34},\ F_{13} + F_{24},\ F_1$$

The key point is the relation with (3). The rotation $x \mapsto qx\bar{p}$ turns the 3 components of F^+ by q only and the 3 components of F^- by p only, and does not mix the two Known .

$$\Lambda^2\mathbb{R}^4 = \Lambda^2_+ \oplus \Lambda^2_- \ (3+3), \qquad q \text{ turns } \Lambda^2_+ \ ,\ p \text{ turns } \Lambda^2_- \quad (6)$$

Which one turns F^+ is fixed by convention. Changing the orientation or the assignment of coordinates to quaternions swaps left and right. This book follows (2) and $\varepsilon_{1234} = 1$. In this convention, the matrix multiplying by i on the left is proportional to $dx1 \wedge dx2 + dx3 \wedge dx4$ and self-dual, and the matrix multiplying on the right is proportional to $dx_1 \wedge dx_2 - dx_3 \wedge dx_4$ and anti-self-dual Computed .

Example. $q = \cos a + i \sin a$ turns F^+ about the first axis by the angle $2a$, and returns to the start at $a = 180°$ ($q = -1$). The double cover of 1-02 is visible on F^+. We checked with 200 random elements that (5) holds, that the components mixing the two are at most 2.2×10^{-16}, and that under left multiplication F^- is fixed while the mirror $x_4 \mapsto -x_4$ swaps the two Computed .

Figure 1 Moving the left angle a turns only F⁺, and moving the right angle b turns only F⁻. "Reflect in the mirror" reflects the initial field by $x_4 \to -x_4$, swapping F⁺ and F⁻, before turning (the roles of a and b do not change). Lengths do not change

```
a = 0°, b = 0°, mirror no, F⁺ = (0.7071, 0.0000, 0.6364), F⁻ = (0.7071, 0.0000, 0.2121)
```

Computed The initial field is $F = dx_1 \wedge dx_2 + 0.6\, dx_1 \wedge dx_4 + 0.3\, dx_2 \wedge dx_3$. The figure computes (2)(4) on the spot, agreeing with the python3 values (including the rows with the mirror). The dotted lines are orbits.

The mirror and the Minkowski signature

A mirror changes the sign of the right-hand side of (5) (case 04). The mirror image of an instanton is an anti-instanton, and this is also why the θ term (the term added to the action by multiplying the integral of the right-hand side of (5) by the coefficient θ) changes sign under a mirror (3-05) Known . With the Minkowski signature, $** = -1$, so there are no real self-dual fields (One sign different); this is one of the reasons for thinking in imaginary time (4-02).

> 04

Changing the scale does not change the action — the dimension in which the coupling is dimensionless

Keeping the shape of the field and multiplying only its size by $1/\lambda$ gives $A_\lambda(x) = \lambda A(\lambda x)$, $F_\lambda(x) = \lambda^2 F(\lambda x)$ (a gauge field has the dimension of inverse length). Replacing $y = \lambda x$,

$$S_d[A_\lambda] = \frac{1}{2g^2}\int_{\mathbb{R}^d} \lambda^4 |F(\lambda x)|^2 \, d^d x = \lambda^{4-d}\, S_d[A] \quad (7)$$

so only in four dimensions does the action not depend on size. This means the classical theory has no preferred length, and it is another face of the $[g^2] = 4 - d$ of 2-04 Known . The problem statement's 'the mass gap is not classically visible' (Jaffe–Witten § 6.6) is also related to this (4-07).

One consequence. For a solution (stationary point) of finite action, the action does not change at first order under deformation of size either.

$$\frac{d}{d\lambda} S_d[A_\lambda]\Big|_{\lambda=1} = (4-d)\, S_d[A] = 0 \quad (8)$$

Hence if $d \neq 4$ then $F = 0$, and non-trivial solutions of finite action exist only in four dimensions (assuming enough decay for the deformation to be allowed; an argument of the same kind is attributed to Derrick 1964) Known .

