Volume 1 Preparation

Part 3 Yang–Mills theory (classical)

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
3-01	The gauge principle	2
3-02	Non-abelian gauge fields	11
3-03	The action and the equations	24
3-04	Gauge theory as geometry	36
3-05	Instantons and topology	47
3-06	Detour: Donaldson and exotic $\mathbb{R}^4$	59
3-07	Yang–Mills inside the Standard Model	70

The gauge principle — demanding a local symmetry requires a tool for comparison, and that tool carries the force

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

What this chapter shows — That demanding a (local) symmetry that turns the phase point by point breaks the derivative, and that a tool for comparing neighbouring points (a connection) repairs it. That this tool becomes a field, couples to the conserved charge, and carries a force. That for SU(2) the gauge field itself is also turned.

Prerequisite chapters — 2-03 (gauge transformations, phase factors) and 1-04 (SU(2), the adjoint representation). Beyond these we use only Noether's theorem from 2-01 and the indices of 2-04 § 07. We take $\hbar = c = 1$, $\varepsilon_0 = 1$ (Heaviside–Lorentz units).

Said first — The gauge principle is a guide for choosing the form of a theory, not a theorem (section 07). The non-abelian field strength and self-interaction are left to 3-02. Only classical fields are treated, and this chapter shows nothing about the mass gap of Yang–Mills.

Order of this chapter

1. Global symmetry — charge is conserved
2. Turning locally — the derivative breaks
3. A tool for comparison — the connection and the covariant derivative (Figure 1, Table 1)
4. Making A a dynamical field — the field strength, and the mass term that cannot be written
5. A force appears — coupling to the conserved charge (Figure 2)
6. The same demand for SU(2) — a field that is itself turned, and the problem of mass
7. What the principle decides, and what it does not (Table 2)
8. Detour: "gauge" was originally a measuring rod
9. Where this book is going

01

Global symmetry — turning everything by the same amount

In 2-03 the potentials $\mathbf{A}$ came first, and the replacement of phase (G') was found afterwards. This chapter goes the other way: it derives the electromagnetic field by demanding that "the theory does not change even if the phase is turned arbitrarily point by point". This procedure is the **gauge principle**.

The starting point is a field $\psi(t, \mathbf{x})$ with complex values, the simplest model of a charged particle. Take the Lagrangian density (2-01, 2-02) to be

$$\mathcal{L}_0 = |\partial_t \psi|^2 - |\nabla \psi|^2 - V(|\psi|^2) \quad (1)$$

(for example $V = M^2|\psi|^2$, where M is the mass of the particle). Turning by a constant α as $\psi \mapsto e^{i\alpha}\psi$ leaves $\mathcal{L}_0$ invariant (a **global U(1) symmetry**; U(1) is the group of complex numbers of absolute value 1, 1-02). Noether's theorem (2-01) gives a conservation law, written with a constant q multiplied in.

$$\rho = iq\big(\psi^*\partial_t\psi - \psi\,\partial_t\psi^*\big), \qquad \mathbf{J} = -iq\big(\psi^*\nabla\psi - \psi\nabla\psi^*\big), \qquad \partial_t\rho + \nabla\cdot\mathbf{J} = 0 \ (2)$$

Example. For a plane wave $\psi = e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)}$ ($\omega > 0$), $\rho = 2q\omega$, $\mathbf{J} = 2q\mathbf{k}$. q becomes the charge in section 05. A global symmetry says that "nothing changes if the reference of phase at every point is shifted all at once".

02

Turning locally breaks the derivative

Let the angle of turning change from point to point, $\psi \mapsto e^{i\alpha(t,\mathbf{x})}\psi$ (a **local** transformation). $|\psi|^2$ does not change, but the derivative acquires an extra term (∂_μ are the components of ∂_t and ∇, $\mu = 0,1,2,3$).

$$\partial_\mu\big(e^{i\alpha}\psi\big) = e^{i\alpha}\big(\partial_\mu\psi + i(\partial_\mu\alpha)\,\psi\big)$$

Example. The gradient energy $|\nabla\psi|^2$ of a constant field $\psi = 1$ is 0, but turning by $\alpha = kx$ gives $\psi = e^{ikx}$, $|\nabla\psi|^2 = k^2$. Only the reference was changed, yet (1) counts it as a different state.

The derivative is the limit of the difference $\psi(x+\varepsilon n) - \psi(x)$ of two neighbouring points divided by ε (n is the unit vector of the direction). If the two values are turned by different angles, the difference mixes "the change in the field" with "the change in the reference". Values at different points cannot be subtracted as they stand.

Why demand local transformations? The abstract of Yang and Mills (1954) says that the usual invariance under isospin rotations (section 06) is not consistent with the idea of localised fields (the content of the abstract is in 3-02 § 01) Known . This is a demand (a principle), not a theorem, and whether nature follows it is decided by experiment (section 07).

03

A tool for comparison — the connection and the covariant derivative

To compare values at different points, we need a correspondence: "which reference of phase at the point y corresponds to the reference at the point x". We call this the **comparator** $U(y,x) \in U(1)$, write $U(y,x)\psi(x)$ for $\psi(x)$ carried to y, and stipulate that under transformations

$$U(y,x) \mapsto e^{i\alpha(y)}\,U(y,x)\,e^{-i\alpha(x)} \ (3)$$

Carrying in the opposite direction is the inverse element, $U(x,y) = U(y,x)^{-1}$. Then the difference $\psi(y) - U(y,x)\psi(x)$ is merely multiplied by $e^{i\alpha(y)}$. Carry before comparing: that is the point.

For two nearby points the comparator is close to 1, and we name the rate of its deviation the field A_μ: $U(x+\varepsilon n, x) = \exp\big(iq\varepsilon\, n^\mu A_\mu(x)\big) + O(\varepsilon^2)$. The opposite direction is its inverse, so $U(x, x+\varepsilon n) = \exp\big(-iq\varepsilon\, n^\mu A_\mu(x)\big) + O(\varepsilon^2)$. Carrying the value at $x+\varepsilon n$ to x with this before subtracting, and taking the limit of the difference quotient with n in the direction μ,

$$D_\mu\psi = \lim_{\varepsilon\to 0}\frac{U(x,x+\varepsilon n)\,\psi(x+\varepsilon n) - \psi(x)}{\varepsilon} = \big(\partial_\mu - iqA_\mu\big)\psi \ (4)$$

This is the **covariant derivative**. Expanding (3) in $y = x + \varepsilon n$ determines how A_μ changes.

$$\psi \mapsto e^{i\alpha}\psi, \qquad A_\mu \mapsto A_\mu + \frac{1}{q}\,\partial_\mu \alpha, \qquad D_\mu\psi \mapsto e^{i\alpha} D_\mu \psi \quad (5)$$

Example. Putting $A_0 = -\phi$, $(A_1, A_2, A_3) = \mathbf{A}$ (the notation of 2-04), $\alpha = q\chi$, (5) becomes $\mathbf{A} \mapsto \mathbf{A} + \nabla\chi$, $\phi \mapsto \phi - \partial_t \chi$, $\psi \mapsto e^{iq\chi}\psi$. These are the gauge-transformation formula of 2-03 and the phase transformation of the wave function (G'). From the demand we have arrived at the electromagnetic potentials and their gauge transformations. Multiplying comparators along a path gives the phase factor of 2-03 § 05, in geometric language parallel transport (1-07 § 03).

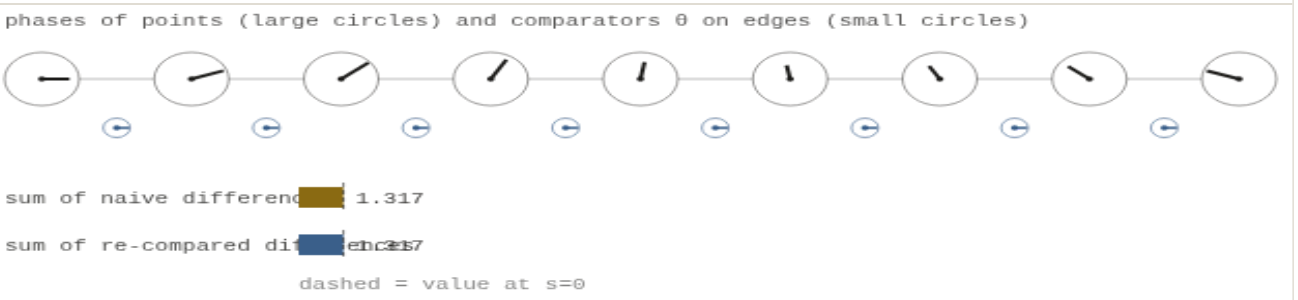
Seen on a chain

If space is cut into a sequence of points $j = 0, 1, 2, \ldots$, the comparator becomes a $U_j = U(j{+}1, j) = e^{i\theta_j}$ on each edge (2-03 § 08, the link variables of 5-01), and instead of the derivative we use the **re-compared difference**. (3) becomes $\theta_j \mapsto \theta_j + \alpha_{j+1} - \alpha_j$, and

$$e^{-i\theta'_j}\psi'_{j+1} - \psi'_j = e^{i\alpha_j}\left(e^{-i\theta_j}\psi_{j+1} - \psi_j\right)$$

The sum of squared absolute values does not change, while the sum of naive differences $\psi_{j+1} - \psi_j$ does.

Figure 1 Changing s turns the phase at each point (large circles). With "Connection: on" the comparators on the edges (small circles) also turn by (3), and the sum of re-compared differences does not change. With "Connection: off" it changes



Computed The numbers in the figure are computed on the spot. Point by point with the connection on, moving s from 0 to 1 in steps of 0.25, the sum of naive differences is 1.317, 8.070, 19.536, 22.541, 17.700, while the sum of re-compared differences stays 1.317 (with the connection off both are the former; turning everything by the same amount, both are 1.317).

Computed **Table 1** A ring of 12 points (random field and comparators) before and after transformation. W is the product of the comparators once round (SU(2) is in section 06), and W' is the value after the transformation. "Max. entry of $W' - W$" is the maximum absolute value of the entries of the difference matrix.

Group	Turning	Sum of naive differences before → after	Sum of re-compared differences before → after	Max. entry of $W' - W$
U(1)	point by point	65.474 → 50.856	83.239 → 83.239	0
SU(2)	same everywhere	82.514 → 82.514	75.704 → 75.704	1.39
SU(2)	point by point	82.514 → 85.310	75.704 → 75.704	0.239

The change in the sum of re-compared differences is within 3×10^{-14} in every case (rounding error). Turning U(1) by the same amount everywhere, no column changes.

> 04

Making A a dynamical field — the field strength, and the term that cannot be written

To make A_μ a field obeying its own equation, we add to the action a term containing derivatives of A that does not change under (5). The material is the comparator of a closed path, in which, by (3), all the α cancel. Going from x in the direction μ, then in the direction ν, once round a square of side ε, and back to x,

$$U_{\square} = \exp\!\big(iq\varepsilon^2 F_{\mu\nu}\big) + O(\varepsilon^3), \qquad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \quad (6)$$

The same quantity also appears in $[D_\mu, D_\nu]\psi = -iqF_{\mu\nu}\psi$. The comparator carries from start to end, like the parallel transport of 1-07, and rewriting in the geometric notation $A_{\text{geom}} = -iqA$, $F_{\text{geom}} = -iqF$, (6) becomes $\exp(-\varepsilon^2 F_{\text{geom}})$, agreeing with (10) of 1-07 and $U_p \approx \exp(-a^2 F_{\mu\nu})$ down to the sign (in 5-01, which reads the lattice links in the opposite direction, the sign of the exponent is reversed). **Example.** The components are $F_{0i} = -E_i$, $F_{ij} = \varepsilon_{ijk} B_k$ (2-04), and F is the electric and magnetic fields themselves. The electric and magnetic fields were the **curvature** of the comparator (the deviation after going round a small loop, 1-07 § 05). Listing the invariant terms with the lowest order of derivatives,

$$\mathcal{L} = |D_t\psi|^2 - |\mathbf{D}\psi|^2 - V(|\psi|^2) + \tfrac{1}{2}\big(|\mathbf{E}|^2 - |\mathbf{B}|^2\big), \qquad D_t = \partial_t + iq\phi, \;\; \mathbf{D} = \nabla - iq\mathbf{A} \quad (7)$$

The last term is $-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}$ ((8) of 2-04). The way of coupling that merely replaces the ∂ of (1) by D is called **minimal coupling**.

The term that cannot be written. The term $-\frac{1}{2}m^2 A_\mu A^\mu = \frac{1}{2}m^2(\phi^2 - |\mathbf{A}|^2)$ that would give A a mass changes under (5) by terms in $-\frac{m^2}{q}\big(\phi\,\partial_t\alpha + \mathbf{A}\cdot\nabla\alpha\big)$ and the second order of α. A mass term cannot be written for a field that comes from the gauge principle (as in 2-03 § 04) Known . This affects the distance a force reaches in section 05, and becomes a difficulty in section 06.

05

A force appears — coupling to the conserved charge

Expanding (7) in q, the quantities of section 01 appear as they are.

$$\mathcal{L} = \mathcal{L}_0 - \phi\,\rho + \mathbf{A}\cdot\mathbf{J} + q^2\big(\phi^2 - |\mathbf{A}|^2\big)|\psi|^2 + \tfrac{1}{2}\big(|\mathbf{E}|^2 - |\mathbf{B}|^2\big) \quad (8)$$

The gauge field couples to the conserved quantity (2) of the global symmetry, with the same q that appeared in the transformation. This is exactly the form of (L) of 2-03.

Varying (8) with respect to ϕ gives Gauss's law.

$$\nabla\cdot\mathbf{E} = \rho - 2q^2\phi\,|\psi|^2 = iq\big(\psi^* D_t\psi - \psi\,(D_t\psi)^*\big)$$

The right side is not the ρ of (2) itself but that with its ∂_t replaced by D_t (terms containing ϕ are added). For a point charge Q at rest at the origin, we put this right side equal to $Q\,\delta^3(\mathbf{x})$. In the static case $\mathbf{E} = -\nabla\phi$, and from $-\nabla^2\phi = Q\,\delta^3(\mathbf{x})$, $\phi = Q\,G_0(r)$. The energy of a charge Q' at distance r is

$$\mathcal{U}(r) = QQ'\,G_m(r), \qquad G_0(r) = \frac{1}{4\pi r}, \qquad G_m(r) = \frac{e^{-mr}}{4\pi r} \;\; \text{(with a mass term added provisionally)} \quad (9)$$

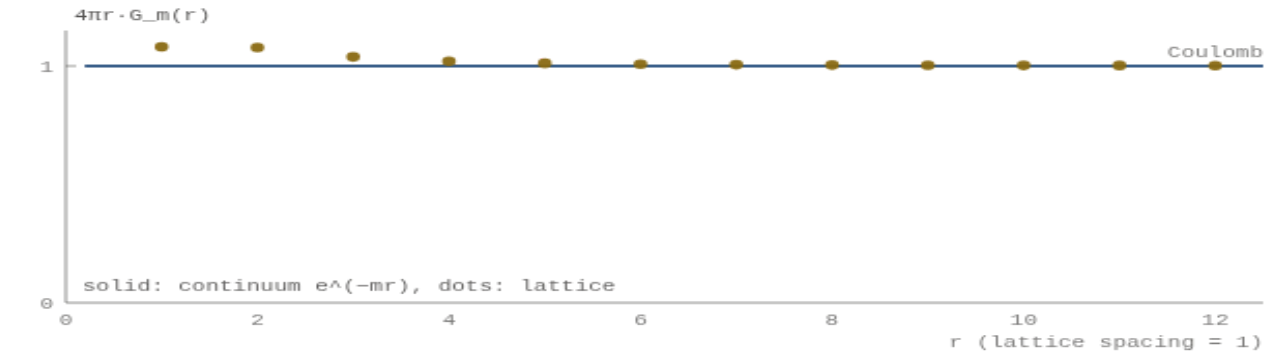
Example. For $m = 0$, charges of the same sign repel, and the force reaches far away like $1/r^2$. Merely demanding a local symmetry has produced a force between distant charges. In the theory (7) (when the minimum of V is at $\psi = 0$), adding the forbidden mass term provisionally would give $(-\nabla^2 + m^2)\phi = \rho$,

and G_m (Yukawa type) does not reach much further than about $1/m$. Without it, the force is of Coulomb type and reaches far.

However, gauge invariance does not guarantee that a force reaches far. Even classically, with a Higgs field the distance reached becomes finite while gauge invariance is kept (3-07) Known . In terms of (8), when the minimum of V is at $|\psi|^2 = v^2 > 0$ and $|\psi|^2$ stays there, the term $q^2(\phi^2 - |\mathbf{A}|^2)|\psi|^2$ has the same form as the mass term $\frac{1}{2}m^2(\phi^2 - |\mathbf{A}|^2)$ of $m^2 = 2q^2v^2$ ((9) of 3-07 puts the minimum of the size of the field at $v/\sqrt{2}$, so the same mass is written $m = qv$).

Figure 2 is $(-\Delta_{\text{lat}} + m^2)G_m = \delta$ found numerically on a cubic lattice (spacing 1) ($\Delta_{\text{lat}} f$ is the sum of the values at the six neighbours minus $6f$).

Figure 2 Changing m changes 4πr·G_m(r) (1 for Coulomb). The solid line is the continuous e^(−mr), the dots the cubic lattice. m > 0 is the case with a mass term added by hand (not gauge invariant)



m = 0.00, rough reach 1/m = ∞, r = 4: lattice 1.020, continuum 1.000

Computed The dots are computations on this machine, the solid line is computed on the spot. Static values: for $r = 1,\ 4,\ 12$ the lattice $4\pi r\, G_0$ is 1.082, 1.020, 1.002, approaching Coulomb further than a few lattice spacings away. $4\pi r\, G_{0.5}$ is 0.670, 0.147, 0.0027 (the continuous $e^{-0.5r}$ is 0.607, 0.135, 0.0025), and the rate of decay along the lattice axis is $\mu = 0.4949$, from $\cosh\mu = 1 + m^2/2$.

06

The same demand for SU(2)

The proton and the neutron have almost equal masses and feel the nuclear force in almost the same way. The isospin view gathers the two into a two-component field $\psi = (\psi_1, \psi_2)$, and the theory of the form (1) does not change under the mixing rotations $\psi \mapsto \Omega\psi$ ($\Omega \in SU(2)$, 1-04; to distinguish it from the V of (1), the matrix is written Ω, as in 3-02). Yang–Mills (1954) imposed the demand of section 02 on this symmetry.

Making the comparator an SU(2) matrix with $U(y,x) \mapsto \Omega(y)\,U(y,x)\,\Omega(x)^{-1}$, and for two nearby points $U(x + \varepsilon n, x) = \exp(ig\varepsilon\, n^\mu A_\mu) + O(\varepsilon^2)$, $A_\mu = A_\mu^a T^a$ ($T^a = \sigma^a/2$; relative to the geometric notation of 1-07, $A_{\text{geom}} = -igA = -i\hat{A}$; $\hat{A} = gA$ is the "coupling absorbed" notation of 3-02 § 02),

$$D_\mu = \partial_\mu - igA_\mu, \qquad A_\mu \mapsto \Omega A_\mu \Omega^{-1} - \frac{i}{g}(\partial_\mu\Omega)\Omega^{-1}, \qquad D_\mu\psi \mapsto \Omega\, D_\mu\psi \quad (10)$$

Example. For $\Omega = e^{i\alpha}$ (1×1) we return to (5) (g playing the role of q), and in geometric notation this is the $A \mapsto \Omega A\Omega^{-1} - (d\Omega)\Omega^{-1}$ of 1-07 § 07 (1-07 writes the transformation matrix as g, but here g is the coupling constant). There are three differences from U(1).

1. **The number of fields.** As many as the dimension of the Lie algebra are needed: 1 for U(1) (the photon), 3 for SU(2) (A^1, A^2, A^3), 8 for SU(3) (gluons), $N^2 - 1$ for SU(N).

2. **The gauge field itself is turned.** The $\Omega A \Omega^{-1}$ of (10) turns A even when Ω is constant (the adjoint representation, 1-04 § 06). Unlike the photon, it carries itself the "charge" it couples to. This is why in Table 1 the W of SU(2) changes as $\Omega_0 W \Omega_0^{-1}$ even under a global rotation, and only $\operatorname{tr} W$ remains; self-interaction comes from here (3-02).
3. **There is one coupling constant.** For U(1) a different q can be put freely for each field. For a simple non-abelian group (1-03 § 08) there is only one coupling constant g, and the strength of the coupling to each matter field is determined by g and the generators T^a of that field's representation ($\sigma^a/2$ for a two-component field, $(T^a)_{bc} = -i\varepsilon_{abc}$ for a three-component field in the adjoint representation) Known . The outline of the reason is that the non-linear term of the field strength $-ig[A_\mu, A_\nu]$ (section 09, 3-02) ties the normalisation of A to g and fixes it, and that if the group is simple the invariant form of the kinetic term is unique up to a constant multiple (2-04 § 07, 1-03).

The problem of mass. For SU(2) too, no mass term can be written for the gauge field. In Yang's recollection, at a seminar in Princeton in late February 1954, Pauli asked "what is the mass of this field?" (the story is in the detour of 3-02). The nuclear force is short-range, which does not fit the Coulomb-type reach of section 05. § 1 of the Jaffe–Witten problem statement also writes that the absence of mass in the classical waves was a serious obstacle Known , and says it was overcome for the weak force by adding a Higgs field (3-07), and for the strong force by a property of the quantum theory itself (asymptotic freedom, 4-06) Physics . Showing as mathematics the mass gap expected in the latter (4-11) is the task of the problem statement.

07

What the principle decides, and what it does not

The gauge principle is a strong constraint, but it does not determine the theory uniquely (**Table 2**).

What it decides	What it does not decide
That a connection A is needed, and its transformation laws (5)(10)	Which group (choosing U(1), SU(2), SU(3) is up to experiment)
That derivatives can enter only through D, and that the current coupled to is conserved	The ratios of charges for U(1) (if the group is a circle, the n of the representation $e^{i\alpha} \mapsto e^{in\alpha}$ is an integer and charges are in integer ratios; whether it is a circle or the line $\mathbb{R}$ lies outside the principle)
That no mass term can be written for the gauge field	That the kinetic term is F^2 (F^4 is also invariant; the reason only low-dimension terms are kept is in 4-04)
That configurations carried into each other by gauge transformations represent the same physics	Choosing minimal coupling (merely replacing ∂ by D). The principle does not forbid non-minimal ways of coupling ($\lvert\psi\rvert^2 F_{\mu\nu}F^{\mu\nu}$ is also invariant)

An example for the second row of the table. If the group is a circle, the charge of every field is an integer multiple of a single unit. If the group is the line $\mathbb{R}$, the ratios of charges can be any real numbers, and the principle alone gives no reason for charges to line up as integer multiples of one unit. An example for the third row. A term like F^4 is of higher order in the fields than F^2, and in four dimensions its coefficient has the dimension of length to the fourth power, so it is thought to have less effect the longer the distance (2-04 § 08, 4-04 § 06) Physics .