Example. Seen on a lattice it is the same. On a n-dimensional lattice of spacing ϵ, the sum of quantities attached to the small squares (plaquettes) of the lattice plays the role of the action, and it is multiplied by the inverse coupling β (the coefficient in front of the action, corresponding to $1/g^2$; the constant factor depends on normalisation, and in Wilson's convention $\beta = 2N/g^2$) (5-01). The plaquette sum is approximated by $\frac{\epsilon^{4-n}}{4} S_{\mathrm{YM}}(A)$ (S_{YM} is the integral of the action of (7) with $1/2g^2$ removed). So classically β need only keep $\beta\,\epsilon^{4-n}$ constant to compensate for the shrinking of the sum as ϵ^{4-n} ($\beta \propto \epsilon^{n-4}$), and at $n = 4$ it does not move. In three dimensions $\beta \propto 1/\epsilon$, agreeing with $[g^2] = 1$. In physics the four-dimensional β is believed to move like $\log(1/\epsilon)$, but there are doubts too, and in mathematics it is unsolved, as Chatterjee (arXiv:1803.01950 § 3, 2018) writes Physics . The construction published in 2026-11 (6-00) includes in its statement that the ratio of the bare coupling of the chosen family of lattices to log(1/ε) approaches a fixed value (10-06; the committee's assessment is that it is proved in Lean; a statement about the one-loop leading part only). This running is asymptotic freedom (4-06).

Figure 2 Move the size ρ and the dimension d. On the left, a cross-section of the action density (half the action inside the dotted circle); on the right, S(ρ)/S(1) = ρ^(d−4) on a log–log plot, horizontal only for d = 4. "Play" moves ρ

back and forth

```
ρ = 1.000, d = 4, S(ρ)/S(1) = 1.000, [g²] = 0, central density 48/ρ⁴ = 48.00, total action 8π² ≈ 78.957
```

Computed The left is (9), the right the ρ^{d-4} of (7) ($\rho = 1/\lambda$), computed on the spot (agreeing with the ratios of integrals in python3). $d \neq 4$ is the ratio when the same shape is shrunk, not a solution.

05

Collapsing to a point — the borderline dimension

From here on we deal with fields whose values are in a Lie algebra, so we fix $|F|^2$. Write $F_{\mu\nu} = \sum_a F^a_{\mu\nu} t_a$ with $t_a = -iT_a$ ($\mathrm{tr}(t_a t_b) = -\frac{1}{2}\delta_{ab}$), the T_a of 1-04 multiplied by $-i$, and

$$|F|^2 = \sum_{\mu<\nu} \sum_a \left(F^a_{\mu\nu}\right)^2, \qquad S = \frac{1}{2g^2}\int |F|^2\, d^4x = -\frac{1}{2g^2}\int \sum_{\mu,\nu} \mathrm{tr}\, F_{\mu\nu} F_{\mu\nu}\, d^4x$$

Since t_a is anti-Hermitian, $\operatorname{tr} F^2$ is negative, so a minus sign is attached. (5) is also read as summed over each a.

The simplest self-dual solution (the BPST instanton, 3-05) has its action density gathered roundly at the origin. The action density $s = \frac{1}{2}|F|^2$ in units of $1/g^2$ and the fraction of the action inside radius R are as follows for the solution of size ρ ($d^4x = 2\pi^2 r^3 dr$; $2\pi^2$ is the area of the unit S^3) Known . The form of s was recomputed from the BPST connection with the normalisation above Computed .

$$s(x) = \frac{48\,\rho^4}{(|x|^2+\rho^2)^4}, \qquad \int_{\mathbb{R}4} s\,d^4x = 8\pi^2, \qquad \frac{1}{8\pi^2}\int_{|x|<R} s\,d^4x = 1 - \frac{3}{(1+u)^2} + \frac{2}{(1+u)^3}$$

Computed **Table 1** Numerical integration of (9). The total does not depend on ρ, and half always lies inside radius ρ.