A local "symmetry" differs from an ordinary symmetry in producing a constraint (Gauss's law) and a redundancy of description (2-01 § 08). How to divide out the redundancy is the subject of 3-04 and 4-08.

There is one more difference from an ordinary symmetry. A symmetry is **spontaneously broken** when the action (in statistical mechanics, the energy) is symmetric but the equilibrium state (the expectation values) is not. An example is the spontaneous magnetisation of 2-08 § 04. Elitzur (1975) argued that in lattice gauge theory without gauge fixing a local symmetry is not spontaneously broken (the expectation values of

quantities that are not gauge invariant are 0), and showed it concretely in simple systems on abelian lattices Known .

08

Detour: "gauge" was originally a measuring rod

A section you may skip.

The words "gauge transformation" and "gauge invariance" first appear in physics in Weyl's 1918 paper attempting to derive gravity and electromagnetism from a single geometry, where they meant changing the standard of length (calibration; the German Eich-, the verb eichen) (O'Raifeartaigh–Straumann § I) Known . Weyl voiced his dissatisfaction with Riemannian geometry, in which the magnitudes of vectors at two distant points can be compared directly (English translation, ibid. § II).

> A true infinitesimal geometry should, however, recognize only a principle for transferring the magnitude of a vector to an infinitesimally close point ...

(A truly infinitesimal geometry should recognise only a principle for transferring the magnitude of a vector to an infinitely near point.) This is the same idea as the comparator of section 03, except that what is carried is length, not phase. Weyl's theory has a real-valued field w_μ expressing at what rate the standard of length changes from point to point, and carrying a vector along a path γ from a point p to a point q changes its length ℓ as

$$\ell(q) = \exp\Big(-\int_\gamma w_\mu \, dx^\mu\Big)\, \ell(p)$$

((2.2) of O'Raifeartaigh–Straumann; the sign depends on conventions). Going once round a closed path and back, the length is multiplied by a real factor $\exp\big(-\oint w_\mu \, dx^\mu\big)$. As with (6), in a region without holes this factor is 1 for every loop when $\partial_\mu w_\nu - \partial_\nu w_\mu = 0$, and Weyl regarded w_μ as the electromagnetic potential (ibid. § II).

Einstein objected at once: the frequency of an atomic clock would change with the path it had followed, contradicting the stability of atomic spectra Known .

When quantum mechanics appeared, London (1927) realised that making the exponent imaginary turns it into a change of the phase of the wave function (that invariance had been found by Fock (1926)). In his 1928 book and 1929 paper Weyl raised this as "the principle of gauge invariance", and wrote that the electromagnetic field necessarily accompanies the wave field of matter (Jackson–Okun § III.B) Known . The name stayed, derived as it was from the calibration of length. As a metaphor it is "a theory in which the measuring rod can be chosen freely at each point", but precisely, what can be chosen at each point is not a scale of length but a complex phase (in general, a group element). The first non-abelian model was Klein (1938), but it was forgotten, and today's line is held to start from Yang–Mills (1954) (ibid. § V).

09

Where this book is going

In 3-02 the square of (6) is turned with SU(N). Because comparators depend on the order of multiplication, a commutator is added to the field strength, and self-interaction appears.

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - ig[A_\mu, A_\nu]$$

Imposing Gauss's law in the quantum theory is 4-01. On the lattice of 5-01 the comparators U_j are the variables, and the action is built from the products around the squares of (6) (plaquettes). That g becomes dimensionless in four dimensions is in 2-04 § 08 and the four-dimensions article.

To sum up: the demand for a local symmetry calls in a gauge field, that field couples to the conserved charge and carries a force, and classically its mass term cannot be written by hand (adding a Higgs field

changes matters; 3-07). The problem statement asks to show that, in the four-dimensional non-abelian case, a mass gap nevertheless arises in the quantum theory (0-03).

What this chapter can say

Item	What is said	What is not said
Sections 01–05: the covariant derivative (4), the transformation law (5), the field strength (6), that the mass term is not invariant, coupling to the conserved current, the two forms of (9)	Standard derivation for classical fields Known . The numbers in Table 1 and Figures 1 and 2 Computed	The lattice values are for spacing 1; the continuum limit is not taken
The distance a force reaches	In the theory (7) (minimum of V at $\psi = 0$), that without a mass term the force is of Coulomb type and reaches far Known	That the force reaches far does not follow from gauge invariance alone. The content of the Higgs mechanism is in 3-07
Section 06: the SU(2) transformation law (10), the number of fields, being turned in the adjoint, that there is one coupling constant	(10), the number of fields, the turning in the adjoint Known . Only an outline of why there is one coupling constant (2-04 § 07, 1-03)	The formula for self-interaction (3-02)
The problem of mass	That no mass term can be written by hand for the classical gauge field Known . How it was overcome for the strong force is Physics	The mass gap of the quantum theory (unsolved; this chapter shows nothing)
Section 07: the standing of the principle, Elitzur's result	That the principle is a guide for choosing the form of a theory. Elitzur's result at the level of the abstract (argued, and shown in simple systems on abelian lattices)	Deriving the principle as a theorem. The content of Elitzur's argument for general groups (the text has not been checked)
The history in section 08	What is written in the two reviews Known	The texts of the original papers (bibliographic data only)

Sources and reproduction

Item	Kind	Source / tools
Table 1, Figure 1	Computed	python3, NumPy (`covariant_diff.py`). The figure is computed on the spot
The dots of Figure 2, the values in section 05	Computed	python3, SciPy (`lattice_green.py`: $G_m(r) = \int_0^\infty e^{-(m^2+6)t} I_r(2t) I_0(2t)^2 dt$, modified Bessel functions; reproduces the identity $G_0(0) - G_0(1) = 1/6$ to 10 digits). Cross-check by FFT, `check_fft.py`
Motivation, Pauli's question	Known	C. N. Yang, R. L. Mills, Phys. Rev. 96 (1954) 191 (abstract checked). C. N. Yang, Selected Papers 1945–1980 with Commentary (1983) p. 525 (bibliographic data only; checked through the quotation in O'Raifeartaigh–Straumann § VII)
The obstacle of massless classical waves, and overcoming it by the Higgs field and asymptotic freedom	Known	A. Jaffe, E. Witten, Quantum Yang–Mills Theory (the Clay problem statement) § 1 (text checked)
Local symmetries are not spontaneously broken (Elitzur)	Known	S. Elitzur, Phys. Rev. D 12 (1975) 3978 (abstract checked)

Item	Kind	Source / tools
The history in section 08	Known	J. D. Jackson, L. B. Okun, Rev. Mod. Phys. 73 (2001) 663, arXiv:hep-ph/0012061 (§ III.B and § V checked). L. O'Raifeartaigh, N. Straumann, Rev. Mod. Phys. 72 (2000) 1, arXiv:hep-ph/9810524 (§ I, § II, § V and § VII of the arXiv version checked). Original papers bibliographic data only (from the reference lists of the two reviews): H. Weyl, Sitzungsber. Akad. Wiss. Berlin (1918) 465 / V. Fock, Z. Phys. 39 (1926) 226 / F. London, Z. Phys. 42 (1927) 375 / H. Weyl, Gruppentheorie und Quantenmechanik (Hirzel, Leipzig, 1928) / H. Weyl, Z. Phys. 56 (1929) 330 / O. Klein, in New Theories in Physics (conference in Warsaw, 1938)

Next chapter: 3-02 Non-abelian gauge fields.

Non-abelian gauge fields — when the order of multiplication matters, the field becomes its own source

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

What this chapter shows — Allowing the $\mathrm{SU}(N)$ transformation to be chosen point by point determines the covariant derivative and the transformation law, and a commutator term is added to the field strength F. Because of that term the gauge field carries its own charge, and cubic and quartic self-interactions and non-linear equations arise. For fields uniform in space, the commutator alone makes the x^2y^2 valley.

Prerequisite chapters — 3-01 (the gauge principle; especially the SU(2) case of § 06). We use the structure constants of 1-03, the adjoint representation of 1-04, and 2-04 (raising and lowering indices, and the signature of the metric). The same content written in the language of geometry is 1-07.

Said first — This is a classical chapter, treating neither quantum theory nor the mass gap. The action and the equations of motion are left to 3-03, and the energy density is used first by analogy with electromagnetism. Pauli's letters and Yang's recollection in § 01 are quoted from reviews.

Order of this chapter

1. Turning at each point — from the SU(2) of 3-01 to SU(N); the abstract of Yang and Mills. Detour: the spring of 1954
2. Conventions — generators, and translating between three notations
3. The covariant derivative — the two terms of the transformation law; seen on a chain (Figure 1)
4. The field strength — the commutator of covariant derivatives; pure gauge
5. Colour electric and colour magnetic fields — recounting the measurable quantities

6. Self-interaction — quadratic, cubic and quartic terms
7. Fields uniform in space — the x^2y^2 valley (Figure 2); the classical theory has no scale
8. The problem of mass — why it cannot be added by hand

01

Turning at each point — the question of Yang and Mills

In 3-01 § 06, we imposed the demand that turning point by point be allowed on the symmetry (isospin) that mixes, by SU(2) matrices, the two-component field ψ combining the proton and the neutron, and obtained the covariant derivative and the transformation law (10). This chapter extends that to $\mathrm{SU}(N)$ and follows in turn what arises from the commutator.

The abstract of the paper of Yang and Mills (Phys. Rev. 1954) points out that the usual invariance, turning by the same matrix at every point, is not consistent with the idea of localised fields, and continues: “The possibility is explored of having invariance under local isotopic spin rotations.” Rephrasing the first sentence of the abstract in this book's reading: that the choice of which direction to call "proton" at one point should fix the naming at distant points as well does not fit the laws of fields, which are written in terms of relations between nearby points. So we allow the transformation matrix $\Omega = \Omega(x)$ to be chosen point by point.

According to the abstract, the consequence is a **b** field that plays the same role for isospin as the electromagnetic field, obeys non-linear differential equations, and whose quanta are particles of spin 1, isospin 1 and charge $\pm e$ or 0 Known . § 03– § 06 turn this summary into formulas. The group carrying the strong force later turned out to be SU(3) (Jaffe–Witten's problem description: “the gauge group is G = SU(3)”). However, what was gauged was not isospin but SU(3) of colour, a different degree of freedom (isospin remains an approximate symmetry, not turned point by point). Mathematically, one just replaces the group (3-07, the detour of 1-04).

Detour: the spring of 1954

A subsection you may skip.

According to an essay by Straumann, Pauli had obtained the same field strength from the framework of extra dimensions (17-06), in notes attached to a letter to Pais in July 1953 and in two seminars in Zürich in November. But in a letter to Pais in December he wrote “one will always obtain vector mesons with rest mass zero”, and did not publish. Yang recalls that at a

seminar in Princeton in late February 1954, just after Yang had written $(\partial_\mu - i\epsilon B_\mu)\psi$ on the blackboard, Pauli asked "What is the mass of this field B_μ?" (quoted from the commentary in Yang's selected papers by Straumann's essay and the review of O'Raifeartaigh–Straumann). The same review mentions that the theory was found independently by other authors, and places Shaw (1954–55) and Utiyama (1954–55, 1956) in its family-tree figure Known . The question of mass is answered in § 08.

02

Conventions — generators, and three notations

In Part 3, $\hbar = c = 1$, and the signature of the metric is $(-,+,+,+)$ (2-04). $\mu, \nu = 0, \dots, 3$ are spacetime indices, $i, j, k = 1, 2, 3$ spatial, and a, b, c indices in the direction of the group (colour); indices appearing twice are summed. The group is $G = \mathrm{SU}(N)$, the generators are Hermitian T^a $(a = 1, \dots, N^2 - 1)$, and the structure constants and normalisation are as follows (1-04).

$$[T^a, T^b] = i f^{abc} T^c, \qquad \mathrm{tr}(T^a T^b) = \tfrac{1}{2}\delta^{ab}$$

Example. For SU(2), $T^a = \sigma^a/2$ (half the Pauli matrices) with $f^{abc} = \varepsilon^{abc}$; for SU(3), $T^a = \lambda^a/2$ (half the Gell-Mann matrices). The gauge transformation matrix is written $\Omega(x) \in G$ (the same letter as in 3-01), and the letter g is kept for the coupling constant. A matter field $\psi(x) \in \mathbb{C}^N$ is taken to change as $\psi \mapsto \Omega(x)\psi$ (the fundamental representation).

This chapter follows the physicists' style, writing the Hermitian $A_\mu = A_\mu^a T^a$ and the coupling g separately. The ϵ and B_μ on Yang's blackboard in the detour of § 01 correspond to this g and A_μ (the conventions for coefficients depend on the paper). The other notations in the book can be translated by the following lines.

Notation	Covariant derivative	Field strength	Chapter
This chapter	$D_\mu = \partial_\mu - igA_\mu$	$F_{\mu\nu}$ (formula (6))	3-02, Part 4
Geometric (anti-Hermitian)	$D = d + \mathcal{A}$, $\mathcal{A} = -igA$	$\mathcal{F} = d\mathcal{A} + \mathcal{A}\wedge\mathcal{A} = -igF$	1-07, 3-04
Coupling absorbed	$D_\mu = \partial_\mu - i\hat{A}_\mu$, $\hat{A} = gA$	$\hat{F} = gF$	2-04

03

The covariant derivative — the two terms of the transformation law

We extend (10) of 3-01 to $\mathrm{SU}(N)$ and follow the derivation again. The trouble is the derivative. If Ω depends on the point, the derivative of the product produces an extra

term.

$$\partial_\mu(\Omega\psi) = \Omega\,\partial_\mu\psi + (\partial_\mu\Omega)\,\psi \quad (1)$$

The second term appears because values at neighbouring points with different frames were subtracted as they stand (1-07 § 01). So, as in 3-01, we demand that the derivative with a correcting field A_μ added change in the same way as ψ.

$$D_\mu = \partial_\mu - igA_\mu, \qquad D'_\mu(\Omega\psi) = \Omega\,D_\mu\psi \quad (2)$$

The left side is $\Omega\partial_\mu\psi + (\partial_\mu\Omega)\psi - igA'_\mu\Omega\psi$, the right side $\Omega\partial_\mu\psi - ig\,\Omega A_\mu\psi$. Equality for every ψ determines A'_μ.

$$A'_\mu = \Omega A_\mu \Omega^{-1} - \frac{i}{g}\,(\partial_\mu\Omega)\,\Omega^{-1} \quad (3)$$

The first term is the rotation by the adjoint representation (1-04 § 06), and the second is an inhomogeneous term coming from the slope of Ω. Differentiating $\Omega\Omega^\dagger = 1$, $(\partial\Omega)\Omega^{-1}$ is anti-Hermitian, and from $\det\Omega = 1$ its trace is 0, so A' also stays in the same Lie algebra. Example. For $\Omega = e^{i\alpha(x)}$ (U(1)) the first term stays A_μ, and $A'_\mu = A_\mu + \partial_\mu\alpha/g$, putting $\chi = \alpha/g$, is exactly the gauge-transformation formula of 2-03 (with the convention $A_\mu = (-\phi, \mathbf{A})$ of 2-04). What is new in the non-abelian case is the first term, which sandwiches A.

Infinitesimal transformations — the gauge field itself turns

Putting $\Omega = 1 + i\alpha^a T^a$ (α small) and expanding (3) to first order, in components

$$\delta A^a_\mu = \frac{1}{g}\,\partial_\mu\alpha^a + f^{abc}A^b_\mu\,\alpha^c = \frac{1}{g}\,(D_\mu\alpha)^a \quad (4)$$

The $(D_\mu\phi)^a = \partial_\mu\phi^a + gf^{abc}A^b_\mu\phi^c$ at the right end is the covariant derivative acting on fields in the adjoint representation. Example. For SU(2), taking α constant with $\alpha^3 = \theta$ (others 0), $\delta A^1 = \theta A^2$, $\delta A^2 = -\theta A^1$, $\delta A^3 = 0$. Even under a rotation independent of position the components of the gauge field turn into each other, so **the gauge field itself carries non-abelian charge** (this is "the gauge field itself is turned" of 3-01 § 06 seen in components). This is in contrast to the photon, which carries no charge, and it is what "isospin 1" in the abstract means.

Seen on a chain

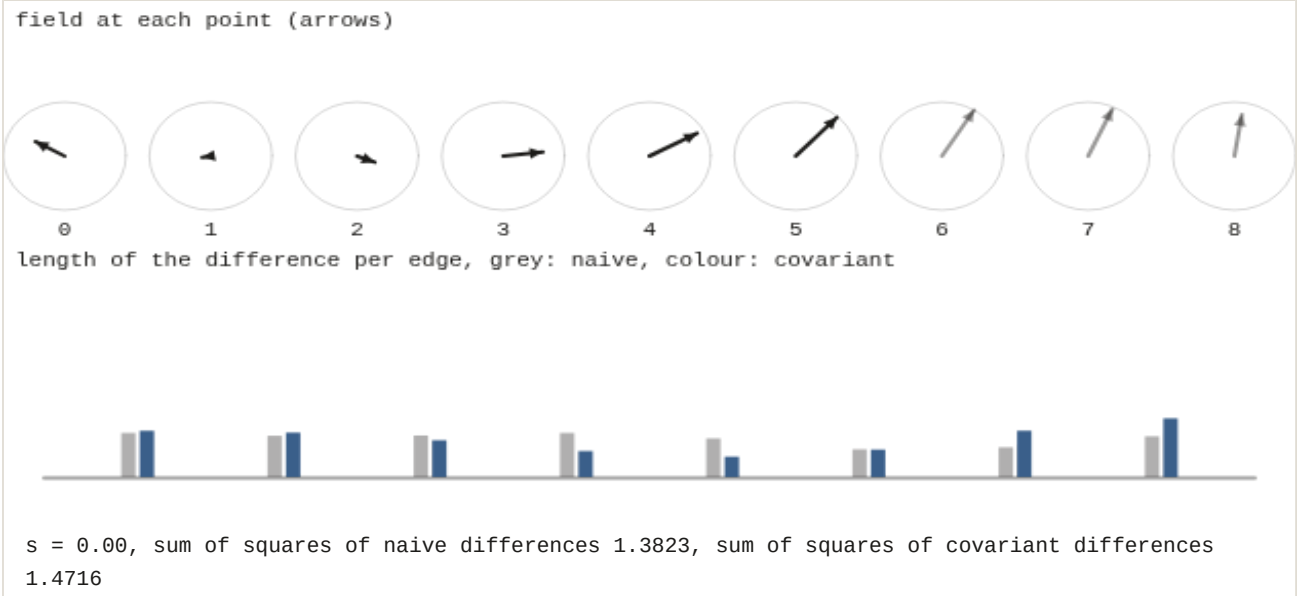
Arranging points in a row makes the meaning of (2) visible. At each point n put a field in the adjoint representation (a real three-component vector) ϕ_n, and on each edge put

the rotation U_n that carries the value at the point $n+1$ into the frame of the point n. This is the non-abelian version of the U(1) elements on the edges of 2-03, and a component of the lattice of 5-01. Written with the comparator of 3-01 it is $U_n = U(n, n{+}1)$, the opposite direction from the $U_j = U(j{+}1, j)$ of the chain of 3-01. The Ω_n and U_n here denote, by the same letters, the rotations of $\mathbf{3} \times \mathbf{3}$ given by SU(2) elements in the adjoint representation (1-04 § 06).

$$\phi'_n = \Omega_n \phi_n, \qquad U'_n = \Omega_n U_n \Omega^{-1}_{n+1}, \qquad U'_n \phi'_{n+1} - \phi'_n = \Omega_n \big(U_n \phi_{n+1} - \phi_n \big) \quad (5)$$

The "covariant difference" at the right end is multiplied by only one rotation, so its length does not change. The naive difference $\phi'_{n+1} - \phi'_n$ has different rotations at the two points, so its length changes. With the spacing of the points h, putting $U_n \approx 1 - igh A$ and letting $h \to 0$, the second formula of (5) becomes (3), and the covariant difference becomes $h D_\mu \phi$. Here A acts as $A = A^a T^a_{\text{adj}}$ using the matrices $(T^a_{\text{adj}})_{bc} = -i f^{abc}$ of the adjoint representation, and the components of $D_\mu \phi$ have the same form as the right end of (4).

Figure 1 The field at nine points (arrows above) and the length of the difference on each edge (bars below). Changing the strength of the gauge transformation with the slider s, the grey bars (naive differences) change, and the coloured bars (covariant differences) do not.



Computed The numbers in this figure were computed on this machine (the same formulas and the same configuration as Table 1, computed in the browser). If JavaScript is disabled, Table 1 is the content of the figure.

Table 1 The strength of the gauge transformation s, and the sums of squares of the differences on the eight edges

s	Naive difference	Covariant difference
0.00	1.3823	1.4716
0.25	3.0524	1.4716

s	Naive difference	Covariant difference
0.50	7.6091	1.4716
0.75	13.5267	1.4716
1.00	18.6570	1.4716

The change in the covariant difference on each edge was at most 3×10^{-16} (rounding error) Computed . That the two differ at $s = 0$ is because U_n is taken to be a small rotation.

04

The field strength — the commutator of covariant derivatives

In electromagnetism $\partial_\mu A_\nu - \partial_\nu A_\mu$ was a quantity unchanged by gauge transformations. In the non-abelian case, because of the first term of (3), this combination is not even covariant. The correct combination comes from the difference of covariant derivatives multiplied in the two orders.

$$[D_\mu, D_\nu]\,\psi = -ig\,F_{\mu\nu}\,\psi, \qquad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu - ig\,[A_\mu, A_\nu] \tag{6}$$

Expanding the left side, the terms in $\partial_\mu\partial_\nu\psi$ and $A\,\partial\psi$ are symmetric in $\mu\nu$ and cancel, so no derivatives of ψ remain. $F_{\mu\nu}$ is a matrix (an element of the Lie algebra) at each point, and in components it is as follows.

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g\,f^{abc} A^b_\mu A^c_\nu \tag{7}$$

For SU(2), regarding the three components $(A^1_\mu, A^2_\mu, A^3_\mu)$ as a vector in the internal space, the last term is the cross product $g\,A_\mu \times A_\nu$. (2) means $D'_\mu = \Omega D_\mu \Omega^{-1}$ as an operator, so the commutator is sandwiched in the same way.

$$F'_{\mu\nu} = \Omega\,F_{\mu\nu}\,\Omega^{-1} \tag{8}$$

F is not invariant; it turns at each point by the adjoint representation (it is covariant). This is a large difference from electromagnetism, where the electric and magnetic fields were themselves measurable quantities.

Example 1 (pure gauge). The $A_\mu = -\frac{i}{g}(\partial_\mu\Omega)\Omega^{-1}$ obtained by transforming $A = 0$ by (3) has $F = 0$, by (8). Conversely, if $F = 0$ on a simply connected region such as $\mathbb{R}^4$, then A can be written in this form Known . The notes Pauli attached to his letter to Pais of July 1953, as presented by Straumann, state this as “The vanishing of the field strength is necessary and sufficient for the $A^a_\mu(x)$ in the whole space to be

transformable to zero." It fails on regions with holes, and the Aharonov–Bohm effect of 2-03 is an example (the global story is in 3-04).

Example 2 (commutator only). For constant fields $A_x = a\,T^1$, $A_y = b\,T^2$, $F_{xy} = -ig\,ab\,[T^1, T^2] = g\,ab\,T^3 \neq 0$. In electromagnetism a constant potential makes no field. A figure showing this as the deviation after going round a small square is Figure 2 of 1-07.

Computed Giving smooth A and Ω for SU(2) and SU(3) and comparing by finite differences, the difference between the two sides of (8) was 5×10^{-9} (the error of step 10^{-4}); dropping the commutator gave 0.78, 1.74; and the F of a pure gauge was at most 10^{-9}.

05

Colour electric and colour magnetic fields — recounting the measurable quantities

Following electromagnetism, we define electric and magnetic fields for each direction of the group. With the same assignment as $F_{0i} = -E_i$, $F_{ij} = \varepsilon_{ijk} B_k$ of 2-04, the three spatial components are written in bold ($\mathbf{A}^a = (A^a_1, A^a_2, A^a_3)$).

$$E^a_i = F^a_{i0}, \qquad B^a_i = \tfrac{1}{2}\varepsilon_{ijk} F^a_{jk} = (\nabla \times \mathbf{A}^a)_i + \tfrac{g}{2}\,\varepsilon_{ijk} f^{abc} A^b_j A^c_k \tag{9}$$

Taking $A_0 = 0$ (temporal gauge), $E^a_i = -\partial_t A^a_i$. The second term of the magnetic field is the commutator part, and a magnetic field arises even if the potential is spatially uniform.

By (8), $\mathbf{E}^a$ and $\mathbf{B}^a$ turn in the colour index at each point. So "component 3 of the colour electric field" changes value with the gauge and is not a measurable quantity. What does not change are squared lengths such as $\sum_a |\mathbf{E}^a|^2 = 2\,\mathrm{tr}(\mathbf{E}\cdot\mathbf{E})$. Following electromagnetism, in this chapter we take the following as the energy density (deriving it from the action is in 3-03).

$$\mathcal{E} = \tfrac{1}{2}\sum_a \left(|\mathbf{E}^a|^2 + |\mathbf{B}^a|^2\right) \tag{10}$$

Example. Typical quantities that do not change with the gauge are $\mathrm{tr}(F_{\mu\nu}F^{\mu\nu})$ and the trace of parallel transport around a closed path (the Wilson loop, 1-07 § 06).

06

Self-interaction — quadratic, cubic and quartic terms

Both the energy (10) and the $F^a_{\mu\nu}F^{a\,\mu\nu}$ put in the action in 3-03 are squares of F. Putting in (7), they split into terms quadratic, cubic and quartic in A.

$$\begin{aligned} F^a_{\mu\nu}F^{a\,\mu\nu} = {} & (\partial_\mu A^a_\nu - \partial_\nu A^a_\mu)(\partial^\mu A^{a\,\nu} - \partial^\nu A^{a\,\mu}) \\ & + 4g\, f^{abc}\,(\partial_\mu A^a_\nu)\, A^{b\,\mu} A^{c\,\nu} \\ & + g^2 f^{abe} f^{cde}\, A^a_\mu A^b_\nu A^{c\,\mu} A^{d\,\nu} \end{aligned} \qquad (11)$$

The first line is Maxwell's $F_{\mu\nu}F^{\mu\nu}$ repeated for the number of colours N^2-1, and on its own it is a collection of Maxwell fields that do not act on each other. The second and third lines contain f^{abc} and vanish for U(1). Put in the action (3-03), they are interactions among gauge fields, and in perturbation theory (4-03) they become vertices where three or four gauge fields meet. In the classical Maxwell equations light does not scatter off light, but the Yang–Mills field scatters off itself.

The consequence is the non-linearity of the equations. Jaffe–Witten write "The Yang–Mills equations are nonlinear—in contrast to the Maxwell equations." Known . Adding two solutions does not in general give a solution. Computed For random A^a_μ, $\partial_\mu A^a_\nu$ the two sides of (11) agreed to 10 digits. The non-zero f^{abc}, counted as ordered, number 6 for SU(2) and 54 for SU(3).

Table 2 Classical differences between electromagnetism and Yang–Mills

	Electromagnetism U(1)	**Yang–Mills SU(N)**
Number of gauge fields	1	N^2-1 (8 for SU(3))
Field strength	Invariant	Covariant (8)
Charge of the gauge field	None	Carries it (4)
Constant potential	Makes no field	Makes a field if the commutator is not 0
Equations	Linear	Non-linear (11)
Classical waves	Mass 0	Mass 0 (§ 08)

07

Fields uniform in space — a valley made by the commutator alone

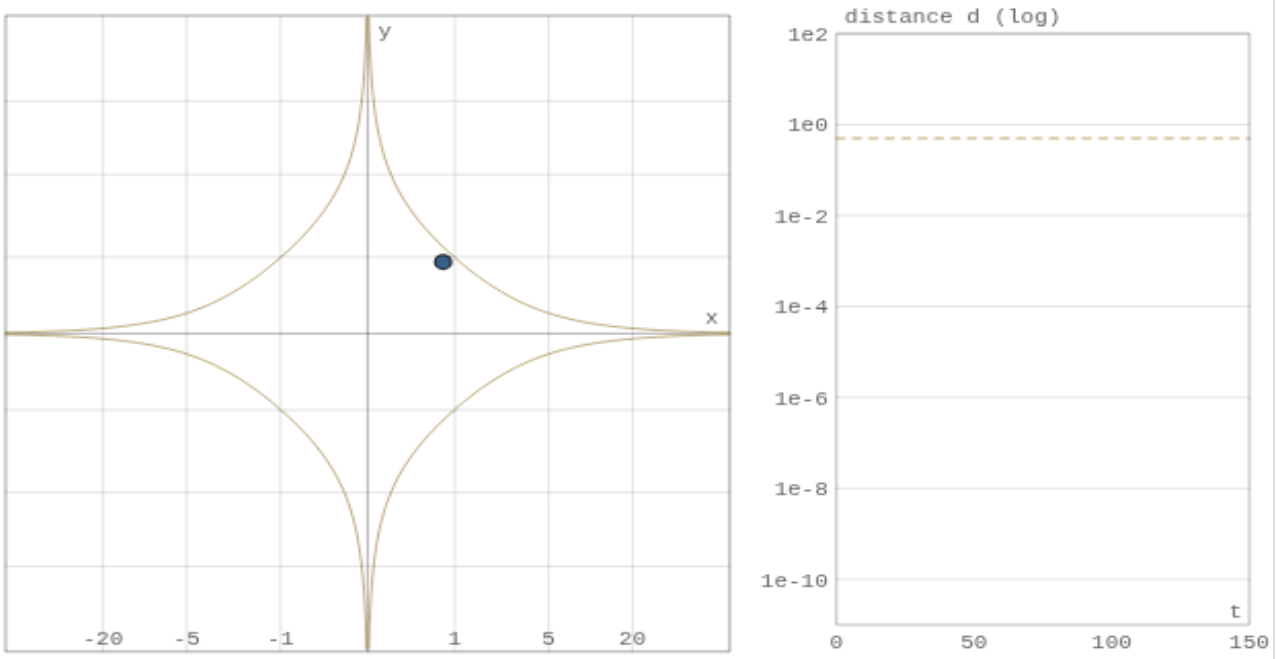
We look at self-interaction in its smallest form. For SU(2) with $A_0 = 0$, and A^a_i functions of time only, the magnetic field of (9) consists of the commutator term alone. Further putting $A^1_1 = x(t)$, $A^2_2 = y(t)$ and the rest 0, the non-zero fields are $E^1_1 = -\dot{x}$, $E^2_2 = -\dot{y}$, $B^3_3 = g\,xy$, and (10) becomes the energy of a particle in the plane.

$$H = \tfrac{1}{2}(\dot{x}^2 + \dot{y}^2) + \tfrac{1}{2} g^2 x^2 y^2, \qquad \ddot{x} = -g^2 x y^2, \quad \ddot{y} = -g^2 x^2 y \tag{12}$$

For uniform fields, the Yang–Mills equations of motion (3-03) become the equations of a particle with 9 components A_i^a moving in $\frac{1}{2}\sum (B_i^a)^2$, and this assignment is consistent with them Computed (by numerical differentiation, the force on the other 7 components is 0, and the Gauss constraint $\varepsilon^{abc} A_i^b E_i^c = 0$ also holds). The Gauss constraint is the condition left by the equation for A_0 lost in the temporal gauge, the non-abelian version of $\nabla \cdot \mathbf{E} = 0$ in electromagnetism. For uniform fields the derivative terms vanish, giving this form (4-01). For comparison, putting the two components in the same colour direction ($A_1^1 = x$, $A_2^1 = y$), the commutator is 0 and $V = 0$, and the particle moves in a straight line, as in electromagnetism.

The non-abelian $V = \frac{1}{2} g^2 x^2 y^2$ is 0 on the coordinate axes, so the valleys go on for ever, narrowing the further in (the width available at energy E is $|y| \le \sqrt{2E}/(g|x|)$). Figure 2 runs two particles at the same time with velocities differing by δ.

Figure 2 Left: orbits of (12) ($g = 1$, initial values $(x, y, \dot{x}, \dot{y}) = (0.8, 0.9, 0.6, -0.3)$, $E = 0.4842$; the scale compresses distant parts, and the thin curve is the contour $V = E$). Right: the distance d between the two particles (vertical axis logarithmic). The slider k changes the initial offset $\delta = 10^{-k}$, and "Make abelian" switches to $V = 0$.



t = 0.00, d = 0.0e+0, time exceeding 0.5 —, drift in E 0e+0 (non-abelian, δ = 1e-6)

Computed The numbers in this figure were computed on this machine (the same integration rule as Table 3: fourth-order Runge–Kutta, step $\min(0.01,\ 0.02/(g\max(|x|, |y|)))$, distance checked every 0.05). If JavaScript is disabled, Table 3 is the content of the figure.

Table 3 The initial offset δ, and the time at which the distance exceeds 0.5

δ	non-abelian	Abelian ($0.5/\delta$)

δ	non-abelian	Abelian ($0.5/\delta$)
1e-2	7.45	50
1e-4	25.05	5e3
1e-6	37.90	5e5
1e-8	54.55	5e7
1e-10	77.15	5e9

In the non-abelian case, each time δ is made 1/10, the separation time grows by only about 8 (a straight-line fit to the 9 points 10^{-2}–10^{-10} gives slope 8.04 (Table 3 shows 5 of them, one per decade); the increments are uneven because separation slows while the orbit is deep in a valley). This means the distance grows exponentially, and on the right it rises almost as a straight line. In the abelian case the distance is $\delta\, t$. The drift of the energy was at most 5×10^{-9} at $t \le 120$ Computed . This system has been studied as classical Yang–Mills mechanics. Savvidy (1983) argued that dynamics of this kind is a K-system (a dynamical system with strong mixing), but Dahlqvist–Russberg (1990) found a family of stable periodic orbits in this x^2y^2 system, refuting earlier claims that the motion is completely ergodic (according to the abstract). What the computation on this machine shows goes only as far as the separation visible in finite time for the initial values above.

The classical theory has no scale

(12) has no constant that could serve as a unit of length or time (g only changes the scale of time). If $x(t), y(t)$ is a solution, then for any $\lambda > 0$

$$x_\lambda(t) = \lambda\, x(\lambda t), \qquad y_\lambda(t) = \lambda\, y(\lambda t), \qquad E_\lambda = \lambda^4 E \quad \text{(13)}$$

is also a solution (substituting, λ^3 appears on both sides). Example. Making the energy 1/16, the motion of the same shape happens twice as slowly. Bringing the energy towards 0, the motion becomes as slow as we like, and the frequencies have no lower bound. Computed Solving the initial values $\lambda = 0.5,\ 0.25,\ 0.125$ separately with the same rule and dividing the position at time $5/\lambda$ by λ, all agreed with $(-0.401421,\ 0.616689)$ to 6 digits.

The same holds for the whole field: keeping only the quadratic terms of (11) around $A = 0$ gives a collection of Maxwell fields, and the frequencies $\omega = |\mathbf{k}|$ continue down to 0. Jaffe–Witten write that the classical equations describe “massless waves that travel at the speed of light” Known . Quantising, deep in the valley the lowest energy of the transverse oscillation $g|x|/2$ acts as a wall, and the spectrum of the operator obtained by quantising (12) is discrete, with a positive lowest eigenvalue Known (Simon 1983, the

detour of 2-05). This is a theorem about a model with finitely many degrees of freedom keeping only uniform fields.

That quantum Yang–Mills as a field theory has a gap is widely accepted in physics, but has not become a theorem Physics .

08

The problem of mass — why it cannot be added by hand

We answer Pauli's question in the detour of § 01 with formulas. For a vector field of mass m, the usual thing is to add $-\frac{1}{2}m^2 A^a_\mu A^{a\,\mu}$ to the Lagrangian density (for the electromagnetic field, the Proca field). But computing the change by (4),

$$\delta\left(A^a_\mu A^{a\,\mu}\right) = \frac{2}{g}\, A^{a\,\mu}\partial_\mu \alpha^a + 2 f^{abc} A^{a\,\mu} A^b_\mu\, \alpha^c = \frac{2}{g}\, A^{a\,\mu}\partial_\mu \alpha^a \quad (14)$$

The second term vanishes because f^{abc} is antisymmetric in a, b while $A^{a\,\mu} A^b_\mu$ is symmetric. The remaining first term comes from the inhomogeneous term of (3), and is not 0 if α depends on the point. The mass term breaks the demand that turning at each point be allowed Known . Computed For random A and small α, the two sides of (14) agreed to 4 digits or more for both SU(2) and SU(3).

So as long as gauge invariance is kept, the waves of the classical Yang–Mills field have no mass. Jaffe–Witten write “The massless nature of classical Yang–Mills waves was a serious obstacle to applying Yang–Mills theory to the other forces”. That the way past it differed for the weak force (the Higgs field, 3-07) and the strong force (asymptotic freedom, 4-06) we saw in 3-01 § 06.

As the first of the properties a theory of the strong force must have at the quantum level, Jaffe–Witten list the mass gap: “there must be some constant $\Delta > 0$ such that every excitation of the vacuum has energy at least Δ” (0-03, 4-11). The classical field has no scale (§ 07), and the gap is thought to arise only on quantisation Physics . Showing existence and the mass gap as theorems for pure Yang–Mills theory is what the problem statement asks, and this chapter shows nothing about it.

What this chapter can say

Item	What is said	What is not said
(2)–(8), (11)	Standard derivations, and numerical checks for SU(2) and SU(3) Known Computed	The equations of motion and the Bianchi identity (3-03), the geometry of the space of connections (3-04)

Item	What is said	What is not said
(12) the x^2y^2 system of uniform fields	Consistency of the assignment, and the exponential separation visible in finite time for these initial values Computed	A proof of irregularity, the value of the Lyapunov exponent (the exponent of the rate at which nearby orbits separate)
(13) that the classical theory has no scale	Exact scaling and numerical checks Computed	A proof, as a field theory, that quantisation produces a scale
(14) the mass term	That it breaks local gauge invariance Known	The content of the Higgs mechanism (3-07)
The mass gap	—	That quantum Yang–Mills theory has a gap (unsolved; this site shows nothing) Physics
The history in § 01	What is written in the abstract, the reviews and the problem description	The text of Yang–Mills, the original wording of Pauli's letters, the content of Shaw's work

Sources and reproduction

Item	Kind	Source / tools
The quotation and summary of the abstract in § 01	Primary source (abstract checked; text not read)	C. N. Yang, R. L. Mills, “Conservation of Isotopic Spin and Isotopic Gauge Invariance”, Phys. Rev. 96 (1954) 191 (received 28 June 1954)
Pauli's 1953 work, the phrase from his letter, the quotation from his notes, Yang's recollection	Review (text checked)	N. Straumann, “On Pauli's invention of non-abelian Kaluza-Klein Theory in 1953”, arXiv:gr-qc/0012054 (Yang's recollection is cited with C. N. Yang, Selected Papers p. 525 as the source). The collected letters (W. Pauli, Wissenschaftlicher Briefwechsel IV/II, Springer 1999) bibliographic data only
The recollection of the February 1954 seminar, Shaw and Utiyama	Review (text checked)	L. O'Raifeartaigh, N. Straumann, “Early History of Gauge Theories and Kaluza-Klein Theories, with a Glance at Recent Developments”, arXiv:hep-ph/9810524 § VII and Figure 1 (published title “Gauge theory: Historical origins and some modern developments”, Rev. Mod. Phys. 72 (2000) 1–23). The original source, C. N. Yang, Selected Papers 1945–1980 with Commentary (1983) p. 525, bibliographic data only
Bibliographic data of Utiyama (1956)	Primary source (abstract checked)	R. Utiyama, “Invariant Theoretical Interpretation of Interaction”, Phys. Rev. 101 (1956) 1597
The quotations on SU(3), non-linearity, massless waves, the serious obstacle and the mass gap	Primary source (text checked)	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory”, the official problem description of the Clay Mathematics Institute, § 1

Item	Kind	Source / tools
The x^2y^2 system: the claim that it is a K-system	Bibliographic data only (checked with Crossref)	Savvidy, “The Yang-Mills classical mechanics as a Kolmogorov K-system”, Phys. Lett. B 130 (1983) 303–307
The x^2y^2 system: a family of stable periodic orbits	Primary source (abstract checked)	P. Dahlqvist, G. Russberg, “Existence of stable orbits in the x^2y^2 potential”, Phys. Rev. Lett. 65 (1990) 2837–2838
Discrete spectrum of the quantum x^2y^2 system	Known (bibliographic data only)	B. Simon, “Some quantum operators with discrete spectrum but classically continuous spectrum”, Ann. Phys. 146 (1983) 209–220 (reference [43] in § 6.6 of Jaffe–Witten, cited as an example in which bound states arise even for potentials with flat directions)
Derivations of (1)–(14)	Known	Standard material. The translation of notation is the table in § 02
Table 1, Figure 1, the numbers in § 04, § 06 and § 08	Computed on this machine	Python 3.12, NumPy 1.26, SciPy 1.11 (`chain_covariance.py`, `formula_checks.py`). Figure 1 computes the same formulas in plain JavaScript
Consistency in § 07, Table 3, the check of (13), Figure 2	Computed on this machine	Python 3.12 (`ym_mechanics.py`). Figure 2 integrates with the same rule in plain JavaScript. Differences in rounding are amplified by irregular motion, so when t goes well beyond 100 the orbits of the figure and the table can diverge

Next chapter: 3-03 The action and the equations ($\mathrm{tr}\, F^2$, the Bianchi identity, the equations of motion, self-duality). The geometry continues in 3-04.

The action and the equations — from one action come the equations, an identity, the energy and self-duality

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

What this chapter shows — Varying the action gives the equation $D_\mu F^{\mu\nu} = 0$, and the way F is built gives the Bianchi identity. In four dimensions the trace of the stress is 0, and it follows that there is no static lump of finite energy. The action in imaginary time is at least $8\pi^2|Q|/g^2$, with equality for self-dual fields.

Prerequisite chapters — 3-02 (the covariant derivative, the field strength, the colour electric and magnetic fields). For variations, 2-01; for the Maxwell action, 2-03 and 2-04; for the Hodge star, 1-06 and 1-12.

Said first — This is a classical chapter, treating neither quantum theory nor the mass gap. § 06 is a known fact about classical fields, and a different matter from whether the quantum theory has a gap. That Q is an integer, and the construction of the BPST solution, are left to 3-05.

Order of this chapter

1. Conventions and the action — why $\operatorname{tr} F^2$ is chosen
2. The equations of motion — the field becomes its own source
3. Writing it with the colour electric and magnetic fields — Gauss's law is a constraint
4. The Bianchi identity — the other half of the equations
5. Energy and momentum — the trace vanishes in four dimensions
6. There is no static lump — Detour: no assumption on the rate of decay is needed, the spherically symmetric field (Figure 1), Detour: a solution of infinite energy
7. The action in imaginary time and self-duality — a lower bound on the action
8. A stretched instanton — Q is unchanged and the action grows (Figure 2)

9. What the classical theory says, and what it does not

01

Conventions and the action — why tr F² is chosen

The conventions are the same as in 3-02 § 02: $\hbar = c = 1$, the metric $(-,+,+,+)$, Hermitian T^a ($\operatorname{tr} T^a T^b = \frac{1}{2}\delta^{ab}$), and $D_\mu = \partial_\mu - igA_\mu$. On a field X with values in the Lie algebra, $D_\mu X = \partial_\mu X - ig[A_\mu, X]$ acts (in components, $\partial_\mu X^a + g f^{abc} A^b_\mu X^c$), and the colour electric and magnetic fields are $E^a_i = F^a_{i0}$ and $B^a_i = \frac{1}{2}\varepsilon_{ijk} F^a_{jk}$ (3-02 § 05). The action replaces the F of the Maxwell action ((8) of 2-04) with the non-abelian field strength and takes the trace.

$$S[A] = -\frac{1}{2}\int \operatorname{tr}\left(F_{\mu\nu}F^{\mu\nu}\right) d^4x = -\frac{1}{4}\int F^a_{\mu\nu} F^{a\,\mu\nu}\, d^4x = \frac{1}{2}\int \sum_a \left(|\mathbf{E}^a|^2 - |\mathbf{B}^a|^2\right) d^4x \tag{1}$$

The form is almost fixed by four requirements: gauge invariance (the rotation by $F' = \Omega F \Omega^{-1}$ disappears under the trace), Lorentz invariance, equations that close at second order in time derivatives, and no length scale in the coefficients of the action (only the terms with mass dimension $[F^2] = 4$; counting dimensions in 2-04 § 08). The last condition is needed because adding a higher-order invariant in F such as $(\operatorname{tr} F^2)^2$ still leaves the equations second order. Its coefficient needs the fourth power of a length, which brings in a scale. Imposing all four, and if the group is simple, the quadratic invariants in F are only linear combinations of (1) and the θ term (3-05), which does not change the equations (2-04 § 07) Known . Example. For the uniform field $A^1_1 = x(t)$, $A^2_2 = y(t)$ (3-02 § 07) the action per volume is $L = \frac{1}{2}(\dot{x}^2 + \dot{y}^2) - \frac{1}{2}g^2x^2y^2$, and its Euler–Lagrange equations (2-01) are (12) of 3-02.

Rescaling as $\hat{A} = gA$, (1) becomes $-\frac{1}{2g^2}\int \operatorname{tr} \hat{F}_{\mu\nu}\hat{F}^{\mu\nu}$ (the same form as the $\frac{1}{4g^2}\int \operatorname{Tr} F \wedge *F$ of Jaffe–Witten, up to conventions of normalisation and sign Known). g has no effect on the classical equations; in the quantum theory, as the combination $g^2\hbar$, it becomes the variable of the perturbative expansion (4-03).