ρ	Central density 48/ρ⁴	Total action / 8π²	Inside radius ρ	Radius containing 90%
4	0.1875	1.000000000000	0.500000	8.1065
1	48	1.000000000000	0.500000	2.0266
1/16	3.146 × 10⁶	1.000000000000	0.500000	0.1267

Example. As $\rho \to 0$, the density collapses to a point while keeping the action at $8\pi^2/g^2$. In three dimensions or fewer, shrinking increases the action (Figure 2), so concentration stops, but in four dimensions it does not stop. That handling this concentration is the crux of the analysis of classical Yang–Mills in four dimensions is said to have been known since Uhlenbeck's work of 1982 Known . § 2 of the problem statement writes 'Asymptotic freedom ensures a decisive regularity in the case when classical Sobolev inequalities are borderline.', stating flatly that asymptotic freedom "ensures" a decisive regularity. This chapter reads it as meaning that asymptotic freedom makes up for the classical borderline seen in (7) and Table 1 (estimates falling just short in four dimensions) (an interpretation).

06

The sphere at infinity and SU(2) — where integers come from

The far region of $\mathbb{R}^d$ is a large sphere S^{d-1}. If the action is finite and the decay sufficient, far away A approaches a form that can be written by a gauge transformation alone (pure gauge; 1-07), and the map of the gauge transformation remains (the same mechanism as the monopole of 1-07). To distinguish it from the coupling constant g, in this section we write that map as h. The classes of maps, with maps that can be continuously deformed into each other regarded as the same, form the homotopy group π_{d-1}.

$$h:\ S^3 \to SU(2) \cong S^3, \qquad k = \deg h \in \pi_3(S^3) = \mathbb{Z}, \qquad S \geq \frac{8\pi^2}{g^2}\,|k| \quad (10)$$

In four dimensions both the far region and SU(2) are S^3, and maps between spheres of the same dimension have an integer, "how many times it covers" (the degree). To be exact, choose one general point of the target S^3, attach $+1$ to each point of its preimage where h preserves orientation and -1 where it reverses it, and add them up (the same value whichever point is chosen). This is the instanton number k.

The inequality follows from (5). With the normalisation of section 05 one can write $k = \frac{1}{16\pi^2}\int (|F^+|^2 - |F^-|^2)\,d^4x$, so $S = \frac{1}{2g^2}\int (|F^+|^2 + |F^-|^2)\,d^4x \geq \frac{8\pi^2}{g^2}|k|$, with equality when self-dual or anti-self-dual. The BPST solution has $k = 1$ (3-05) Known .

Example. According to the tables of homotopy groups, the classes $\pi_{d-1}(S^3)$ into SU(2) = S^3 are $0, 0, \mathbb{Z}, \mathbb{Z}/2$ for $d = 2, 3, 4, 5$, and the classes into U(1) = S^1 are $\mathbb{Z}$ only for $d = 2$ (the number of vortices; the same mechanism as the monopole of 1-07). For a compact simple Lie group π_3 is always $\mathbb{Z}$, and SU(3) too has k Known .

07

Planes meet in a point — intersection forms and exotic $\mathbb{R}^4$

As in case 05, two planes in four dimensions usually meet in one point. Two closed surfaces in a closed four-dimensional manifold also meet in finitely many points, and the sum of ± 1 with orientations included does not change when the surfaces are moved. This is the intersection form (a symmetric matrix of integers).

$$\dim(\Sigma_1 \cap \Sigma_2) = 2 + 2 - 4 = 0, \qquad Q(\Sigma_1, \Sigma_2) = \sum_{p \in \Sigma_1 \cap \Sigma_2} (\pm 1) \quad (11)$$

Example. The intersection form of $\mathbb{CP}^2$ is (1), and that of $S^2 \times S^2$ is $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ Known . Using the space of solutions for anti-self-dual connections ($F^+ = 0$ in (6); the same as self-dual connections with the orientation reversed, and for the convention of the primary sources see 3-06) as a tool, Donaldson arrived at the theorem that "the definite intersection form of an oriented closed smooth four-dimensional manifold can be diagonalised". He is said to have shown the simply connected case in 1983 and later extended it to a form without the assumption of simple connectivity (details in 3-06). Contrasted with Freedman's topological theorem, spaces homeomorphic but not diffeomorphic to $\mathbb{R}^4$ (exotic $\mathbb{R}^4$) appear, and Taubes is credited with first showing that there are continuum-many of them Known . The main treatment is in 3-06.

Detour: knots and Wilson loops

A subsection you may skip.