02

The equations of motion — the field becomes its own source

Shifting A by δA, (6) of 3-02 gives $\delta F_{\mu\nu} = D_\mu \delta A_\nu - D_\nu \delta A_\mu$. By the antisymmetry of $F^{\mu\nu}$, and because integration by parts moves the commutator term onto the other

factor by $\operatorname{tr}([A,X]Y) = -\operatorname{tr}(X[A,Y])$ (δA is 0 far away),

$$\delta S = -2\int \operatorname{tr}\big(F^{\mu\nu} D_\mu \delta A_\nu\big)\, d^4x = 2\int \operatorname{tr}\big((D_\mu F^{\mu\nu})\, \delta A_\nu\big)\, d^4x$$

follows. The stationarity condition, that this is 0 for every δA, is the Yang–Mills equation Known .

$$D_\mu F^{\mu\nu} = 0, \qquad \partial_\mu F^{a\,\mu\nu} = -g\, f^{abc} A^b_\mu F^{c\,\mu\nu} \quad (2)$$

Unlike Maxwell, the product of A and F left on the right-hand side is the colour current carried by the gauge field itself (as announced in 3-02 § 06). (2) is a non-linear equation containing terms up to third order in A, and Jaffe–Witten write "The Yang–Mills equations are nonlinear—in contrast to the Maxwell equations." Known . In differential forms it is $d_A{*}F = 0$ (1-06).

The current is conserved only covariantly

When there is a matter field, a colour current appears on the right-hand side, with the same sign as in Maxwell: $D_\mu F^{\mu\nu} = -j^\nu$. On the left-hand side there is an identity that holds for any A (each term is a commutator of a matrix with itself, so it is 0).

$$D_\nu D_\mu F^{\mu\nu} = \tfrac{1}{2}[D_\nu, D_\mu] F^{\mu\nu} = -\tfrac{ig}{2}\,[F_{\nu\mu}, F^{\mu\nu}] = 0$$

So $D_\nu j^\nu = 0$ is required, and what is conserved in the ordinary sense is the sum of the matter current and the field current, $j^{a\nu} + g f^{abc} A^b_\mu F^{c\,\mu\nu}$ (the difference from the conservation of charge in 2-03).

03

Writing it with the colour electric and magnetic fields — the non-abelian Maxwell equations

Putting $F^{i0} = -E_i$ and $F^{ij} = \varepsilon_{ijk} B_k$ into $D_\mu F^{\mu\nu} = -j^\nu$ gives the laws of Gauss and Ampère–Maxwell (the Maxwell equations with sources of 2-03) with ∇ replaced by the covariant derivative $(\mathbf{D}_i X)^a = \partial_i X^a + g f^{abc} A^b_i X^c$ ($\rho = j^0$).

$$(\mathbf{D}\cdot\mathbf{E})^a = \rho^a, \qquad (\mathbf{D}\times\mathbf{B})^a - (D_t\mathbf{E})^a = \mathbf{j}^a \quad (3)$$

Gauss's law contains no second time derivative; it is a constraint at each time. With $A_0 = 0$, if it holds at the start, it keeps holding by the identity of § 02. The constraint $\varepsilon^{abc} A^b_i E^c_i = 0$ checked in 3-02 § 07 is this equation for a uniform field.

This constraint tells us how many components can be chosen freely. For each colour, A_μ has 4 components. Since no time derivative of A_0 appears in (1) (F_{i0} contains $\partial_0 A_i$ but not $\partial_0 A_0$), A_0 is not a dynamical variable and can be made $A_0 = 0$ by a gauge transformation. Of the 3 remaining components of A_i, the longitudinal one along the direction of propagation, seen in small waves, is removed by Gauss's law and by time-independent gauge transformations that preserve $A_0 = 0$. So $4 - 1 - 1 = 2$ remain for each colour, the same number as the two polarisations of light. For $SU(N)$ there are $N^2 - 1$ colours, so $2(N^2 - 1)$; for $SU(3)$ it is 16 Known (the quantum version is in 4-01).

04

The Bianchi identity — the other half comes from the construction

The source-free half of the Maxwell equations came from building the field out of a potential (the $dF = 0$ of 1-06). In the non-abelian case too, the following holds just because F is the curvature of A Known .

$$D\lambda F_{\mu\nu} + D_\mu F_{\nu\lambda} + D_\nu F_{\lambda\mu} = 0 \quad (4)$$

The reason is the Jacobi identity for operators, $[D_\lambda, [D_\mu, D_\nu]] + [D_\mu, [D_\nu, D_\lambda]] + [D_\nu, [D_\lambda, D_\mu]] = 0$. Putting in $[D_\mu, D_\nu] = -igF_{\mu\nu}$ ((6) of 3-02) and $[D_\lambda, F_{\mu\nu}] = D_\lambda F_{\mu\nu}$, which comes from the product rule, the left-hand side is $-ig$ times (4). In the geometric notation it is (12) of 1-07, and in Jaffe–Witten $0 = d_A F$. In electric and magnetic fields it is two equations, paired with (3).

$$\mathbf{D} \cdot \mathbf{B} = 0, \qquad \mathbf{D} \times \mathbf{E} + D_t \mathbf{B} = 0 \quad (5)$$

Unlike electromagnetism, the ordinary divergence $\nabla \cdot \mathbf{B}^a = -g f^{abc} \mathbf{A}^b \cdot \mathbf{B}^c$ is not necessarily 0; looking at one colour alone, magnetic field lines may end partway. More precisely, $\nabla \cdot \mathbf{B}^a$ is a quantity that changes under gauge transformations, and the gauge-independent statement is the $\mathbf{D} \cdot \mathbf{B} = 0$ of (5). The flux through the six faces of a small cube is measured in a different colour frame on each face; when they are parallel transported (1-07) to one common point before being added, the term of the order of the volume vanishes (the form of the first equation of (5) integrated over the cube). For a random field, the left-hand side of (4) was 0 within the error of the differences (4×10^{-7}), and the cyclic sum with ordinary derivatives was 2.1–2.6 Computed .

05

Energy and momentum — the trace vanishes in four dimensions

The Noether conserved quantity for translations (2-01), tidied into a gauge-invariant form, is the energy–momentum tensor Known .

$$T^{\mu\nu} = F^{a\,\mu\alpha} F^{a\,\nu}{}_{\alpha} - \tfrac{1}{4}\,\eta^{\mu\nu} F^{a}_{\alpha\beta} F^{a\,\alpha\beta} \quad (6)$$

$T^{00} = \tfrac{1}{2}\sum_a (|\mathbf{E}^a|^2 + |\mathbf{B}^a|^2)$ is the energy density ((10) of 3-02) and $T^{0i} = \sum_a (\mathbf{E}^a \times \mathbf{B}^a)_i$ the momentum density; since they are summed over colours they do not change under gauge transformations. The energy density is at least 0, and it is 0 only where $F = 0$ (pure gauge, 3-02 § 04). Taking the divergence and using (4), two derivative terms of F cancel, and

$$\partial_\mu T^{\mu\nu} = (D_\mu F^{\mu\alpha})^a\, F^{a\,\nu}{}_{\alpha} \quad (7)$$

remains. It is 0 on solutions, and energy and momentum are conserved. If there is matter the right-hand side is $-j^{a\alpha} F^{a\,\nu}{}_{\alpha}$, and the field loses as much as the colour Lorentz force does on the matter. From $\eta_{\mu\nu}\eta^{\mu\nu} = d$ (the dimension of spacetime), the trace is

$$T^{\mu}{}_{\mu} = \left(1 - \frac{d}{4}\right) F^{a}_{\alpha\beta} F^{a\,\alpha\beta} \quad (8)$$

and it vanishes identically only for $d = 4$. This is the stress side of saying that the classical action has no length scale (2-04, 1-12) Known . In the quantum theory, perturbative calculation gives the trace a F^2 term proportional to the β function (the rate at which the coupling constant moves with the scale; 4-06) (the trace anomaly), and this is taken to be the same phenomenon as the dimensional transmutation of 4-07 Physics . A figure that places the coefficient of the trace, $1 - d/4$, next to other quantities that switch at four is the dimension knob of 1-12 § 10.

For random SU(2) and SU(3) fields, the difference between $D_\nu D_\mu F^{\mu\nu}$ and the two sides of (7) was within the error of the differences (within 2.2×10^{-5}), T^{00}, T^{0i}, the trace and the form of (3) were within rounding error, and the coefficient in (8) was $d = 3, 5, 6$ also for $1 - d/4$ Computed .

06

There is no static lump — trace 0 and the balance of forces

If particles were to be represented by classical fields, the candidates would be solutions whose energy stays in one place (static lumps). For pure Yang–Mills fields in $3 + 1$ dimensions, the only time-independent solution that is smooth everywhere and

has finite energy is the vacuum Known . We follow the argument. If $T^{\mu\nu}$ is time-independent, (7) gives $\partial_k T^{kj} = 0$ ($T^{\mu\nu}$ is symmetric by (6), so the divergence is the same whichever index it is taken on). Together with $\partial_k x^i = \delta_{ik}$ this gives $\partial_k(x^i T^{kj}) = T^{ij}$, so integrating over a ball of radius R and using the divergence theorem moves the volume integral to a surface integral (usable because the field is smooth inside the ball).

$$\int_{|\mathbf{x}|<R} T^{ij}\, d^3x = \oint_{|\mathbf{x}|=R} x^i T^{kj} n_k \, dS$$

The right-hand side is of the order of the product of the area of the sphere $4\pi R^2$, the length of the arm R, and $|T|$ on the sphere. For example, if T decays faster than r^{-3}, it goes to 0 as $R \to \infty$ and $\int T^{ij} d^3x = 0$. In fact finite energy alone is enough (the detour below). Together with $-T^{00} + \sum_i T^{ii} = 0$ from (8),

$$E = \int T^{00}\, d^3x = \sum_i \int T^{ii}\, d^3x = 0 \quad (9)$$

follows. The energy density is a continuous function that is at least 0, so it is 0 everywhere, and $F = 0$ everywhere. In this form the outward pressure of the lump, $\sum_i T^{ii}$, equals the energy, and there is nothing to push back. Papers that put related conclusions in their titles are Deser (1976) “Absence of static solutions in source-free Yang–Mills theory” and Coleman (1977) “There are no classical glueballs” (bibliographic data only; the methods of proof have not been checked in the text). Adding a constant with the dimension of mass (a Higgs field and so on) makes the trace non-zero, and this argument no longer applies.

Detour: no assumption on the rate of decay is needed

A subsection you may skip.

Writing (6) in electric and magnetic fields, for each colour $T^{ij} = \frac{1}{2}\delta_{ij}(|\mathbf{E}|^2 + |\mathbf{B}|^2) - E_i E_j - B_i B_j$. Sandwiching between unit vectors $\mathbf{n}$ gives $\frac{1}{2}(|\mathbf{E}|^2 + |\mathbf{B}|^2) - (\mathbf{n}\cdot\mathbf{E})^2 - (\mathbf{n}\cdot\mathbf{B})^2$, and even summed over colours it lies between $-T^{00}$ and T^{00}. Since it is a symmetric matrix, $|T^{kj} n_k| \le T^{00}$, and the size of the surface integral is at most $R\,f(R)$. Here $f(R) = \oint_{|\mathbf{x}|=R} T^{00} dS$ satisfies $\int_0^\infty f(R)\, dR = E$.

If $R\,f(R) \ge c > 0$ always held for large R, $\int^\infty c/R\, dR$ would diverge, contradicting the finiteness of E. So a sequence of radii $R_n \to \infty$ with $R_n f(R_n) \to 0$ can be chosen, and along that sequence the surface integral tends to 0. The left-hand side also converges from $|T^{ij}| \le T^{00}$ to $\int T^{ij} d^3x$, so (9) holds with only "smooth everywhere and of finite energy" Known .

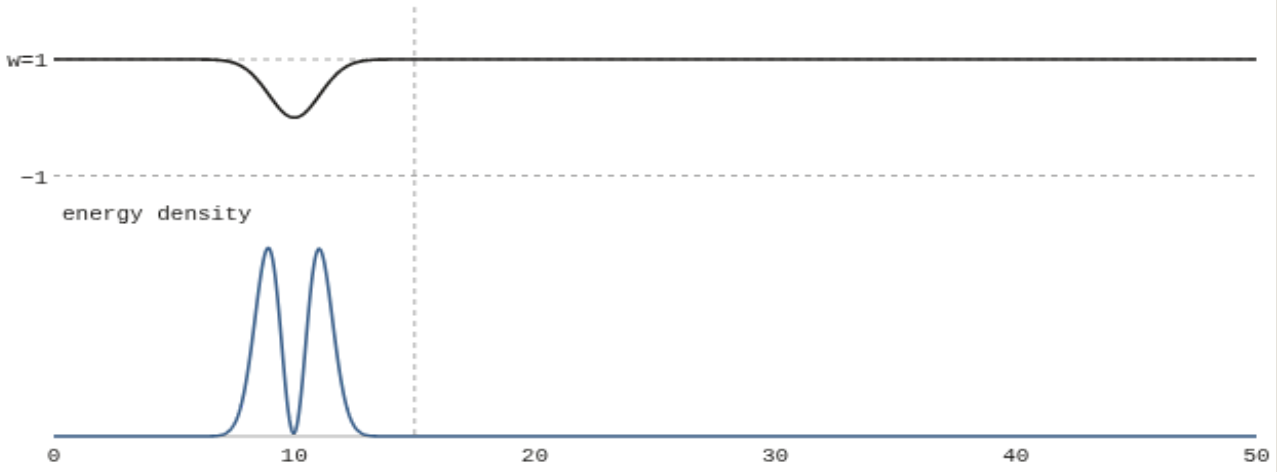
Seen in a spherically symmetric field

In SU(2), taking the form $A_0 = 0$, $A_i^a = \varepsilon_{aij}x_j\,(1 - w(r,t))/(gr^2)$, which is unchanged when space and colour are rotated together, the problem becomes one in a single variable $r = |\mathbf{x}|$.

$$E = \frac{4\pi}{g^2}\int_0^\infty \Big[w_t^2 + w_r^2 + \frac{(1-w^2)^2}{2r^2}\Big]\,dr, \qquad w_{tt} = w_{rr} + \frac{w(1-w^2)}{r^2} \quad (10)$$

$w = \pm 1$ is a pure gauge with energy 0. On a three-dimensional field with differences, we checked that the energy density (to accuracy 10^{-8}), Gauss's law and the rest of (3) reduce to (10) Computed .

Figure 1 The knob a changes the depth of the dip ($w(r,0) = 1 - a\,e^{-((r-10)/1.5)^2}$), and "Play" advances time. Top: w; bottom: the energy density per dr; the dotted line is $r = 15$



a = 1, t = 0.00, E(t)/E(0) = 1.0000, fraction in r<15 = 1.000

Computed The numbers in this figure were computed on this machine. It solves the same differences as Table 1 ($h = 0.05$, leapfrog with $\Delta t = h/2$) in the same order, and agrees with python3 to 12 digits. If JavaScript is disabled, Table 1 is the content of the figure.

Table 1 Time evolution of $a = 1.0$. The position of a peak is the maximum of the energy density per dr. The drift in the total energy $0.84922 \times 4\pi/g^2$ is at most 2×10^{-4} relative, for $t \le 40$

t	Fraction in $r < 15$	Outgoing peak	Incoming → reflected peak
0	1.0000	–	–
10	0.4996	18.95	1.10
20	0.4983	28.95	10.90
30	0.0000	38.90	20.90

The dip splits in two and travels at the speed of light, 1; the incoming peak reflects at the origin at $t \approx 10$, and by $t = 30$ nothing is left in $r < 15$. The same held for $a = 0.3$, 1.8 Computed . What is shown covers only these initial values and a finite range of time.

Detour: there is a static solution, but its energy is infinite

A subsection you may skip.

The $w = 0$ of (10) is a time-independent solution, and its magnetic field has the form of a point magnetic charge, $|\mathbf{B}| = 1/(gr^2)$ Computed . It is singular at the origin and the energy $\int 4\pi r^2\, dr/(2g^2r^4)$ diverges at the origin, so it lies outside the assumptions of (9) (smooth everywhere, finite energy). It is the solution said to be called by the names of Wu and Yang.

07

The action in imaginary time and self-duality — a lower bound on the action

We move to imaginary time $t = -i\tau$, $x_4 = \tau$ (2-04 § 09). From $A_0\, dt = A_4\, d\tau$ we get $A_0 = iA_4$, and also $\partial_t = i\partial_4$, so $F_{i0} = iF_{i4}$. Rewriting as $E_i^a = F_{i4}^a$, the sign of $|\mathbf{E}^a|^2$ in (1) flips, and iS becomes $-S_E$. The Euclidean action S_E is a sum of positive quantities, and in the quantum theory it enters the weight e^{-S_E} (4-02).

$$S_E[A] = \frac{1}{4}\int F^a_{\mu\nu}F^a_{\mu\nu}\, d^4x = \frac{1}{2}\int \sum_a \left(|\mathbf{E}^a|^2 + |\mathbf{B}^a|^2\right) d^4x \;\geq 0 \qquad (11)$$

The Hodge star $(*F)_{\mu\nu} = \frac{1}{2}\varepsilon_{\mu\nu\rho\sigma}F_{\rho\sigma}$ ($\varepsilon_{1234} = 1$, 1-12) exchanges $\mathbf{E}$ and $\mathbf{B}$ in this notation, so self-dual $F = *F$ means $\mathbf{E}^a = \mathbf{B}^a$, and anti-self-dual means $\mathbf{E}^a = -\mathbf{B}^a$. At each point $\frac{1}{2}(|\mathbf{E}|^2 + |\mathbf{B}|^2) = \frac{1}{2}|\mathbf{E} \mp \mathbf{B}|^2 \pm \mathbf{E}\cdot\mathbf{B}$, so

$$S_E = \frac{1}{2}\int \sum_a |\mathbf{E}^a \mp \mathbf{B}^a|^2\, d^4x \pm \frac{8\pi^2}{g^2}Q \;\geq\; \frac{8\pi^2}{g^2}|Q|, \qquad Q = \frac{g^2}{8\pi^2}\int \sum_a \mathbf{E}^a\cdot\mathbf{B}^a\, d^4x \qquad (12)$$

with equality only when self-dual ($Q > 0$) or anti-self-dual ($Q < 0$) Known . For a solution of the Yang–Mills equations with finite action, it is said that it extends to a field on S^4 and Q is an integer (Uhlenbeck 1982, bibliographic data only). General fields that are not solutions are treated in 3-05 § 03. In the geometric notation of 3-05 (anti-Hermitian A and c_2) it is $Q = -c_2$ (the detour of 3-05).

A self-dual field is automatically a solution, because $D_\mu F_{\mu\nu} = \pm\frac{1}{2}\varepsilon_{\mu\nu\rho\sigma}D_\mu F_{\rho\sigma}$ is 0 by the cyclic sum (4). Instead of the second-order (2), it is enough to solve the first-order $F = \pm *F$, and its solutions have the least action within the same Q. Jaffe–Witten § 2 write that classical non-abelian gauge theory has played a large role in the study of three- and four-dimensional manifolds Known . The representative of this in four dimensions is the first-order (anti-)self-dual equation, and Donaldson (1983) used the space of its solutions for four-dimensional topology (3-06) Known . Real self-dual fields exist only in imaginary time (1-06).

08

A stretched instanton — Q is unchanged and only the action grows

As an example we borrow the BPST solution (Belavin–Polyakov–Schwartz–Tyupkin 1975), constructed in 3-05 Known . In 't Hooft's symbols ($\eta_{aij} = \varepsilon_{aij}$, $\eta_{ai4} = -\eta_{a4i} = \delta_{ai}$), the solution of size ρ has the following form.

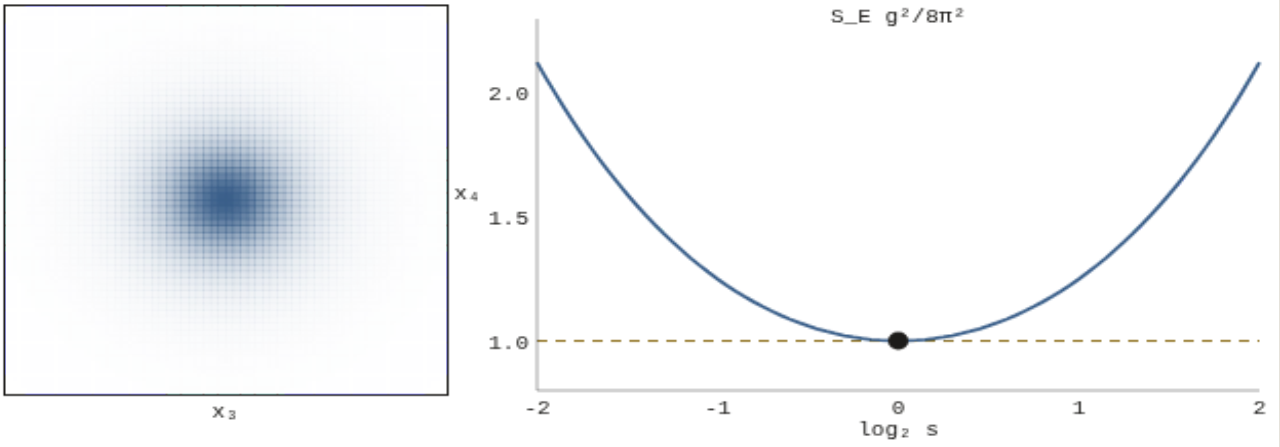
$$A^a_\mu = \frac{2}{g}\,\frac{\eta_{a\mu\nu} x_\nu}{|x|^2+\rho^2}, \qquad F^a_{\mu\nu} = -\frac{4}{g}\,\frac{\eta_{a\mu\nu}\,\rho^2}{(|x|^2+\rho^2)^2} \quad (13)$$

Since η_a is self-dual, $\mathbf{E}^a = \mathbf{B}^a$, $S_E = 8\pi^2/g^2$, $Q = 1$ (Table 1 of 1-12). With differences we checked the right-hand formula of (13), $F = *F$ and $D_\mu F_{\mu\nu} = 0$ to within 6×10^{-10} Computed . We deform this in the direction of imaginary time. Pulling back by $M = \mathrm{diag}(1,1,1,s)$ as $A^{(s)}_\mu(x) = M_{\nu\mu} A_\nu(Mx)$, for $s>1$ the shape of the field shrinks by a factor $1/s$ in the x_4 direction, and for $s<1$ it stretches by a factor $1/s$. $\mathbf{E}^{(s)} = s\,\mathbf{E}(Mx)$, $\mathbf{B}^{(s)} = \mathbf{B}(Mx)$, and only the components $Fi4$ of $\mathbf{E}$ that have x_4 as an index are multiplied by s (because s multiplies the derivative in x_4 and A_4). For $s \neq 1$ it is not a solution. Substituting $y = Mx$ ($d^4x = d^4y/s$) gives the following.

$$S_E^{(s)} = \frac{8\pi^2}{g^2}\cdot\frac{1}{2}\Big(s+\frac{1}{s}\Big), \qquad Q^{(s)} = 1 \quad (14)$$

The action is least at $s=1$ and grows whether stretched or shrunk (the inequality of arithmetic and geometric means). Q does not change because the integrand $\mathrm{tr}\,F\wedge F$ is a four-form that does not use the metric, and its integral does not change under orientation-preserving changes of coordinates (1-06). The action uses the Hodge star, that is, the metric, so it changes.

Figure 2 Moving $\log_2 s$ with the knob stretches and shrinks the cross-section of the action density on the left ($x_1 = x_2 = 0$, vertical axis imaginary time x_4) (vertically shrunk for $s>1$), and the point on the right moves along the curve $\frac{1}{2}(s+1/s)$. The dashed line is the lower bound $Q=1$



s = 1.000 |E|/|B| = 1.000 S_E·g²/8π² = 1.0000 Q = 1

Computed The numbers in this figure were computed on this machine. The figure computes (14) and the closed formula for the action density on the spot (agreeing with Table 2).

Table 2 The instanton stretched in the x_4 direction, integrated numerically in four-dimensional polar coordinates in x without a change of variables ($g = \rho = 1$; 96 radial points, $48 \times 64 \times 64$ angular points) Computed . S_E agrees with (14) to 6 digits, and $\geq |Q| = 1$ for every s

s	0.25	0.5	0.8	1	1.25	2	4
$SEg^2/8\pi^2$	2.125000	1.250000	1.025000	1.000000	1.025000	1.250000	2.125000
Q	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

09

What the classical theory says, and what it does not

The classical Yang–Mills field has no static lump of finite energy (§ 06), no lower bound on the frequency of waves (3-02 § 07), and in imaginary time the minimum of the action in each Q is attained by self-dual fields. None of these is a mass gap. Jaffe–Witten § 2 call the mass gap and confinement "quantum properties that do not have classical analogs" Known . Since the classical theory has no scale, the scale of the gap is thought to enter in the process of quantisation (4-07) Physics . This chapter shows nothing about the quantum theory or about the mass gap.

A bridge to the chapters ahead. In 3-04 we study the shape of the space in which fields related by gauge transformations are regarded as the same (the gauge orbit space), and the reason why the gauge cannot be fixed in one way globally. That the Q of § 07 is an integer and the construction of the BPST solution are in 3-05, and the story of how the space of solutions of self-dual fields distinguishes four-dimensional topology is the detour of 3-06. The trace and scale of § 05 continue on the quantum side in 4-06 (asymptotic freedom) and 4-07 (dimensional transmutation).

What this chapter can say

Matter	What is said	What is not said
(1)–(8)	Derivations and numerical checks Known Computed	Matter fields (3-07), the quantum trace Physics
(9)	The argument about time-independent solutions that are smooth everywhere and of finite energy Known	Removing the smoothness assumption (the $w = 0$ in the detour of § 06 is singular)
(10), Figure 1, Table 1	The scattering seen over finite time for three initial values Computed	A proof for general initial values

Matter	What is said	What is not said
(12)–(14), Table 2	Derivations and 6-digit numerical checks Known Computed	That Q is an integer (3-05)
The mass gap	—	That the quantum Yang–Mills theory has a gap (open; this site shows nothing) Physics

Sources and reproduction

Item	Kind	Source / tools
Quotations and mentions in § 01, § 02, § 04, § 07, § 09	Primary source (text checked)	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory” (2000), the official problem description of the Clay Mathematics Institute, § 1 and § 2. Collected in *The Millennium Prize Problems* (2006)
Non-existence of static solutions, the BPST solution, extension of solutions of finite action	Bibliographic data only (Crossref)	S. Deser, “Absence of static solutions in source-free Yang–Mills theory”, Phys. Lett. B 64 (1976) 463–464 / S. Coleman, “There are no classical glueballs”, Commun. Math. Phys. 55 (1977) 113–116 / A. A. Belavin, A. M. Polyakov, A. S. Schwartz, Yu. S. Tyupkin, “Pseudoparticle solutions of the Yang-Mills equations”, Phys. Lett. B 59 (1975) 85–87 / K. Uhlenbeck, “Removable singularities in Yang-Mills fields”, Commun. Math. Phys. 83 (1982) 11–29
Donaldson in § 07	Bibliographic data only (quoted in the text of 3-06)	S. K. Donaldson, “An application of gauge theory to four-dimensional topology”, J. Differential Geom. 18 (1983) 279–315
The name of the $w = 0$ solution	Not checked	The primary source has not been checked. Candidate: T. T. Wu, C. N. Yang, “Some solutions of the classical isotopic gauge field equations”, in H. Mark, S. Fernbach (eds.), *Properties of Matter Under Unusual Conditions* (1969), p. 349 ff. (bibliographic data only; whether it contains this solution has not been checked)
Uniqueness of the action (with the scale condition), the variation, (4)–(9), (12)	Known	Standard material. Derivations in the text. Conventions as in 3-02 § 02
Numbers in § 03–§ 05, § 06 and § 08, Table 1, Table 2, Figure 1, Figure 2	Computed on this machine	Python 3.12, NumPy 1.26 (`identity_checks.py`, `radial_reduction.py`, `radial_sim.py`, `instanton_stretch.py`). Figure 1 solves the same differences in plain JavaScript (checked against Node.js)

Next chapter to read: 3-04 Gauge theory as geometry. For instantons, 3-05; for the lattice action, 5-01.

2026-10-01 · chapter Yang–Mills and the mass gap Part 3 Yang–Mills theory (classical) the space of connections and the orbit space

Gauge theory as geometry — divide a straight space by orbits, and representatives can no longer be chosen continuously

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

What this chapter shows — The set of all connections $\mathcal{A}$ is straight, but the orbit space, in which configurations related by gauge transformations are identified, is curved and has corners. Gauge fixing is choosing a slice of the orbit space, and a slice meets the same orbit many times (Gribov copies). By Singer's theorem, for semisimple groups such as $SU(N)$ over S^4 (and S^3) there is no continuous slice.

Prerequisite chapters — 1-07 (connections, gauge transformations, holonomy) and 3-02 (the non-abelian field strength). Haar measure (§ 04), homotopy groups and principal bundles (§ 07) are defined where they appear.

Said first — This is a chapter of classical geometry, treating neither quantum theory nor the mass gap. Singer's theorem is presented as a statement and the outline of its proof. The figure and tables are computations in lattice models with finitely many edges; they do not count copies in the four-dimensional continuum theory.

Order of this chapter

1. The space of configurations — each whole field as one point
2. The space of connections — straight, with no holes
3. The gauge group and orbits — tangents, and the elements that fix
4. The orbit space — curved, with corners (Table 1). Detour: the orbit space and the gap
5. Gauge fixing is a slice — copies and the horizon (Figure 1, Table 2)
6. Going once round leads to another copy (Table 3)
7. Singer's theorem — the outline in four steps (Table 4)

8. How it has been avoided

01

The space of configurations — the whole field seen as one space

In mechanics, the set of all positions of a particle was seen as one space (the configuration space) (2-01). For fields too, we regard one connection A as one point and write the set of all of them as $\mathcal{A}$ (infinite-dimensional). Configurations related by gauge transformations represent the same physics (3-01), so the true configuration space is the collection of bundles of related configurations. We call one such bundle an **orbit**, and the set of all of them the **orbit space** $\mathcal{A}/\mathcal{G}$. The gauge transformation is (4) of 1-07 and its lattice version ((7) of 1-07).

$$g \cdot A = gAg^{-1} - (dg)\,g^{-1}, \qquad (g \cdot U)\ell = g_y\, U_\ell\, g_x^{-1} \quad (\ell : x \to y) \tag{1}$$

The U_ℓ in (1) is the parallel transport carrying from x to y along the edge ℓ ((7) of 1-07). The $U_{x,\mu}$ of 5-01 is read in the direction carrying from $x + a\hat{\mu}$ to x, so in the notation of this chapter it corresponds to U_ℓ ($\ell : x + a\hat{\mu} \to x$) ($U_\ell = U_{x,\mu}$; on a reversed edge, the inverse).

Example (lattice). On a lattice with group elements on the edges (5-01), a configuration is a tuple G^E of E elements, one per edge, and a gauge transformation is a tuple G^V of V elements, one per vertex; both are finite-dimensional. There are two questions. What shape is the orbit space? Can one representative be chosen from each orbit continuously (§ 07)?

02

The space of connections — straight, with no holes

Conventions. As in 1-07, $G = SU(N)$, $\mathfrak{g}$ is the anti-Hermitian traceless matrices, and a connection is a $\mathfrak{g}$-valued 1-form $A = \sum_\mu A_\mu\, dx^\mu$ (the $\mathcal{A} = -igA$ of the table in 3-02 is written A in this chapter, and $\mathcal{A}$ is used for the set of all connections). The base space M is $\mathbb{R}^4$, S^4, S^3 or the circle. The difference of two connections is a $\mathfrak{g}$-valued 1-form, and adding a 1-form to a connection gives a connection again (1-07 § 03). So $\mathcal{A}$ is a vector space with no fixed origin (an affine space).

$$A_t = (1-t)A_0 + tA_1 \in \mathcal{A}, \qquad g \cdot A_t = (1-t)\, g \cdot A_0 + t\, g \cdot A_1 \tag{2}$$

The right-hand formula follows because the $-(dg)g^{-1}$ in (1) appears only once, with weight $(1-t)+t=1$. Moving t from 1 to 0 shrinks the whole of $\mathcal{A}$ continuously to one point A_0. We call this **contractible**, and it is the starting point of § 07. On the tangent directions a, b we put an inner product.

$$\langle a, b\rangle = -\int_M \sum_\mu \operatorname{tr}\left(a_\mu b_\mu\right) d^d x \quad (3)$$

For anti-Hermitian matrices $-\operatorname{tr}(a_\mu a_\mu) \geq 0$, and the gauge transformation $a \mapsto gag^{-1}$ preserves this inner product. **Example.** The lattice $U(1)^E$ is a torus and has loops that cannot be shrunk, so it is not contractible (that the configuration space of a lattice is not contractible matters in § 08).

03

The gauge group and orbits — tangents to orbits, and the elements that fix

The set of all gauge transformations $\mathcal{G}$ (smooth $g: M \to G$, with pointwise product) is an infinite-dimensional group. The tangent to an orbit comes from putting $g = e^{t\xi}$ ($\xi : M \to \mathfrak{g}$) into (1) and differentiating.

$$\left.\frac{d}{dt}\right|_{t=0} e^{t\xi}\cdot A = -D_A\xi, \qquad D_A\xi = d\xi + [A,\xi] \quad (4)$$

D_A is the covariant derivative in the adjoint representation (3-02 § 03). Integrating (3) by parts (the inner product of two ξ is $-\int \operatorname{tr}(\xi\eta)\, d^d x$, and boundary terms are dropped) determines the adjoint D_A^*.

$$\langle D_A\xi, a\rangle = \langle \xi, D_A^* a\rangle, \qquad D_A^* a = -\sum_\mu \left(\partial_\mu a_\mu + [A_\mu, a_\mu]\right) \quad (5)$$

That a direction a is orthogonal to the orbit is the same as $D_A^* a = 0$, and at $A = 0$ it is the Landau condition $\partial_\mu a_\mu = 0$ Known . **The elements that fix.** Multiplying $g\cdot A = A$ by g on the right gives $gA - dg = Ag$, that is,

$$g\cdot A = A \iff dg + [A, g] = 0 \quad (6)$$

which is the equation of parallel transport (1-07 § 04). g is determined by its value at one point m, and $g(m)$ must commute with the holonomy of A (Singer 1978 § 2) Known . So the based gauge group $\mathcal{G}_m = \{g : g(m) = 1\}$ moves every connection (acts freely), and the only elements that fix an **irreducible** connection, one whose holonomies have no common invariant subspace, are the centre $\mathbb{Z}_N$ (1-04). A **reducible** connection such

as $A = 0$ is fixed by a larger group (all constant G); its orbit becomes thinner, and it is a corner of the orbit space.

04

The orbit space — curved, with corners

On a lattice of L edges forming a ring, the only gauge-invariant quantity is the conjugacy class of the holonomy round the ring, $U = U_{L-1} \cdots U_0$ (1-07 § 06). For $U(1)$ it is the angle Φ ($U = e^{i\Phi}$), and the orbit space is a circle. For $SU(2)$ it is fixed by the eigenvalues $e^{\pm i\theta}$ ($0 \le \theta \le \pi$), and the orbit space is the segment $[0, \pi]$; the ends $U = \pm 1$ are corners where the elements that fix jump up to the whole of $SU(2)$. Seen as $SU(2) \cong S^3$ (unit quaternions, 1-12), the conjugacy class with angle θ is a sphere of radius $\sin\theta$, which collapses to a point at the ends. The measure of weight 1 on a group that does not change under multiplication is called **Haar measure**. For $SU(2)$ it is the uniform distribution on S^3 (1-08 § 03). Pushed down to the orbit space, it gets the weight of the area of the orbit (numerator) divided by the volume of S^3 (denominator) (the Weyl integration formula) Known .

$$\rho(\theta)\, d\theta = \frac{4\pi \sin^2\theta}{2\pi^2}\, d\theta = \frac{2}{\pi} \sin^2\theta\, d\theta \quad (7)$$

Computed **Table 1** The distribution of the angle of the holonomy θ for a ring whose 4 edges are each chosen at random with the Haar measure of $SU(2)$ (2 million samples; sampling error at most 0.0004).

Interval of θ	Samples	Integral of (7)
0°–30°	0.0289	0.0288
30°–60°	0.1666	0.1667
60°–90°	0.3047	0.3045
90°–120°	0.3044	0.3045
120°–150°	0.1667	0.1667
150°–180°	0.0288	0.0288

It agrees with (7), and the same holds for a single edge (a product of Haar-distributed elements is again Haar-distributed). Orbits near the corners are small, and the probability of landing there is small too. This weight plays the same role as the Faddeev–Popov determinant multiplied in when the gauge is fixed (the determinant of M in (9) of § 05). More precisely, when conjugation in a finite-dimensional group is fixed to diagonal form, the Faddeev–Popov determinant becomes a constant times $\sin^2\theta$ (the Weyl factor). The determinant in infinite-dimensional gauge fixing is a formal analogue of it (4-08 § 02).

Detour: the orbit space and the gap

A subsection you may skip.

In quantum theory the kinetic energy corresponds to the Laplacian on the orbit space (4-01). Karabali–Nair (1996) summarise Feynman's (1981) point as "for a finite-dimensional compact manifold the Laplacian has a discrete spectrum, the gap between the first and second eigenvalues being the analogue of the mass gap", and present a (2+1)-dimensional argument that a gap appears if gauge-invariant distances do not grow without bound Physics . Strictly this is not a theorem, and the same paper cautions that it would "have to be understood with regularizations and all that".

05

Gauge fixing is a slice — Gribov copies and the horizon

Fixing the gauge means choosing a map $s : \mathcal{A}/\mathcal{G} \to \mathcal{A}$ that picks one representative from each orbit (a **slice**). A rule often used is to choose the point on the orbit closest to the origin. From (4)(5),

$$\frac{d}{dt}\Big|_{t=0} \|e^{t\xi} \cdot A\|^2 = -2\langle A, D_A\xi\rangle = -2\langle D_A^* A, \xi\rangle \quad (8)$$

and this is 0 for every ξ when $\partial_\mu A_\mu = 0$, the Landau gauge. Singer calls the surface $\mathcal{S}_A = \{A + \tau : D_A^*\tau = 0\}$ through A orthogonal to the orbits the generalised Coulomb gauge, and states that near A the surface meets each orbit exactly once (a local slice) Known . Far away the surface becomes tangent to orbits. The tangent $D_{A+\tau}\xi$ to the orbit at $A + \tau$ lies in the surface when the following operator has a non-zero solution.

$$M(\tau)\,\xi = D_A^* D_{A+\tau}\,\xi \quad (9)$$

There the Faddeev–Popov determinant (the determinant of $M(\tau)$) is 0. For $A = 0$ it is $M(\tau) = -\partial_\mu D_{\tau,\mu}$. The range where M is positive is called the **Gribov region**, its boundary the **horizon**, and each point where an orbit meets the surface a **Gribov copy**. Singer's Theorem 8 shows that in every direction of the surface, far enough out, there is a point where it is tangent to an orbit. Gribov (1978) found, for $SU(2)$ in $\mathbb{R}^4$, that the Coulomb gauge meets the orbit of the origin once more far away (as presented by Singer) Known . How to count them in quantum theory is in 4-08 § 05 and § 06.

Example (Figure 1). A ring of L edges of $U(1)$. With $g_0 = 1$ at the base point and the rest $g_x = e^{i\alpha x}$, the angle of an edge transformed by (1) is $\psi_\ell = \theta_\ell + \alpha_{\ell+1} - \alpha_\ell$ (indices go round with L). We measure the distance from the origin with the same function as Mehta et al. (2014).

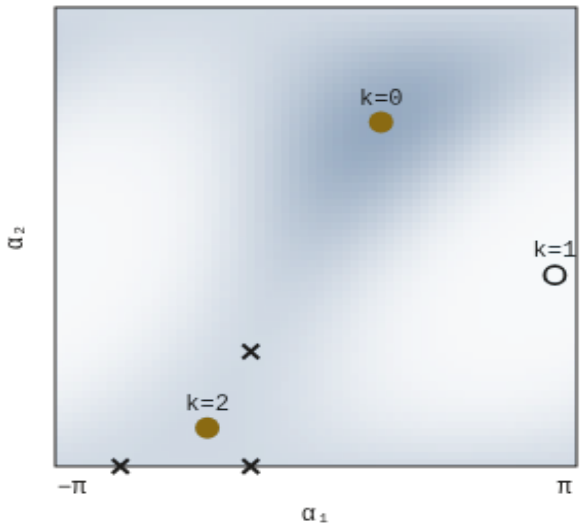
$$E(\alpha)=\sum_{\ell=0}^{L-1}\bigl(1-\cos\psi_\ell\bigr),\qquad \frac{\partial E}{\partial\alpha_j}=\sin\psi_{j-1}-\sin\psi_j \tag{10}$$

For small ψ, $E\approx\frac12\sum_\ell\psi_\ell^2$, the lattice version of (8). At stationary points all the $\sin\psi_\ell$ are equal; there are the uniform points $\psi_\ell=\varphi$ and points where only some edges have $\pi-\varphi$. The uniform points are the following L.

$$\varphi_k=\frac{\Phi+2\pi k}{L}\quad(k=0,\dots,L-1),\qquad E=L\,(1-\cos\varphi_k) \tag{11}$$

The matrix of second derivatives there (the lattice Faddeev–Popov matrix) is $\cos\varphi_k$ times the Laplacian of the ring, so for $\cos\varphi_k>0$ it is a minimum (a copy inside the Gribov region), and for negative values a maximum. Non-uniform points are not minima ($L=3$–9) Computed .

Figure 1 Changing Φ with the knob moves the shading of E (darker is smaller) on the orbit (a torus) of $L=3$ and its critical points (● minimum, × saddle, ○ maximum; k is the label in (11)). "Play" takes Φ once round, and the representative $k=0$ arrives at the original position of $k=1$. For $L=3$ it crosses the horizon on the way, so where it arrives is a maximum (§ 06).



Φ = 135°, minima: 2, E = 0.879, 2.224

Φ	Minimum	Saddle	Maximum	E at the minima	φ of the minima
0°	1	3	2	0.0000	0°
99°	2	3	1	0.4840, 2.8430	33°, −87°
135°	2	3	1	0.8787, 2.2235	45°, −75°
180°	2	3	1	1.5000, 1.5000	60°, −60°

Computed **Table 2** Critical points of the orbit of the representative $\theta=(0,0,\Phi)$, listed by (10) and classified by second derivatives. The numbers in this figure were computed on this machine (the figure computes the same formulas on the spot). If JS is disabled, this table is the content of the figure.

Only when Φ is between 90° and 270° are $\cos\varphi_0 > 0$ and $\cos\varphi_2 > 0$ both possible, and there are two copies inside the Gribov region. When Φ crosses 90° going up, the point $k = 2$ and three saddle points merge into one, and the second derivative becomes 0. Beyond this, $k = 2$ becomes a minimum. This is the horizon. At $\Phi = 180^\circ$ the E of the two minima are equal, and the boundary of the region that collects only the closest points (the fundamental modular region) is here. That the signed sum of critical points is 0 (Neuberger's 0/0) is in 7-11 § 05.

06

Going once round leads to another copy — keep choosing representatives continuously and they disagree

Taking Φ once round in Figure 1, the uniform representative of (11) moves continuously from $\varphi = \Phi/L$ to $(\Phi + 2\pi)/L$ and arrives at the representative $k = 1$. The orbit has returned to where it was, but the chosen representative is another copy. For $L = 3$ in Figure 1, it crosses the horizon at $\Phi = 270^\circ$ on the way, and where it arrives is a maximum (a copy outside the Gribov region). Arriving at another copy while staying inside the region happens for $L \geq 5$ (Table 3). The two copies are related by the gauge transformation that turns the point x by the angle $\alpha_x = 2\pi x/L$ (the transformation of winding number 1 on the continuous circle, transferred to the lattice) (on the last edge too, $\alpha_0 - \alpha_{L-1} \equiv 2\pi/L$).

$$\psi_\ell \;\mapsto\; \psi_\ell + \alpha_{\ell+1} - \alpha_\ell \equiv \psi_\ell + \frac{2\pi}{L} \pmod{2\pi}, \qquad \varphi_k \mapsto \varphi_{k+1} \quad (12)$$

Keep choosing representatives continuously, and after one round they disagree: this is the smallest form of the mechanism that prevents a continuous global gauge fixing. Whether the path hits the horizon ($\cos\varphi = 0$, that is, $\Phi = \pi L/2$) on the way depends on L.

Computed **Table 3** The result of following the minimum $\Phi = 0$ once round along Φ, and the number of copies inside the Gribov region (minimum–maximum over 4000 values of Φ).

L	Copies inside the region	Φ at which the horizon is hit	Representative after one round
3	1–2	270°	— (stops being a minimum on the way)
4	2	360° (end of the round)	—
5	2–3	Not hit	$\varphi = 2\pi/5$ ($k = 1$)
8	4	Not hit	$\varphi = 2\pi/8$ ($k = 1$)

For $L \geq 5$ the representative stays inside the Gribov region during the round and arrives at the copy $k = 1$; for $L \leq 4$ it hits the horizon on the way or at the end. In

neither case does it return to the original representative Computed .

Example (the continuous circle). For a $U(1)$ connection $A = i\,a\,dx$ on a circle of length ℓ, the Landau condition makes a constant, and the gauge transformation $e^{2\pi i n x/\ell}$ of winding number n takes a to $a - 2\pi n/\ell$, so there are infinitely many copies. Only $a\ell \bmod 2\pi$ is gauge-invariant; choosing a representative continuously means choosing a real number continuously from an angle (taking a continuous logarithm), which is impossible because after one round it is off by 2π Known .

07

Singer's theorem — a contractible space cannot be a product with a group that has holes

Singer (1978) takes as base space S^4 ($\mathbb{R}^4$ with one point at infinity added; this corresponds to the condition that gauge transformations approach a constant value far away) or S^3 (space at one time), and for compact connected semisimple groups (groups whose Lie algebra is a direct sum of simple ones; in the paper, for simplicity, $SU(N)$, $N > 1$) states in the abstract "It is shown that no continuous choice of exactly one connection on each orbit can be made.", concluding "No gauge fixing is possible." Known . Precisely, it says there is no **continuous global** gauge fixing; local slices and discontinuous choices are not forbidden.