A Wilson loop is the trace of the holonomy of a closed path (1-07), and in three-dimensional Chern–Simons theory its expectation value gives a knot invariant (the Jones polynomial) (Witten 1989) Known . In four dimensions every loop comes untied (case 03), and Yang–Mills is not topological either, so the expectation value is determined by the size and shape of the path, and in theories without matter fields an area law is taken as the sign of confinement Physics (5-02).

08

The 24-cell is a finite subgroup of SU(2)

Multiplying the vertices of the 24-cell (two coordinates ± 1, two 0) on the left by $(1+i)/\sqrt{2}$ and scaling by $1/\sqrt{2}$ gives the following 24 unit quaternions Computed . They form a finite subgroup of SU(2) closed under multiplication (the binary tetrahedral group; its image in SO(3) is the 12 rotations of the regular tetrahedron) Known .

$$\bar{T} = \left\{ \pm 1,\ \pm i,\ \pm j,\ \pm k,\ \tfrac{1}{2}(\pm 1 \pm i \pm j \pm k) \right\}, \qquad |\bar{T}| = 24 \ \text{(12)}$$

The 120 vertices of the 600-cell likewise form the binary icosahedral group $\bar{I}$.

Example. In lattice gauge theory (5-01) elements of a finite group can be placed on the links. Petcher–Weingarten 1980 computed with $\bar{T}$, $\bar{O}$, $\bar{I}$, and according to the abstract, for $\bar{I}$ the mean plaquette, Wilson loops and estimates of the string tension agreed with SU(2) over a wide range of couplings. At sufficiently weak coupling it is said to behave differently from SU(2) because of the gaps between elements Physics (sources and details in 5-10).

Computed **Table 2** All three checked to be closed under products. "Angle to the farthest point" is the maximum (an estimate) of the angle from 200,000 random points on S^3 to the nearest element.

Group	Order	Nearest-neighbour distance	Number of edges	Angle to the farthest point
$\bar{T}$ (24-cell)	24	1	96	44.7°
$\bar{O}$	48	$\sqrt{2-\sqrt{2}} \approx 0.7654$	144	31.2°
$\bar{I}$ (600-cell)	120	$1/\varphi \approx 0.6180$	720	22.2°

09

The area of the sphere, and integrals that begin to diverge in four dimensions

The volume of the ball in case 08 comes from the formula for the area of the unit sphere. This formula makes sense even when d is not an integer, and the integrals of the form first met in perturbation theory (4-03) also become functions of d.

$$|S^{d-1}| = \frac{2\pi^{d/2}}{\Gamma(d/2)}, \qquad I(d) = \int_{\mathbb{R}^d} \frac{d^d k}{(k^2+1)^2} = |S^{d-1}| \int_0^\infty \frac{k^{d-1} dk}{(k^2+1)^2} = \pi^{d/2}\,\Gamma\!\left(2-\frac{d}{2}\right) \quad (13)$$

Γ is the gamma function, the extension of the factorial to real numbers ($\Gamma(n) = (n-1)!$, $\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$, $\Gamma(x+1) = x\Gamma(x)$). The integrand is k^{-4} far away and the volume $k^{d-1}\,dk$, so it converges only for $d < 4$, and four dimensions is the borderline where divergence begins, logarithmically. In the dimension where the coupling is dimensionless, the divergence is exactly logarithmic (4-04). Computing at $d = 4-\epsilon$ (dimensional regularisation, 't Hooft–Veltman 1972, Bollini–Giambiagi 1972) and cutting off the momentum at K give

$$I(4-\epsilon) = \frac{2\pi^2}{\epsilon} + O(1), \qquad \int_{|k|<K} \frac{d^4 k}{(k^2+1)^2} = 2\pi^2 \ln K - \pi^2 + O(K^{-2}) \quad (14)$$

and $1/\epsilon$ and $\ln K$ correspond with the same coefficient $2\pi^2 = |S^3|$.

Computed **Table 3** The formula (13) and numerical integration.

d	I(d) formula	I(d) numerical integration	(4−d)·I(d)
2	3.141593	3.141593	6.283185
3	9.869604	9.869604	9.869604
3.99	1957.039052	–	19.570391
3.999	19722.225249	–	19.722225

Example. The last column approaches $2\pi^2 = 19.739209$, and for the cut-off, $I_K - 2\pi^2 \ln K = -9.869604$ at $K = 10^4$ ($-\pi^2$). The volumes of balls in the article ($V_5 = 5.2638$ and so on) also agree with the same formula Computed .