Homotopy groups. $\pi_k(X)$ is the set of classes of based continuous maps from the k-dimensional sphere into X, grouped by continuous deformation into one another. π_0 is the connected components, π_1 the classes of loops, and $\pi_3(SU(2))$ is the degree of 1-12, with $\pi_k(X \times Y) = \pi_k(X) \times \pi_k(Y)$. The proof has four steps (Singer 1978 § 2).

1. $\mathcal{A}$ is contractible (§ 02). So is $\mathcal{A}^*$, with the set of reducible connections C removed: all its homotopy groups vanish, $\pi_j(\mathcal{A}^*) = 0$ (Singer's Theorem 2). The removed set C leaves, at every point, infinitely many directions orthogonal to it, so the image of a sphere can be pushed off C.
2. $\hat{\mathcal{G}} = \mathcal{G}/\mathbb{Z}_N$ (the centre $\mathbb{Z}_N$ fixes every connection, so it is divided out) acts freely on $\mathcal{A}^*$, and $\mathcal{A}^* \to \mathcal{N}^* = \mathcal{A}^*/\hat{\mathcal{G}}$ is a principal bundle with fibre $\hat{\mathcal{G}}$ (Singer's Theorem 1). A **principal bundle** is a bundle in which over each point of the base (here $\mathcal{N}^*$) sits a copy of the group $\hat{\mathcal{G}}$ (the **fibre**; here one orbit), and $\hat{\mathcal{G}}$ acts freely and transitively within each copy (every element other than 1 moves every point, and any point can be moved to any point). If there is a section, the whole is the product of the base and the group.
3. If there were a continuous gauge, it would be a section of the principal bundle, the bundle would be the product $\mathcal{A}^* \cong \mathcal{N}^* \times \hat{\mathcal{G}}$, and from $\pi_j(\mathcal{A}^*) = 0$ and the product

formula, $\pi_j(\hat{\mathcal{G}}) = 0$ would follow for every j.

4. But some π_j of $\hat{\mathcal{G}}$ is not 0 (Theorem 3). Contradiction (Corollary 4).

$$\pi_j(\mathcal{G}_m) \cong \pi_{j+r}\big(SU(N)\big) \qquad (M = S^r)\ \text{(13)}$$

The tool of the fourth step is (13) (Theorem 5). Singer notes that $\mathcal{G}_m$ acts freely on the whole of $\mathcal{A}$, and by (13) some homotopy group of it is not 0, so $\mathcal{A} \to \mathcal{A}/\mathcal{G}_m$ has no section either.

Known **Table 4** $\pi_j(\mathcal{G}_m)$ $(M = S^r)$ read from (13). The non-abelian columns follow Singer's text (which cites Toda's tables).

M	$U(1)$	$SU(2)$	$SU(N)$, $N \geq 3$
S^3	All 0 (note)	$\pi_0(\mathcal{G}_m) = \pi_3(SU(2)) = \mathbb{Z}$	$\pi_0(\mathcal{G}_m) = \pi_3(SU(N)) = \mathbb{Z}$
S^4	All 0 (note)	$\pi_0(\mathcal{G}_m) = \pi_4(SU(2)) = \mathbb{Z}_2$, $\pi_2(\mathcal{G}_m) = \pi_6(SU(2)) = \mathbb{Z}_{12}$	$\pi_0(\mathcal{G}_m) = \pi_4(SU(N)) = 0$ (the fourth step uses $\pi_1(\hat{\mathcal{G}}) \neq 0$)

Note: the $U(1)$ column is read from the homotopy groups $\pi_j(\mathcal{G}_m) = \pi_{j+r}(U(1))$ of the space of based maps (the π_k of $U(1)$ is 0 for $k \geq 2$); Singer's Theorem 5 is a statement for $SU(N)$.

For S^3, the contradiction comes just from the fact that gauge transformations of different winding numbers cannot be connected (the circle of § 06 is the abelian version, by $\pi_0(\mathcal{G}_m) = \pi_1(U(1)) = \mathbb{Z}$). The higher homotopy groups of $U(1)$ are 0, so for S^3 and S^4 the fourth step does not work, and Singer too writes that in the abelian case "An integral submanifold provides the gauge for quantum electrodynamics." The obstruction comes from the topology made by non-commutativity and the condition at infinity Known .

08

How it has been avoided — the lattice, local gauges, the orbit space itself

The Faddeev–Popov procedure (4-08 § 03) is correct on a local slice but cannot be used globally as it is. Singer named the route of covering the orbit space by regions where local slices work and adding them up with a partition of unity (assigning to each piece of the cover a weight such that the weights add up to 1 everywhere), and wrote "This partition of unity argument would be useful if the covering used could be made explicit." Later work has avoided the obstruction by three routes.

- **The lattice.** Wilson's lattice integrates with Haar measure, so it is finite without fixing the gauge (4-02, 5-01) Known . The gauge that sets all edges of a tree joining all vertices (a maximal tree) to 1 is a global continuous slice for the based gauge group Known . Set $g = 1$ at the root vertex, and moving along the tree to neighbouring

vertices, fix g in turn so that each edge becomes 1 (if g_x is fixed on the edge $\ell : x \to y$, then $g_y = g_x U_\ell^{-1}$). g is determined uniquely by products and inverses of the edge elements, and depends continuously on U. Since the based gauge group acts freely and transitively on the configurations of the tree edges, each orbit has exactly one representative, and it can be chosen continuously. $SU(N)^E$ is not contractible ($\pi_3(SU(N)) = \mathbb{Z}$; for $SU(2)$, the degree of 1-12), so this does not contradict Singer's theorem. Instead the symmetry of the lattice is broken. Fixing a covariant gauge BRST-symmetrically (so that a formal symmetry remains even after fixing) produces 0/0 (7-11, 4-08 § 08; for BRST, 4-08 § 04).

- **Local gauges.** On Bałaban's renormalisation group on the three-dimensional periodic lattice, the problem description writes “One defines these gauges in phase cells: The choices vary locally in space-time, as well as on different length scales.” (Jaffe–Witten § 6.5) Known . It is a method of changing slices by place and by scale (7-06 § 05, 7-05 § 05). Bałaban's four-dimensional work is said to use the same idea (7-07 § 07; the specific choice in the four-dimensional papers has not been checked in the text).
- **The orbit space itself.** Stochastic quantisation does not fix the gauge and takes the orbit space as the state space. A review writes “The space O is a non-linear space if G is non-Abelian, which makes non-trivial even the construction of the state space” (Chevyrev 2022), and Chandra–Chevyrev–Hairer–Shen, on the two-dimensional torus, took as state space the space of orbits of connections as distributions (1-11) and constructed a stochastic dynamics on it (abstract) Known (7-09).

7-11 § 05 treats this as one of the obstacles other than divergence (the obstacle of the curved state space), and separates it into the part settled on the lattice and the 0/0 that remains when weak-coupling analysis needs gauge fixing.

What this chapter can say

Item	What is said	What is not said
Connections and orbits	Derivation of (1)–(9) Known	How to topologise the function spaces; the proof of the local slice theorem
Lattice models	Tables 1–3, Figure 1 Computed	The number or arrangement of copies in the four-dimensional continuum theory
Singer's theorem	The statement and the four-step outline (checked in the original paper) Known	Reproducing the proofs of Singer's Theorems 2 and 5
Detour	Feynman's overall picture Physics	Not a theorem

Item	What is said	What is not said
Yang–Mills	The meaning of the global obstruction to gauge fixing	Nothing is shown about the mass gap or the construction of the theory

Sources and reproduction

Item	Kind	Source / tools
Tables 1–3, Figure 1	**Computed on this machine**	Python 3.12, NumPy (`ring_copies.py`, `ring_bruteforce_check.py`, `su2_classes.py`). The figure computes (10)(11) on the spot
Singer's statement, quotations, Table 4	Primary source (text checked)	I. M. Singer, "Some remarks on the Gribov ambiguity", Commun. Math. Phys. 60 (1978) 7–12
Gribov copies	Bibliographic data only	V. N. Gribov, "Quantization of non-Abelian gauge theories", Nucl. Phys. B 139 (1978) 1–23 (the content in the text follows Singer's account in § 1; Singer cites the 1977 SLAC Translation 176 version)
The function (10) and its second derivatives	Primary source (§ I–II checked)	D. Mehta, N. S. Daleo, J. D. Hauenstein, C. Seaton, Phys. Rev. D 90 (2014) 054504, arXiv:1406.6678
Detour	Text § 1 checked / Feynman bibliographic data only	D. Karabali, V. P. Nair, Nucl. Phys. B 464 (1996) 135–152, arXiv:hep-th/9510157; R. P. Feynman, Nucl. Phys. B 188 (1981) 479
Quotations in § 08	Text checked / abstract	A. Jaffe, E. Witten, "Quantum Yang–Mills Theory" (Clay, 2000) § 6.5; I. Chevyrev, "Stochastic quantisation of Yang–Mills", J. Math. Phys. 63 (2022) 091101, arXiv:2202.13359 § 1; A. Chandra, I. Chevyrev, M. Hairer, H. Shen, Publ. Math. IHÉS 136 (2022) 1–147, arXiv:2006.04987
(7), values of homotopy groups	Known (bibliographic data only)	Weyl integration formula: T. Bröcker, T. tom Dieck, *Representations of Compact Lie Groups*, Graduate Texts in Math. 98, Springer (1985), Chapter IV; homotopy groups of spheres: H. Toda, *Composition Methods in Homotopy Groups of Spheres*, Ann. of Math. Studies 49, Princeton (1962) (following Singer's bibliography)

Next chapter to read: 3-05 Instantons and topology (the integer that separates S^4 bundles, and the orbits of least action). The quantum theory of gauge fixing is 4-08.

Instantons and topology — the integer in the lower bound on the action comes from the winding number at infinity

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

What this chapter shows — That the integer Q in the lower bound on the action of 3-03 comes from $\operatorname{tr} F \wedge F$ being a total derivative and from the winding number at infinity. The BPST solution written with quaternions, and how one of two instantons shrinks to a point. How the θ term changes the vacuum in quantum theory (the model of a particle on a circle).

Prerequisite chapters — 3-03 (self-duality and the lower bound (12)). For differential forms, 1-06; for pure gauge, 1-07; for the quantum mechanics of § 06 and § 07, 2-05. § 02, § 03, § 05 and § 06 of the detour chapter 1-12 (quaternions, degree, self-dual 2-forms, the fraction of the action inside) are worth reading; the minimum this chapter needs is defined in the text (quaternions and degree in § 03, the fraction inside in § 04).

Said first — § 01– § 05 are known mathematics of classical fields. The quantum discussion of § 06– § 08 is argument at the level of physics, not theorems. This chapter shows nothing about the mass gap.

Order of this chapter

1. An integer that does not change under continuous change — the winding number and conventions
2. tr F∧F is a total derivative — the integral is fixed by the boundary
3. The integer is fixed at infinity — degree (Table 1), SO(3) bundles, Detour: the sign
4. The BPST solution — one line in quaternions (Table 2)
5. Two instantons — one shrinks (Figure 1, Table 3)
6. The θ term — a particle on a circle (Figure 2, Table 4)

7. Vacua are joined by tunnelling — θ vacua (Table 5)
8. The lattice and θ in the problem statement — sectors and θ = π

01

An integer that does not change under continuous change — from the winding number in the plane

Topology deals with properties that do not change under continuous deformation. The winding number of a closed curve $z(t)$ that does not pass through the origin,

$$n = \frac{1}{2\pi i}\oint \frac{dz}{z} \in \mathbb{Z} \quad (1)$$

is an integer, so it cannot change when the curve is moved continuously; it changes only at the moment the curve crosses the origin. Example. $z = e^{it} + c$ (c real) gives 1 for $|c| < 1$ and 0 for $|c| > 1$. Moving c along the real axis, the curve passes through the origin at the moment it crosses $|c| = 1$, and only there does the value jump. The first Chern number worked the same way (1-07 § 08). This chapter derives, from the side of the field, where the integer Q in (12) of 3-03 comes from.

Conventions. A is an anti-Hermitian matrix-valued 1-form, $D = d + A$, $F = dA + A \wedge A$, and g is absorbed into A ($A = -ig\,A_{\text{phys}}$; 3-03 § 01, the table of notation). In this notation the action, the second Chern number and Q take the following form. Q is the Q of 3-03 and the k of 1-12 (the sign is in the detour of § 03).

$$S_E = -\frac{1}{g^2}\int \operatorname{tr} F \wedge *F, \qquad c_2 = \frac{1}{8\pi^2}\int \operatorname{tr} F \wedge F, \qquad Q = -c_2 \quad (2)$$

In the physics notation (the $D_\mu = \partial_\mu - igA_\mu$ of 3-03, without absorbing g into A), with $\tilde{F}_{\mu\nu} = \frac{1}{2}\varepsilon_{\mu\nu\rho\sigma}F_{\rho\sigma}$ (the components of the $*F$ of 3-03),

$$Q = \frac{g^2}{32\pi^2}\int F^a_{\mu\nu}\tilde{F}^a_{\mu\nu}\,d^4x = \frac{g^2}{8\pi^2}\int \sum_a \mathbf{E}^a\cdot\mathbf{B}^a\,d^4x$$

which is exactly the Q of (12) of 3-03.

02

tr F∧F is a total derivative — the Chern–Simons form

For U(1), $F \wedge F = d(A \wedge dA)$, and the integral was fixed by the boundary alone (1-06). In the non-abelian case it holds too, once one correction is added Known .

$$\operatorname{tr} F \wedge F = d\,\mathrm{CS}(A), \qquad \mathrm{CS}(A) = \operatorname{tr}\Big(A \wedge dA + \tfrac{2}{3} A \wedge A \wedge A \Big) \quad (3)$$

Checking it needs only one rule. For a matrix-valued p-form α and q-form β, the trace of the product of the component matrices does not change when the order is swapped, and only the sign from exchanging the differential-form parts remains.

$$\operatorname{tr}(\alpha \wedge \beta) = (-1)^{pq} \operatorname{tr}(\beta \wedge \alpha)$$

Let A^k be the wedge product of A taken k times. Expand $d \operatorname{tr} A^3$ into three terms by the product rule $d(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^p \alpha \wedge d\beta$ (1-06), and move dA to the front in each term.

$$\begin{aligned} \operatorname{tr}(dA \wedge A \wedge A) &= \operatorname{tr}(dA \wedge A^2) \\ -\operatorname{tr}(A \wedge dA \wedge A) &= +\operatorname{tr}(dA \wedge A^2) \qquad (p=1,\ q=3) \\ +\operatorname{tr}(A \wedge A \wedge dA) &= +\operatorname{tr}(dA \wedge A^2) \qquad (p=q=2) \end{aligned}$$

The sum is $d \operatorname{tr} A^3 = 3 \operatorname{tr}(dA \wedge A^2)$. Together with $d \operatorname{tr}(A \wedge dA) = \operatorname{tr}(dA \wedge dA)$ ($d\,dA = 0$), $d\,\mathrm{CS} = \operatorname{tr}(dA \wedge dA) + 2 \operatorname{tr}(dA \wedge A^2)$. Expanding $\operatorname{tr} F \wedge F = \operatorname{tr}\big((dA + A^2) \wedge (dA + A^2)\big)$ gives the same two terms ($\operatorname{tr}(A^2 \wedge dA) = \operatorname{tr}(dA \wedge A^2)$) and $\operatorname{tr} A^4$, which is 0 since $\operatorname{tr} A^4 = \operatorname{tr}(A \wedge A^3) = -\operatorname{tr}(A^3 \wedge A) = -\operatorname{tr} A^4$. The name comes from Chern–Simons (1974).

There are two consequences. First, if A is smooth on the ball B_R, Stokes' theorem (1-06) gives

$$\frac{1}{8\pi^2} \int_{B_R} \operatorname{tr} F \wedge F = \frac{1}{8\pi^2} \int_{S^3_R} \mathrm{CS}(A) \quad (4)$$

Example. For the solution of § 04 the right-hand side is $R = \rho, 3\rho, 10\rho$ at -0.5, -0.972, -0.999708, agreeing with the fraction inside in Table 1 of 1-12 (Table 2) Computed . Second, for δA that is 0 far away, the Bianchi identity (3-03 § 04) gives $\delta \operatorname{tr} F \wedge F = 2\,d \operatorname{tr}(\delta A \wedge F)$, and the integral does not change. Adding c_2 to the action does not change the equations, and the θ term of § 06 starts here.

03

The integer is fixed at infinity — pure gauge and degree

If the action is finite, far away $A \to -(dg)g^{-1}$ ($g : S^3 \to SU(2)$, pure gauge). For solutions, the basis is the abstract of Uhlenbeck (1982), "every Yang-Mills field over R4 with bounded functional (L2 norm) may be obtained from a field on S4" Known ; for general fields we assume decay far away (1-07). For pure gauge $dA = -A^2$, so CS $= \frac{1}{3}$ $\operatorname{tr}(g^{-1}dg)^3$, and the $R \to \infty$ of (4) is the following formula.

$$c_2 = \frac{1}{24\pi^2}\int_{S^3} \operatorname{tr}\left(g^{-1}dg\right)^3 = -\deg g, \qquad Q = \deg g \quad (5)$$

Checking with the identity map and with numbers

We use quaternions (1-12 § 02; this chapter needs only the following rules). $x = x_1 + x_2 i + x_3 j + x_4 k$ is multiplied with $i^2 = j^2 = k^2 = ijk = -1$, $\bar{x} = x_1 - x_2 i - x_3 j - x_4 k$, $|x|^2 = x\bar{x}$. The set of quaternions of length 1 is S^3, which under multiplication is the same group as SU(2). The degree $\deg g$ is the integer obtained by counting the preimages of a generic target point, with $+1$ for points that preserve orientation and -1 for points that reverse it (1-12 § 06).

The orientation of S^3 is the one that, as the boundary of the ball B^4, puts the outward normal first (at the point $x = 1$, $dx_2 \wedge dx_3 \wedge dx_4$ is positive). For the identity map $g(x) = x$, at the point $x = 1$, $g^{-1}dg = i\,dx_2 + j\,dx_3 + k\,dx_4$. The signed sum over permutations of i, j, k is -6, so $\operatorname{tr}(g^{-1}dg)^3 = -12\,dx_2 \wedge dx_3 \wedge dx_4$ (the quaternion 1 is the 2×2 identity matrix, with trace 2). It is invariant under left multiplication, so it is the same everywhere, and multiplying by the area $2\pi^2$ and dividing by $24\pi^2$ gives $c_2 = -1$. For general g too, the integrand is -1 times the pullback of the normalised volume form, and the integral is -1 times the degree Known .

Table 1 Numerical integration of $-\frac{1}{24\pi^2}\int_{S^3} \operatorname{tr}(g^{-1}dg)^3$ ($x \in S^3$; 64×64×128 points) Computed . The value changes only when the map before normalisation passes through 0

g	Integral	Degree
x^{-2}, x^{-1}, 1, x, x^2, x^3	-2, -1, 0, 1, 2, 3 (errors below 10^{-5} here too)	n
$\frac{x+c}{\lvert x+c\rvert}$ ($c = 0.5, 0.9$ / $1.1, 1.5$), $\frac{x^2+ci}{\lvert x^2+ci\rvert}$ ($c =$ 0.7 / 1.3)	1, 1 / 0, 0; 2 / 0.000002	1, 2 at $\lvert c\rvert <$ 1

Fields of finite action split into classes (sectors) by Q, which cannot be connected to one another continuously. This is the statement that SU(2) bundles over S^4 are classified by c_2 (1-07).

A group divided by its centre — SO(3) bundles

Taking as group $SO(3) = SU(2)/\mathbb{Z}_2$, SU(2) divided by its centre $\{\pm 1\}$ (1-04 § 07), the Lie algebra is the same, so the local formulas do not change. The difference shows in how the bundle is glued. On S^4 (fields on $\mathbb{R}^4$ that approach pure gauge far away), the gluing is a map from S^3 to the group, and since S^3 is simply connected, a map to SO(3) lifts to a map to SU(2). So every SO(3) bundle over S^4 comes from an SU(2) bundle Known .

On a closed four-dimensional manifold X with $H^2(X;\mathbb{Z}_2) \neq 0$ (for example the four-dimensional torus T^4), there are SO(3) bundles that do not lift to SU(2) bundles. The obstruction is the second Stiefel–Whitney class $w_2 \in H^2(X;\mathbb{Z}_2)$, and SO(3) bundles are said to be classified by w_2 and the first Pontryagin class p_1 (Dold–Whitney 1959) Known . Measuring the number by $-p_1/4$, so that it agrees with c_2 for bundles coming from SU(2), it is not necessarily an integer for $w_2 \neq 0$, and half-integers also appear for T^4. In physics this is said to be known as the fractional topological charge of 't Hooft's twisted boundary conditions (1979, 1981) Known . The integer Q of this chapter is about SU(2) bundles (or S^4). Centre symmetry is in 18-03.

Detour: the Q of physics and the c₂ of mathematics have opposite signs

A subsection you may skip.

Physics writes c_2 with $Q = \frac{g^2}{8\pi^2}\int \mathrm{tr}\, F_{\mathrm{phys}} \wedge F_{\mathrm{phys}}$ and mathematics with $F = -ig\, F_{\mathrm{phys}}$, so the signs differ by $(-i)^2 = -1$. Physics instantons are self-dual with $Q = 1$, while the mathematical literature often treats the anti-self-dual case ($c_2 \geq 0$) (3-06); the two are mirror images (1-12).

For how this book uses this, see 25-01 § 01 (the lattice setting of the small twisted T^2).

04

The BPST solution — written in one line with quaternions

With the quaternions $x = x_1 + x_2 i + x_3 j + x_4 k$ of § 03, reading i, j, k as $-i\sigma_1, -i\sigma_2, -i\sigma_3$, the purely imaginary quaternions are elements of su(2). With $dx = \sum_\mu e_\mu dx_\mu$ ($e_\mu = 1, i, j, k$), the solution of size ρ has the following form Known .

$$A = \mathrm{Im}\,\frac{x\, d\bar{x}}{|x|^2 + \rho^2}, \qquad F = \frac{\rho^2\, dx \wedge d\bar{x}}{(|x|^2 + \rho^2)^2} \quad (6)$$

The i components of $dx \wedge d\bar{x}$ are $-2(dx_1 \wedge dx_2 + dx_3 \wedge dx_4)$, and the components of j, k are also self-dual (1-12 § 03). So $F = *F$, and it is a solution (3-03 § 07). Far away, with $g = x/|x|$, $A \to \mathrm{Im}(x\, d\bar{x})/|x|^2 = g\, d\bar{g} = -(dg)g^{-1}$; g is the identity map of S^3 (degree 1),

so $Q=1$, $c_2=-1$. The density is $48\rho^4/(|x|^2+\rho^2)^4$, the fraction of the action inside radius R is $1-3/(1+u)^2+2/(1+u)^3$ ($u=R^2/\rho^2$; (9) of 1-12), and $S_E=8\pi^2/g^2$ is equality in the lower bound (the same property as (13) of 3-03). The abstract of BPST (1975) reads “We find regular solutions of the four dimensional euclidean Yang-Mills equations” Known . The mirror image $\mathrm{Im}(\bar{x}\,dx)/(\cdots)$ is anti-self-dual, with $Q=-1$.