Using $\Gamma(x+1) = x\Gamma(x)$, $(4-d)\,I(d) = 2\pi^{d/2}\Gamma(3-\frac{d}{2})$, and the right-hand side is smooth across $d = 4$. The divergence is only in the factor $1/(4-d)$, and the remaining coefficient is $2\pi^2\Gamma(1) = 2\pi^2$ at $d = 4$. For the cut-off, $I_K = \pi^2\big(\ln(1+K^2) + \frac{1}{1+K^2} - 1\big)$, and the growth $2\pi^2 K^4/(K^2+1)^2$ for each increase of $\ln K$ by 1 approaches the same $2\pi^2$. Figure 3 sets these two side by side.

Figure 3 Bringing d towards 4 with the left slider, I(d) (dotted) diverges but (4−d)·I(d) (solid) approaches 2π². Enlarging the cut-off K with the right slider, the growth of I_K per ln K approaches the same 2π²

```
d = 3.0000, I(d) = 9.8696, (4−d)·I(d) = 9.8696 | K = 10.00, I_K = 35.7775, I_K − 2π² ln K =
-9.6737 (2π² = 19.7392)
```

Computed The left computes $I(d) = \pi^{d/2}\Gamma(2 - \frac{d}{2})$ and $(4-d)I(d) = 2\pi^{d/2}\Gamma(3 - \frac{d}{2})$, the right $2\pi^2 K^4/(K^2+1)^2$ and I_K, on the spot (the gamma function by Lanczos's approximation; agreeing with the python3 values). The left slider moves the u of $4 - d = 10^{-u}$.

10

The dimension slider — rows that line up at four and rows that do not

We line up the d-dependences scattered through the book on a single slider. Each row divides by the sign of a power $p(d)$, and at the borderline $p = 0$ the power vanishes, leaving a logarithm or invariance. The six rows of the four of spacetime reach their borderline at the same $d = 4$. The remaining four rows switch somewhere other than four, or for a different reason. Known To be exact, the six rows line up because of the same counting of dimensions (the 2 of the propagator $1/k^2$, twice), and this does not mean that construction in four dimensions is impossible.

Figure 4 Moving d changes the sentences of all eleven rows at once. Green = below, thick line = borderline, ochre = beyond, grey = a d the source does not treat

Computed The borderlines were computed on this machine; the sentences are copies from each chapter. If JavaScript is disabled, see Table 4.

Table 4 The rows of Figure 4.

Quantity (chapter)	Label	Borderline	Below (small d)	At the borderline	Beyond (large d)
$[g^2]$ and the type of divergence (4-04 Table 2)	four of spacetime	4	finitely many divergent diagrams	finitely many types, at all orders; the coupling runs logarithmically	types increase without limit
Trace coefficient 1−d/4 (3-03 (8))	four of spacetime (classical)	4	non-zero	identically 0 (no scale in the classical theory)	non-zero (opposite sign)
Instanton action ρ^(d−4) ((7), 3-05)	four of spacetime (classical)	4	spreading out reduces the action	independent of ρ	collapsing reduces the action
Wilson–Fisher, Σg_k^2 (4-04 (11), 7-05)	four of spacetime	4	fixed point $\hat{u}^*>0$, Σg_k^2 finite	$\hat{u}$ goes to 0 at a logarithmic rate, Σg_k^2 grows as the logarithm of the number of steps	$\hat{u}$ goes to 0 as a power
Room in stochastic quantisation (7-09 Table 2)	four of spacetime	4	positive room (locally subcritical)	room 0, the symbols do not fit into finitely many	negative room
Delta well, analogy with D=d−2 (4-07 Table 5)	four of spacetime (analogy)	4	D=1: g settles at 2	D=2: 1/g is proportional to ln K	D=3: g is tuned as a power
Variance at a point G_a(0) (1-09 Table 3)	contrast (two)	2	finite	diverges as ln(1/a)	diverges as a^(2−d)
Variance of the Wilson loop (7-06 Table 2)	contrast (three)	3	finite (area)	proportional to ln(1/a)	proportional to 1/a
Overlap of vacua, D=d−1 (4-09 § 07)	four of space	5	finite	diverges logarithmically (spacetime of five dimensions)	—