Table 2 Checks of (6) and its mirror image ($\rho=1$; differences and numerical integration) Computed

Item	(6)	Mirror image
Difference from the closed formula, $F \mp *F$, c_2, Q	1.6×10⁻¹⁰, 0, -1.00000000, +1	1.8×10⁻¹⁰, 0, +1.00000000, -1
Right-hand side of (4) ($R=0.5,1,3,10$)	-0.104, -0.500, -0.972, -0.999708	Opposite sign

The parameters are 8: centre 4, size 1, and constant gauge rotation 3; for $Q=k$ there are $8k$ (Tong's lecture notes: $4kN$ for SU(N)) Known . The construction that gives all self-dual fields of finite action by linear algebra is ADHM (1978).

05

Two instantons — an action independent of the distance, and one that shrinks

A self-dual field has the least action in each Q, so for $Q=2$, wherever two lumps are placed the action is $16\pi^2/g^2$, and no force acts between the lumps. There is a concrete family. Here only, in the component notation of 3-03 ($A_\mu=A^a_\mu T^a$, $D_\mu=\partial_\mu-igA_\mu$), we set $g=1$. $\bar{\eta}$ is 't Hooft's symbol η of 3-03 § 08 with the sign of η_{ai4} changed ($\bar{\eta}_{aij}=\varepsilon_{aij}$, $\bar{\eta}_{ai4}=-\bar{\eta}_{a4i}=-\delta_{ai}$).

$$\phi=1+\sum_{i=1}^{k}\frac{\rho_i^2}{|x-a_i|^2},\qquad A^a_\mu=-\bar{\eta}_{a\mu\nu}\,\partial_\nu\ln\phi \tag{7}$$

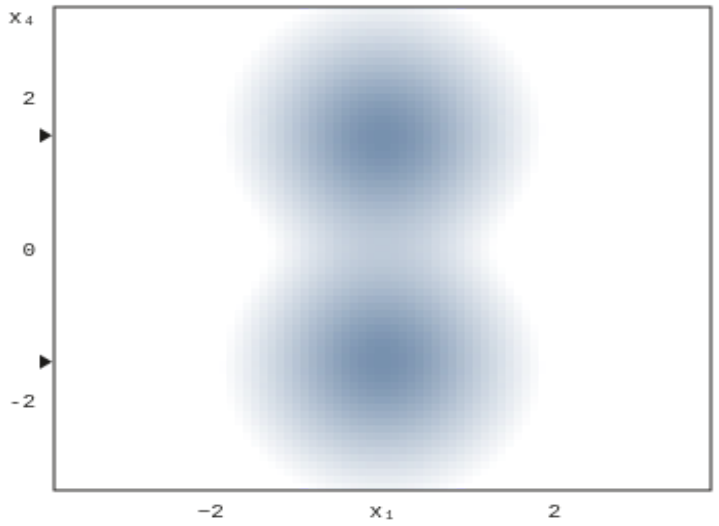
$1/|x|^2$ is a four-dimensional harmonic function, and this family is self-dual and $Q=k$ regardless of a_i,ρ_i. The A of (7) diverges at the poles a_i, but this singularity can be removed by a gauge transformation singular at that point. $k=1$ is thereby taken to the solution of § 04 translated to $a1$.

The density can be written with ϕ alone. The notation is $\Box=\sum_\mu\partial_\mu^2$ (the Euclidean four-dimensional Laplacian; not the d'Alembertian of 2-04), $\psi_\mu=\partial_\mu\phi$, $H_{\mu\nu}=\partial_\mu\partial_\nu\phi$, $|H|^2=\sum_{\mu\nu}H_{\mu\nu}^2$, $\psi\cdot H\psi=\sum_{\mu\nu}\psi_\mu H_{\mu\nu}\psi_\nu$, $|\psi|^2=\sum_\mu\psi_\mu^2$.

$$s=\tfrac{1}{4}F^a_{\mu\nu}F^a_{\mu\nu}=-\tfrac{1}{2}\,\Box\Box\ln\phi=\frac{|H|^2}{\phi^2}-\frac{4\,\psi\cdot H\psi}{\phi^3}+\frac{3|\psi|^4}{\phi^4}\qquad(x\ne a_i) \tag{8}$$

With two poles, the F from differences satisfies $F = *F$ and (8) with relative error 10^{-9} Computed . Corrigan–Fairlie (1977) treated solutions of this type as a relation between solutions of a scalar field theory of φ^4 type and solutions of the SU(2) gauge field. Jackiw–Nohl–Rebbi (1977) extended the known family of $5k$ parameters, representing the position and size of each lump, to solutions with $5k + 4$ parameters (both from the abstracts) Known .

Figure 1 Changing the distance d between the two poles with the knob changes the action density (log shading) on the cross-section $x_2 = x_3 = 0$ (horizontal x_1, vertical x_4, triangles are the poles). For small d a sharp peak rises at the origin



d = 3.00, s(0) = 2.100, 48/(d/2)⁴ (approximation for small d) = 9.481, total action = 16π² (Table 3)

Computed The numbers in this figure were computed on this machine. (8) is computed on the spot ($\rho_i = 1$, poles at $x_4 = \pm d/2$).

Table 3 The action of two poles ($\rho_1 = \rho_2 = 1$) (in units of $8\pi^2$; adaptive integration) Computed

d	4	1	0.5	0.2	0.1	0
Total action	2.000000	2.000000	2.000000	2.000000	2.000000	1.000000
Inside $\lvert x\rvert < d$	1.945544	1.140005	0.925267	0.896362	0.895881	—

The total action is $16\pi^2$ for every d. For small d the density at the origin approaches $48/(d/2)^4$ (7.661×10^6 against 7.68×10^6 at $d = 0.1$), and the 0.896 inside radius d is the fraction $1 - \frac{3}{25} + \frac{2}{125}$ for the BPST solution of size $d/2$ ((9) of 1-12). One lump shrinks while keeping its action, and at $d = 0$ it becomes the $Q = 1$ solution of $\phi = 1 + 2/\lvert x\rvert^2$, one lump's worth vanishing into a point. This is the concentration of 1-12 § 05 happening within a family of solutions (3-06).

06

The θ term — an angle invisible to the equations

By § 02, adding a constant multiple of Q to the action does not change the classical equations. In the quantum weight it stays imaginary even in imaginary time (4-02; the sign depends on conventions).

$$e^{-S_E[A]} \longrightarrow e^{-S_E[A]+i\theta Q[A]} \quad (9)$$

Known (i) In perturbation theory (the expansion around $A = 0$; 4-03) the integral of a total derivative is 0 at every order, and it cannot be seen. (ii) Since Q is an integer, θ is an angle of period 2π. (iii) In spacetime it is proportional to $\mathbf{E}^a \cdot \mathbf{B}^a$ (§ 01, 2-04); $\mathbf{E}$ changes direction under spatial reflection and $\mathbf{B}$ under time reversal, so $\theta \neq 0, \pi$ breaks spatial reflection and time reversal. The abstract of Jackiw–Rebbi (1976) describes θ vacua (§ 07) as "a family of quantum mechanical vacua, parametrized by a CP-nonconserving angle" (CP is the transformation combining the exchange of particles and antiparticles with spatial reflection). At θ = 0 and π it is invariant under time reversal (Gaiotto–Kapustin–Komargodski–Seiberg 2017, hereafter GKKS, abstract: "time reversal invariant at θ = 0 and θ = π").

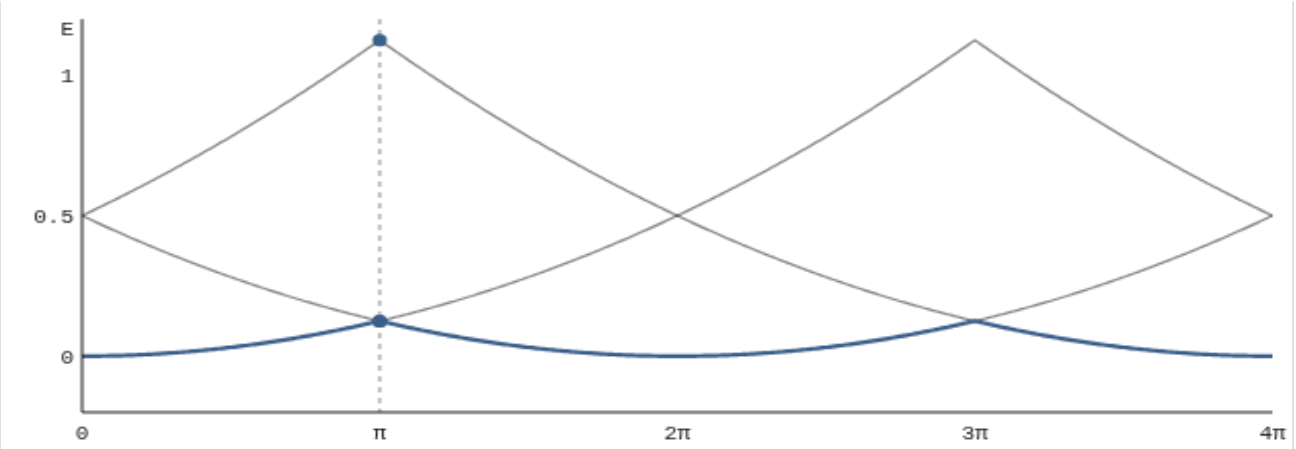
A particle on a circle — the smallest model of θ

There is quantum mechanics with the same structure (GKKS Appendix D.1). To a particle on the angle $q \sim q + 2\pi$ we add a term linear in the velocity.

$$L = \frac{1}{2}\dot{q}^2 + \frac{\theta}{2\pi}\dot{q}, \qquad E_n = \frac{1}{2}\left(n - \frac{\theta}{2\pi}\right)^2 \quad (n \in \mathbb{Z}) \quad (10)$$

The added term is $\frac{d}{dt}\frac{\theta q}{2\pi}$ and does not change the equations (2-01), but $\theta q/2\pi$ shifts by θ in one round, and the weight of periodic paths in imaginary time gets a factor $e^{i\theta \times \text{winding number}}$. This is the same form as (9). With periodic wave functions, the eigenvalues of the eigenfunctions e^{inq} of $H = \frac{1}{2}(-i\,d/dq - \theta/2\pi)^2$ are (10) (2-05). θ plays the role of a magnetic flux threading the circle (GKKS footnote 27; 2-03), and "for θ = π and only in this case the ground state is two-fold degenerate". Adding $-K\cos q$ opens the crossing, while for $-K\cos 2q$ the two-fold degeneracy remains exactly (same appendix: "still maintains some exact two-fold degeneracy").

Figure 2 Moving θ (horizontal axis 0–4π) with the knob moves the vertical line and the levels (dots). Switching the potential with the buttons between 0, $-K\cos q$ and $-K\cos 2q$ ($K = 0.4$) shows whether the crossing at $\theta = \pi$ opens or remains



θ = 1.00π, E_0 = 0.125, $E_1 - E_0$ = 0.000

Computed The numbers in this figure were computed on this machine. V = 0 is (10); the others are a table diagonalised with e^{inq} ($|n| \le 40$) (13 points of $0 \le \theta \le \pi$), extended by periodicity and symmetry.

Table 4 $E_1 - E_0$ ($K = 0.4$, $I = \hbar = 1$) Computed

Potential	θ = 0	π/2	π
0	0.5	0.25	0
$-K\cos q$	0.60436	0.46493	0.38481
$-K\cos 2q$	0.32981	0.21083	0

07

Vacua are joined by tunnelling — θ vacua

From here on the level is that of physics Physics . Setting $A_4 = 0$, the classical vacua are three-dimensional pure gauges $-(dh)h^{-1}$ (with $h : \mathbb{R}^3 \to SU(2)$ approaching 1 far away, regarded as $\mathbb{R}^3 \cup \{\infty\} = S^3$; a different sphere from the S^3 at infinity of § 03), numbered by the degree n, and the gauge transformation T of degree 1 shifts n by one. Applying (4) to a slab of imaginary time, Q is the difference of the vacuum numbers at $\tau \to \pm\infty$, and the BPST solution is a path in imaginary time from n to $n+1$. The weight is

$$e^{-8\pi^2/g^2}: \qquad g^2 = 1,\ 2,\ 4,\ 8 \text{ give } 5.1\times10^{-35},\ 7.2\times10^{-18},\ 2.7\times10^{-9},\ 5.2\times10^{-5}$$ (1

Computed , and since every derivative at $g^2 = 0$ is 0, it appears at no order of perturbation theory (the double well of 2-05, 4-07, 4-12). In 1976 't Hooft computed “the tunneling process described by the Belavin-Polyakov-Schwarz-Tyupkin field configuration”, and Jackiw–Rebbi and Callan–Dashen–Gross described the family of vacua (abstracts) Known . The states that are invariant under small gauge transformations (4-01) and change only by a phase under T are the θ vacua.

$$|\theta\rangle = \sum_{n\in\mathbb{Z}} e^{in\theta}|n\rangle, \qquad T|\theta\rangle = e^{-i\theta}|\theta\rangle \tag{12}$$

For the particle on a circle, the valleys $q = 2\pi n$ of the line before it is wrapped onto the circle are $|n\rangle$, $q \mapsto q + 2\pi$ is T, and (12) has the form of a Bloch wave (GKKS also read q as the Chern–Simons quantity of Yang–Mills). When the valleys are deep the lowest level becomes a narrow band, fixed by the action $S_0 = \int_0^{2\pi} \sqrt{2K(1-\cos q)}\,dq = 8\sqrt{K}$ of the path from valley to valley. The band has the form $E_0(0) + \frac{W}{2}(1-\cos\theta)$, and W approaches the asymptotic form of the Mathieu equation (a rewriting of DLMF 28.8.2, $\frac{16}{\sqrt{\pi}} K^{3/4} e^{-S_0}(1 - \frac{7}{64\sqrt{K}})$) (Table 5).

Table 5 The width $W = E_0(\pi) - E_0(0)$ of the lowest band of $-K\cos q$ (bisection with 60-digit decimal arithmetic; the "—" at the right end means the deviation cannot be measured in double precision) Computed

K	W	W / asymptotic form	Deviation from $1-\cos\theta$ / W
1	2.66699×10⁻³	0.98887	2.8×10⁻³
4	2.71024×10⁻⁶	0.99782	1.7×10⁻⁶
16	8.89113×10⁻¹³	0.99951	—

In the dilute gas approximation of Yang–Mills (regarding instantons as scattered far apart), the vacuum energy density is said to be proportional to $-\cos\theta\, e^{-8\pi^2/g^2}$ Physics . It is an approximation that regards typical configurations as superpositions of smooth solutions, and it is mathematically unsettled (A dictionary of scales and limits). More precisely, in four dimensions the coupling runs with the scale (4-06), so the integral over the instanton size ρ diverges on the side of large ρ, and the approximation is said to be uncontrolled even at the level of physics Physics . The quantity that does not depend on the approximation is the topological susceptibility $\chi t = \langle Q^2\rangle/V$ (the second derivative of $E(\theta) = -\frac{1}{V}\ln\sum_Q Z_Q e^{i\theta Q}$ at $\theta = 0$).

For how this book uses this, see 25-02 § 03 (counting the string tension with a dilute gas of fractional instantons).

08 Integers on the lattice, and θ in the problem statement

The space of fields on the lattice (5-01) is connected, and “the concept of a topological sector has therefore no a priori well-defined meaning” (Lüscher 2010) Known . Lattice fields with small action density have an integer Q, which is said to return to the Chern number in the classical continuum limit (that they have an integer is from the abstract of Lüscher 1982; 7-11 § 09). For SU(3), the fraction of fields V smoothed by the Wilson flow that satisfy $\max_p \mathrm{Re}\,\mathrm{tr}\{1 - V(p)\} < 0.067$ is 0%, 8% and 70% at $a = 0.1,\ 0.07,\ 0.05$ fm, and the sector is “a dynamical property of the theory” (same, 2010) Physics . The

construction of Magnen–Rivasseau–Sénéor (1993) for SU(2) with a fixed infrared cutoff (a cutoff on the long-distance side) is restricted to the trivial sector (7-07) Known . In HMC the effective dynamical critical exponent of Q^2 is about 5 (Schaefer–Sommer–Virotta 2011; 5-07) Physics .

Neither the θ term nor instantons appear in the problem statement of Jaffe–Witten (full-text search). GKKS argue, from the 't Hooft anomaly of time reversal and centre symmetry (18-03) (an obstruction to making a trivial vacuum while keeping both symmetries), that “at θ = π the vacuum cannot be a trivial non-degenerate gapped state”, and state that the vacuum at θ = 0, by contrast, is trivial, gapped and non-degenerate (both at the level of physics) Physics . "A unique vacuum with a gap" is a claim that depends on θ, and $-K\cos 2q$ and $\theta = \pi$ in Figure 2 are its quantum-mechanical version.

What this chapter can say

Item	What is said	What is not said
Integrality of Q	Solutions of finite action (Uhlenbeck), and fields assumed to decay to pure gauge far away Known	General fields of finite action without the decay assumption
Classical theory and models	(3)–(10), Tables 1–5, Figures 1 and 2 Known Computed	Reproducing the proofs of Uhlenbeck and ADHM. The model's levels are not those of Yang–Mills
Quantum theory	θ vacua, the dilute gas, the lattice, the statements of GKKS Physics	Not theorems. Nothing is shown about the mass gap or the construction of the theory

Sources and reproduction

Item	Kind	Source / tools
BPST, Uhlenbeck, ADHM, the family (7), tunnelling, θ vacua and CP	Abstract	Belavin–Polyakov–Schwartz–Tyupkin, PLB 59 (1975) 85 / Uhlenbeck, CMP 83 (1982) 11 / Atiyah–Hitchin–Drinfeld–Manin, Phys. Lett. A 65 (1978) 185 / Corrigan–Fairlie, PLB 67 (1977) 69 / Jackiw–Nohl–Rebbi, PRD 15 (1977) 1642 / 't Hooft, PRD 14 (1976) 3432 / Jackiw–Rebbi, PRL 37 (1976) 172 / Callan–Dashen–Gross, PLB 63 (1976) 334
Chern–Simons form	Bibliographic data only	Chern–Simons, Ann. of Math. 99 (1974) 48

Item	Kind	Source / tools
Classification of SO(3) bundles, fractional topological charge (SO(3) bundles in § 03)	Bibliographic data only (content not checked against the text)	A. Dold, H. Whitney, Ann. of Math. 69 (1959) 667 / G. 't Hooft, Nucl. Phys. B 153 (1979) 141 / G. 't Hooft, Commun. Math. Phys. 81 (1981) 267
$-\cos\theta$ of the dilute gas, divergence of the size integral	Not checked	The text of the primary sources not checked
Number of parameters, $\theta = \pi$, the particle on a circle, the lattice	Text	Tong, arXiv:hep-th/0509216 § 1.1.2, § 1.2 (eq. (1.13)) / Gaiotto–Kapustin–Komargodski–Seiberg, JHEP 05 (2017) 091 (abstract, Appendix D.1) / Lüscher, JHEP 08 (2010) 071 § 4
Topological charge on the lattice, freezing, the trivial sector	Abstract	Lüscher, CMP 85 (1982) 39 / Magnen–Rivasseau–Sénéor, CMP 155 (1993) 325 / Schaefer–Sommer–Virotta, NPB 845 (2011) 93
The problem statement, the asymptotic form of Mathieu	Text	Jaffe–Witten, *Quantum Yang–Mills Theory* (Clay) / NIST DLMF (28.8.2)
Tables 1–5, (11), figures	Computed	Python 3.12, NumPy, SciPy, decimal (standard library; the 60-digit arithmetic of Table 5) (`winding.py`, `bpst.py`, `two_instantons.py`, `rotor.py`). The figures are plain JavaScript

Next chapter to read: 3-06 Detour: Donaldson and exotic R^4. The dictionary of topological charge is 18-04.

Detour: Donaldson and exotic R⁴ — the space of solutions of classical Yang–Mills tells smoothness in four dimensions apart

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

What this chapter shows — What an exotic structure is. Intersection forms and the map of topology (Freedman, Rokhlin). Donaldson's line of argument, which reads the space of solutions of self-dual connections as a cobordism and derives that definite intersection forms can be diagonalised, and why exotic structures on $\mathbb{R}^4$ follow from it.

Prerequisite chapters — 3-05 (the BPST solution, the limit of shrinking to a point). For smooth structures, 1-05; for an entry to intersection forms, 1-12 § 07; for the orbit space and reducible connections, 3-04 § 04. Homology, signature, cobordism, index and the group of conformal transformations are defined minimally where they appear.

Said first — This is a detour, and a prerequisite for no chapter. It follows only the line of known mathematics and does not reproduce the body of the analysis. It shows nothing about quantum Yang–Mills or the mass gap.

Order of this chapter

1. The same shape, yet different — homeomorphism and diffeomorphism
2. The intersection form — a matrix that counts intersection points
3. A map of the topological world — Freedman and Rokhlin, Detour: in five or more dimensions
4. Donaldson's theorem — the counting lemma (Table 1)
5. The moduli space — a five-dimensional space of solutions (Figure 1)
6. Reading the space of solutions as a cobordism — cones, the end and the signature (Figure 2)
7. Exotic R⁴ — a set that no sphere can enclose (Table 2)

8. Two faces of the same equation — classical and quantum

01

The same shape, yet different — homeomorphism and diffeomorphism

Being the same in the topological sense (homeomorphic) means there is a one-to-one correspondence that is continuous in both directions. For smooth manifolds (1-05) we ask for a correspondence that is smooth in both directions (a diffeomorphism).

$$X \underset{\text{diffeomorphic}}{\cong} Y \implies X \underset{\text{homeomorphic}}{\cong} Y, \qquad \text{the converse does not hold in general} \quad (1)$$

When two are homeomorphic but not diffeomorphic, one is called an "exotic" structure on the other. Structures that differ only in how the atlas is written (the $x \mapsto x^3$ of 1-05) are related by a diffeomorphism, so they do not count. The first example is said to be the seven-dimensional sphere (Milnor 1956) Known .

Example. The smooth structure on $\mathbb{R}^n$ is unique for $n \neq 4$. The introduction of Gompf (1983) writes “it is a standard fact that for n ≠ 4, exotic R^n's cannot exist”, and continues that applying the high-dimensional theory of smoothing, $\mathbb{R}^4$ should have one structure (“there should only be one structure on R4, as it has no cohomology”) Known . In fact there are uncountably many (§ 07, the article on four dimensions). Why is the subject of this chapter.

02

The intersection form — an integer matrix that counts intersection points of surfaces

The closed surfaces inside a closed (compact, without boundary) oriented four-dimensional manifold X fall into classes that do not change under continuous motion, and classes can be added. This collection is $H_2(X;\mathbb{Z})$, and for simply connected X it is $\mathbb{Z}^r$ ($r = b_2$; the Betti number of 1-06 § 08). Precisely, the elements of H_2 are homology classes, and here we use only the intuition (in four dimensions every class is said to be representable by an embedded surface). Two surfaces in general position meet at finitely many points (1-12 § 07), and the sum of ± 1, fixed by orientations, does not change when the surfaces are moved.

$$Q(\Sigma_1, \Sigma_2) = \sum_{p \in \Sigma_1 \cap \Sigma_2} (\pm 1), \qquad Q : \mathbb{Z}^r \times \mathbb{Z}^r \to \mathbb{Z} \ \text{ is symmetric with } \det Q = \pm 1 \quad (2)$$

$\det Q = \pm 1$ (unimodular) is due to Poincaré duality Known . The numbers read from the form are the rank r, the signature $\sigma = b^+ - b^-$ (the number of positive eigenvalues minus the number of negative ones), and the parity (even if every $Q(a,a)$ is even). If $b^- = 0$, it is positive definite.