Quantity (chapter)	Label	`Borderline`	Below (small d)	At the borderline	Beyond (large d)
Complete families (5-11, Lean `no_perfectZ_of_five`)	four by coincidence	—	exist	only for integer d	do not exist
Steps of construction for YM (7-01 Table 2)	steps 1–6	—	d=2: done, done, done, done, partial, —; d=3: done, done (lattice), partial, not yet, not yet, not yet; d=4: done, done (lattice), done, partial, done, not yet (the third to fifth are by the construction published in 2026-11; the volume limit only along a subsequence. The four-dimensional row of 7-01 Table 2 is what the literature before 2026 had reached (done, done (lattice), partial, not yet, not yet, not yet), and here the published construction is laid over it. The committee's assessment is that it is proved in Lean; 6-01)		

What this chapter can say

Content	Status
(3)–(8), the swap under a mirror	Known . (8) assumes decay. Numerical checks are Computed (the surjectivity of (3) is not checked)
The running of $\log(1/\epsilon)$ for β on the lattice / the difference between finite groups and SU(2) at weak coupling	Physics . For the former, the source states plainly that it is unsolved in mathematics. The source for the latter is in 5-10
(9), Table 1, the normalisation of $\lvert F\rvert^2$ / the reading of the problem statement in section 05	The form of the density and the integrals are Computed (that it is a solution, 3-05) / this chapter's interpretation
(10)(11), the theorems of section 07	Known (original papers bibliographic data only)
(12)–(14), Tables 2 and 3, Figure 3	Computed (the covering angle is an estimate from samples)
Figure 4, Table 4	Copies from each chapter. The arrangement is this book's
The mass gap, the continuum limit	This chapter shows nothing

Sources and reproduction

Item	Kind	Source / tools
Tables 1–4, the normalisation of section 05, the reference values of Figures 1–3	Computed	python3 (NumPy, SciPy). The figures are JavaScript in the browser (checked by running the same functions with node)
The eight cases, the volume of the ball, the 24-cell	This site	What happens only in four dimensions
The lattice and β, quotations from the problem statement	Known	S. Chatterjee, arXiv:1803.01950 § 3 / A. Jaffe, E. Witten, Quantum Yang–Mills theory (the problem description of the Millennium Prize Problem of the Clay Mathematics Institute, PDF at claymath.org) § 2, § 6.6 (text checked for both)
Section 08	Known	D. Petcher, D. H. Weingarten, Phys. Rev. D 22 (1980) 2465 (abstract checked; the source for the difference at weak coupling is in 5-10)
(3), the π_{d-1} of section 06, the claims and attributions of section 07	Known	Wikipedia "Rotations in 4-dimensional Euclidean space", "Homotopy groups of spheres", "Donaldson's theorem" (which takes the 1983 version to assume simple connectivity and the extension to general fundamental groups to be the 1987 paper), "Exotic R4" (secondary sources)
Bibliographic data only (text not read)	Known	Conway–Smith, *On Quaternions and Octonions* (2003) / Belavin–Polyakov–Schwartz–Tyupkin, Phys. Lett. B 59 (1975) 85 / Derrick, J. Math. Phys. 5 (1964) 1252 / Uhlenbeck, Commun. Math. Phys. 83 (1982) 11, 31 / Donaldson, J. Differential Geom. 18 (1983) 279 / Donaldson, ibid. 26 (1987) 397 / Freedman, ibid. 17 (1982) 357 / Taubes, ibid. 25 (1987) 363 / Witten, Commun. Math. Phys. 121 (1989) 351 / 't Hooft–Veltman, Nucl. Phys. B 44 (1972) 189 / Bollini–Giambiagi, Nuovo Cim. B 12 (1972) 20

Next chapter: 2-01 Lagrangians and the action. It continues in 3-05, 3-06, 4-07 and 7-10.