Example. In $\mathbb{CP}^2$ two lines $\mathbb{CP}^1$ meet at one point, so (1); $\mathbb{CP}^2$ with the orientation reversed gives (-1); and $S^2 \times S^2$ gives $H = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. The connected sum (remove a ball and glue) becomes a direct sum, and $\#r\,\mathbb{CP}^2$ is the identity matrix $\langle 1 \rangle^r$ (the standard form). The form of the Kummer surface (K3) is $E_8 \oplus E_8 \oplus H^{\oplus 3}$ (E_8 is the negative-definite sign; Gompf writes this $H^{\oplus 3}$ as H of rank 6) Known .

The smallest example of a positive-definite even form is E_8. The following Cartan matrix has determinant 1 and is even and positive definite Computed . Gompf states that the only even definite unimodular form of rank 8 is E_8 ("the unique symmetric form over Z which has rank 8, is even, and is negative definite").

$$E_8 = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 2 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 2 \end{pmatrix} \quad (3)$$

03

A map of the topological world — Freedman's theorem and Rokhlin's restriction

Seen topologically only, four dimensions resembles higher dimensions. Freedman (1982) showed that every symmetric form of determinant ± 1 can be realised as the form of a closed simply connected topological manifold, and that the homeomorphism class is fixed by the form and a $\mathbb{Z}/2$ quantity (the Kirby–Siebenmann obstruction; 0 if smooth) (as presented in Donaldson 1983 § I.1) Known . There is also a topological manifold $|E_8|$ whose form is E_8.

One restriction from smoothness was already known. As Donaldson presents it, "Rohlin's theorem asserts that any even form coming from a smooth simply connected 4-manifold has signature divisible by 16" Known .

$$X \text{ smooth and simply connected, and } Q \text{ even} \implies \sigma(X) \equiv 0 \pmod{16} \quad (4)$$

Example. E_8 ($\sigma = 8$) is excluded by (4), and $|E_8|$ has no smooth structure. $E_8 \oplus E_8$ ($\sigma = 16$) passes (4), and (4) says nothing about odd $E_8 \oplus \langle 1 \rangle$. What settled these is the theorem of § 04.

Detour: why smoothness does not matter in five or more dimensions

A subsection you may skip.

In higher dimensions, the key is said to be the operation of removing an extra pair of intersection points of opposite sign by sliding along a disc spanning a loop joining the two points (called the Whitney trick). In five or more dimensions the disc in general position meets nothing else, but in four dimensions, with $2 + 2 = 4$, the disc itself meets surfaces at points (1-12 § 07). Freedman showed that an infinite tower replacing the disc (a Casson handle) is homeomorphic to the open 2-handle $D^2 \times \mathbb{R}^2$. Precisely, a Casson handle is an open four-dimensional manifold made by thickening a disc with intersection points (a kinky handle) and, for each intersection point, gluing on another at the next stage, repeated infinitely. From Donaldson's theorem Gompf writes “they are not all diffeomorphic to it” Known .

04

Donaldson's theorem — definite forms can be diagonalised

Donaldson's diagonalisation theorem (Theorem 1 of Donaldson 1983) is “Let X be a compact smooth simply connected oriented 4-manifold with the property that the associated form Q is positive definite. Then that form is equivalent, over the integers, to the standard diagonal form” Known . That is, a change of integer basis turns it into the identity matrix. Reversing the orientation, the same holds for negative definite forms, and the simply connected assumption is said to have been removed later (Donaldson 1987). The proof uses one property of numbers, a lemma on the number $n(Q)$ of a in $Q(a, a) = 1$, counting $\pm a$ as one pair (Lemma 2 of the paper) Known .

$$n(Q) = \#\{\pm a \in \mathbb{Z}^r : \ Q(a,a) = 1\}, \qquad n(Q) \le r, \qquad \text{equality} \iff Q \cong \langle 1 \rangle^r \ (5)$$

The proof is by induction on the rank. If $Q(a, a) = 1$, every b splits with integer coefficients as $b = m\,a + c$ ($m = Q(b, a)$, $c \perp a$), and $Q(b, b) = m^2 + Q(c, c)$. Since the form is positive definite, $Q(b, b) = 1$ is $b = \pm a$ or "$m = 0$ and $Q(c, c) = 1$", so $n(Q) = n(Q|_{a^\perp}) + 1$, and the rank also drops by 1. When the a run out, the form is standard if the remaining rank is 0, and otherwise $n(Q) < r$.

Example. Table 1 gives $n(Q)$ counted by enumeration Computed . D_n^+ is the union of the set D_n of all integer vectors with even sum of components and that set shifted by $(\frac{1}{2},$

$\dots, \frac{1}{2}$) (an integer form of determinant 1 when n is a multiple of 4). Even forms that pass (4) are excluded by the theorem if $n(Q) < r$.

Table 1 $n(Q)$ of positive-definite unimodular forms (determinant 1; by enumeration) Computed

Form	r	Parity	$n(Q)$	Norm 2	(4)	Theorem
$\langle 1\rangle^8$ ($\#8\,\mathbb{CP}^2$)	8	Odd	8	112	—	Standard form
E_8	8	Even	0	240	excluded	excluded
$E_8 \oplus \langle 1\rangle$	9	Odd	1	240	—	excluded
D_{12}^+	12	Odd	0	264	—	excluded
$E_8 \oplus E_8$, D_{16}^+	16	Even	0	480	passes	excluded

05

The tool: the moduli space of self-dual connections

Conventions. We use the same geometric notation as 3-05 (anti-Hermitian A, $c_2 = \frac{1}{8\pi^2}\int \operatorname{tr} F \wedge F$). Donaldson (1983) also uses the same sign, with an SU(2) bundle of $c_2(P) = -1$ over a positive-definite X, and self-dual connections Known . Self-dual means $F^- = 0$, where F^- is the anti-self-dual part of F (1-12 § 03; the same condition as the $F = *F$ of 3-03). Many textbooks reverse the orientation and write in terms of negative definite forms and anti-self-duality ($c_2 = k > 0$) (the detour of 3-05, 1-12 § 07). Changing the orientation, it is the same story. The collection of solutions with those related by gauge transformations identified is the moduli space, the part of the orbit space of 3-04 made of solutions.

$$\mathcal{M}_k = \{A : F_A^- = 0\}/\mathcal{G} \subset \mathcal{A}/\mathcal{G}, \qquad k = -c_2(P), \qquad \dim \mathcal{M}_k = 8k - 3(1 - b_1 + b^-) \quad (6)$$

k is the Q of 3-05. b_1 is the number of one-dimensional holes (the Betti number of 1-06 § 08; 0 if simply connected). The dimension comes from the **index** of the linearised equation. The index is an estimate of the dimension of the solution space of a linear equation; for a finite system of linear equations it corresponds to "the number of unknowns − the number of equations" (precisely, the dimension of the kernel minus the dimension of the cokernel; here we use only the name). Donaldson (1983) § II writes this index as $8|c_2(P)| - \frac{3}{2}(\chi(X) - \tau(X))$ Known . τ is the signature σ and $\chi = b_0 - b_1 + b_2 - b_3 + b_4$ is the Euler number (a relative of the $V - E + F$ of 1-06); in closed four dimensions $b_0 = b_4 = 1$, $b_3 = b_1$, $b_2 = b^+ + b^-$, so $\frac{3}{2}(\chi - \tau) = 3(1 - b_1 + b^-)$, that is, (6). In the anti-self-dual convention b^- is replaced by b^+.

For S^4 ($b_1 = b^- = 0$), $8k - 3$. Removing from the $8k$ parameters of 3-05 § 04 the 3 constant gauge rotations (identified because they are related by gauge

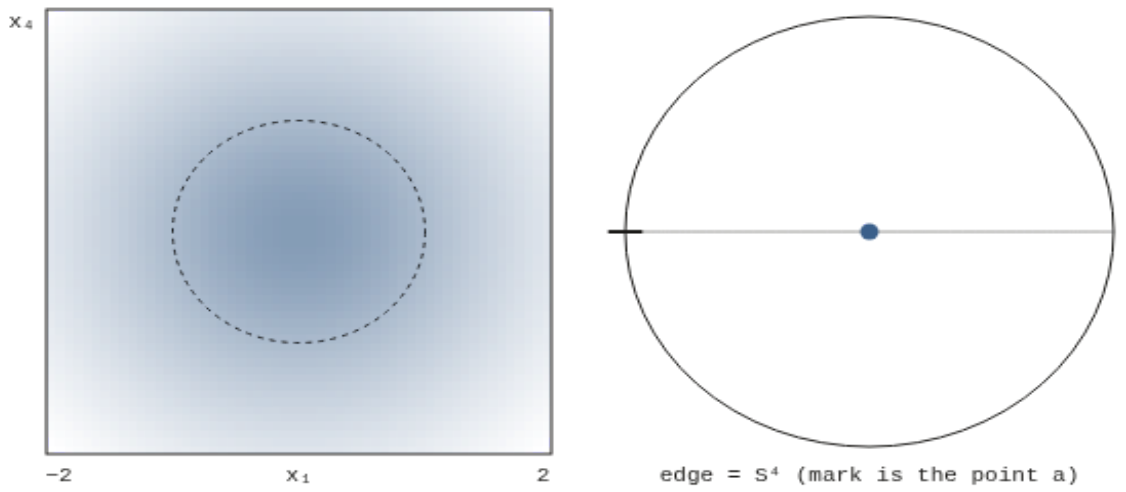
transformations) gives $8k-3$, which for $k=1$ is 5: centre 4 + size 1. For the X of the theorem ($b_1 = b^- = 0$, $k=1$) it is also 5.

Example. For $k=1$ on S^4, Donaldson writes $\mathcal{M}(S^4) = SO(5,1)^+/SO(5)$ as the open five-dimensional ball B^5. $SO(5,1)^+$ is the group of conformal transformations (angle-preserving transformations) of S^4. Donaldson explains that a solution is fixed by a "centre" and a "size", and states for the closed ball $\bar{B}^5 = \mathcal{M}(S^4) \cup S^4$ that “a point x of S4 represents the limit point of a sequence of connections whose curvature becomes concentrated in diminishing balls about x” Known . The limit of "shrinking to a point" in 3-05 § 05 is the edge. Figure 1 reads the centre a (on the x_1 axis) and the size λ as a point $w = a + i\lambda$ of the upper half-plane, and is the slice mapped by the linear fractional transformation $z = (w-i)/(w+i)$ that maps the upper half-plane to the unit disc (coordinates for the figure).

The hyperbolic distance in Figure 1

The hyperbolic distance in the figure's readout is the shortest length joining the points (a,λ) and $(0,1)$ when the length of curves in the upper half-plane $\lambda > 0$ is measured as $\int \sqrt{da^2+d\lambda^2}/\lambda$. This metric $(da^2+d\lambda^2)/\lambda^2$ (a slice of the metric of upper half-space) is invariant under the group of conformal transformations $SO(5,1)^+$, but it is not the L^2 metric that enters naturally on the space of solutions (the one measuring the size of a change of connection δA by $\int |\delta A|^2$). The hyperbolic distance of $a=0$ is $|\ln\lambda|$ ($\ln(1/\lambda)$ if $\lambda \le 1$) Computed . In this metric the edge is infinitely far away. However small the size λ is made, it remains a solution, and the family does not break off before the edge because in four dimensions the action does not depend on the scale (1-12 § 04).

Figure 1 Moving the size λ (log) and the centre a with the knobs moves the action density on the left (cross-section $x_2 = x_3 = 0$, log shading, half the action inside the dotted circle) and the point in the disc on the right (a two-dimensional slice of B^5). As $\lambda \to 0$ the point goes to the position a on the edge (S^4)



λ = 1.00, a = 0.00, inside radius 0.1 = 0.0003, 1 − |z| = 1.00, hyperbolic distance = 0.000

Computed The numbers in this figure were computed on this machine. The density $48\lambda^4/(|x-a|^2+\lambda^2)^4$ ((9) of 1-12) and z are computed on the spot. Static values: shrinking as $\lambda = 1,\ 0.1,\ 0.01$ at $a = 0$ gives $1 - |z| = 1,\ 0.182,\ 0.0198$, hyperbolic distance $0,\ 2.30259,\ 4.60517$, fraction of the action inside radius 0.1 $0.0003,\ 0.5000,\ 0.9997$ (the closed formula and numerical integration agree).

06

Reading the space of solutions as a cobordism — cones, the end and the signature

The $\mathcal{M}$ of (6) consists of three parts (Donaldson 1983 § I.2, § II, § III) Known .

(i) Irreducible points. Connections whose holonomies (1-07) have no common invariant subspace; a small perturbation of the equation makes them an oriented smooth five-dimensional manifold.

(ii) Reducible points. The case where the bundle splits as $E = L \oplus L^{-1}$ and the connection preserves the splitting; from $c_2(E) = -c_1(L)^2$, $Q(\alpha,\alpha) = 1$ ($\alpha = c_1(L)$). Since $b^- = 0$, there is a self-dual connection for each $\pm\alpha$; since $b_1 = 0$, it is unique up to gauge equivalence, and their number is $n(Q)$. The group of gauge transformations that leaves the connection fixed (the group that fixes) is as large as U(1), so these are the "corners" of 3-04 § 04. Precisely, at points where the group that fixes is large, the orbit space stops being a smooth manifold and the neighbourhood takes the shape of a cone. The neighbourhood is $\mathbb{C}^3/S^1$: dividing the unit sphere S^5 of $\mathbb{C}^3$ by the circle S^1 gives $\mathbb{CP}^2$, and adding the radial direction gives a cone over $\mathbb{CP}^2$.

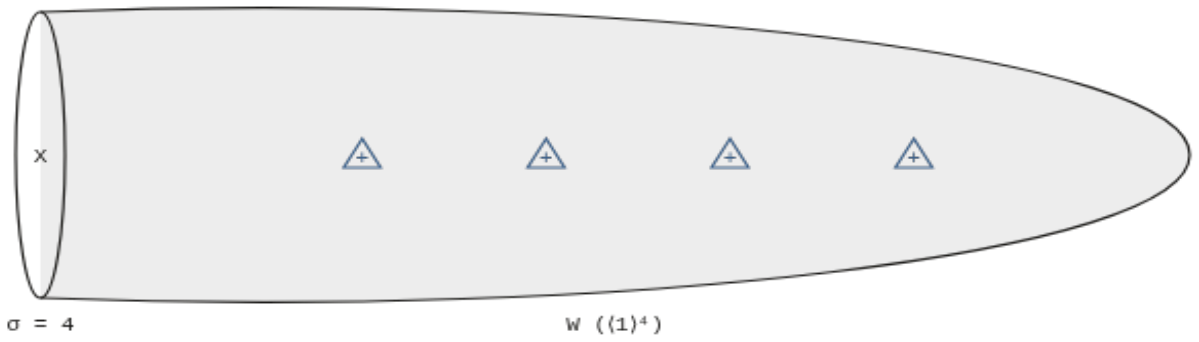
(iii) The end. Solutions of small size look like BPST solutions concentrated at one point of X. Taubes (1982) constructed them under $b^- = 0$. Donaldson § III uses the convergence results of Uhlenbeck (1982) to show that sequences that escape are only those that concentrate. Theorem 11 states that a neighbourhood of the end is diffeomorphic to $X \times (0, \lambda_0)$. The edge of Figure 1 is here a copy of X.

Cutting out the neighbourhoods of the cone vertices and gluing X onto the end gives an oriented compact five-dimensional manifold W whose boundary is X and $n(Q)$ copies of $\pm\mathbb{CP}^2$ (the relation of sharing a boundary is called cobordism). The signature is “an invariant of oriented cobordism”. That it is a cobordism invariant means that the signature of the whole boundary of an oriented compact five-dimensional manifold (here X and the sum of $n(Q)$ copies of $\pm\mathbb{CP}^2$, with orientations) is 0; from this follows $\sigma(X) = \sum \varepsilon_i\, \sigma(\mathbb{CP}^2)$ ($\varepsilon_i = \pm 1$ is how the orientations enter), and the following can be said.

$$r = \sigma(X) = \sum_{i=1}^{n(Q)} \varepsilon_i\, \sigma(\mathbb{CP}^2) \ \le\ n(Q)\ \le\ r \qquad (\varepsilon_i = \pm 1,\ \sigma(\mathbb{CP}^2) = 1) \tag{7}$$

The reasons for the four relations are as follows. The first equality: since the form is positive definite, $b^- = 0$, so $\sigma = b^+ = r$. The second equality: cobordism invariance (the paragraph above). The first inequality: each $\varepsilon_i \le 1$, so the sum is at most the number of terms $n(Q)$. The second inequality: (5). Both ends are r, so everything in between is equality, $n(Q) = r$, and the equality case of (5) gives the standard form. Example. For $E_8 \oplus E_8$ there are no cones and the boundary is only X, so $\sigma = 16$ cannot be 0, a contradiction. Figure 2 is the balance for each form. Most of the paper is the analysis of (i)–(iii).

Figure 2 Choosing a form with the buttons changes the schematic of W assuming there were a smooth X with that form (the left edge is X, triangles are cones) and the balance of (7). If there are too few cones, a contradiction (dashed line)



Q = ⟨1⟩⁴ r = σ(X) = 4, number of cones n(Q) = 4 (7) : 4 ≤ 4 holds → standard form (4) : odd, so it says nothing

Computed The numbers in this figure were computed on this machine. The rank and $n(Q)$ are the enumerated values of Table 1 ($\langle 1 \rangle^4$ is $n(Q) = 4$). Without JavaScript, putting r and $n(Q)$ from Table 1 into (7) gives the same verdict.

07

Exotic R⁴ — a set that no smooth sphere can enclose

The introduction of Taubes (1987) writes “Donaldson's theorem … has the surprising corollary that there exists an exotic smooth structure on R4. This corollary was deduced by M. Freedman” Known . The line has three steps (as presented by Taubes). (a) $|E_8 \oplus E_8|$ cannot be smoothed (§ 04). (b) With one point removed it can be smoothed (Freedman), and the neighbourhood of the removed point is diffeomorphic, outside a compact set, to the infinity of some smooth R homeomorphic to $\mathbb{R}^4$. (c) If R were standard, one could cut along a large smooth S^3 and cap it with a ball, and $|E_8 \oplus E_8|$ would become smooth. So the following can be said.

$$R \underset{\text{homeomorphic}}{\cong} \mathbb{R}^4, \qquad R \underset{\text{diffeomorphic}}{\not\cong} \mathbb{R}^4 \tag{8}$$

Gompf (1983) made a R^4_{DF} of the same kind from K3 and $\#3\, S^2 \times S^2$, and writes “there is a compact set in R4DF … which cannot be enclosed by any smooth S3. This

behavior is impossible in the standard R4." Known .

Example. Gompf § 2 uses $\mathbb{CP}^2\#9\mathbb{CP}^2$ (form $\langle 1\rangle \oplus 9\langle -1\rangle$, basis $e_0, \dots, e_9$) instead of the Kummer surface to make another exotic $\mathbb{R}^4$. The key is $x = 3e_0 + e_1 + \cdots + e_8$. $x \cdot x = 9 - 8 = 1$, and the orthogonal complement $E_8 \oplus \langle -1\rangle$ (negative definite and not diagonalisable) contradicts the theorem, so cutting off by surgery (cutting out part of a manifold and gluing it back differently) the part of form $\langle 1\rangle$ spanned by x is possible topologically but cannot be done smoothly. Table 2 is the check Computed .

Table 2 The orthogonal complement of x in Gompf § 2 (inside $e_0, \dots, e_8$) Computed

Basis	$e_i - e_{i+1}$ $(i = 1, \dots, 7)$ and $e_0 + e_1 + e_2 + e_3$ (norm -2)
Gram matrix	-1 times (3) (agreeing after changing the sign of the last basis vector). Determinant 1, even, 8 negative eigenvalues
Number of vectors of norm -2, -4	240, 2160 (the same as E_8)

How many are there? Taubes showed, by gauge theory on manifolds with periodic ends, that the R_r obtained as images of balls of radius r under the homeomorphism of (b) are not diffeomorphic to one another for distinct r larger than some r_0. Theorem 1.1 is "There exists an uncountable family of diffeomorphism classes of oriented 4-manifolds which are homeomorphic to R4." Known . r runs over the interval $r > r_0$, and distinct r give distinct classes, so there are at least as many as there are real numbers (the classes of manifolds have at most the cardinality of the reals, so it is said to be exactly the cardinality of the continuum). On the other hand, whether S^4 has an exotic structure is said to be unknown.

08

Two faces of the same equation — a classical tool and a quantum problem

Although the equation is the same as in 3-05, Donaldson uses the shape of the collection of solutions as a tool, and uses neither probability nor time evolution. What the problem statement (0-03) asks about is the probability measure weighted by the same action (4-02) and the spectrum of its Hamiltonian.

Classical: $\mathcal{M} = \{F_A = *F_A\}/\mathcal{G}$ — study its shape, Quantum: $d\mu(A) = \frac{1}{Z} e^{-SE[A]} \mathcal{D}$

On the quantum measure, smooth classical solutions are expected not to be typical configurations (by analogy with the Cameron–Martin space of 1-09) Physics . The "four" that matters in the problem is the four at which the dimension of the coupling constant vanishes (0-04). Example. § 2 of the problem statement says "Classical properties of gauge theory are within the reach of established mathematical methods", and after describing the role in the study of four-dimensional manifolds, continues

“one does not yet have a mathematically complete example of a quantum gauge theory in four-dimensional space-time” Known . This was written in 2000; the committee's assessment is that the construction published in 2026-11 built a quantum Yang–Mills theory in four dimensions (6-00).

There are two bridges at the level of physics. The abstract of Witten (1994) writes, on Donaldson's invariants, “instead of defining four-manifold invariants by counting SU(2) instantons, one can define equivalent four-manifold invariants by counting solutions of a non-linear equation with an abelian gauge group” (background in 17-11) Physics . § 5 of the problem statement says that the clustering that follows from the mass gap “may make it possible to apply mathematical results established on R4 to any 4-manifold”, and concludes that the gap “may also be important in mathematical applications of four-dimensional quantum gauge theories to geometry” Physics .

What this chapter can say

Item	What is said	What is not said
Intersection forms	Definitions and examples, the proof of (5), Tables 1 and 2 Known Computed	Proofs of Poincaré duality and of the theorems of Rokhlin and Freedman
Donaldson's theorem	The statement and the line of (6)(7), Figures 1 and 2 Known Computed	The analysis of perturbation, orientation, convergence and the end. The hyperbolic distance of Figure 1 is that of a metric invariant under conformal transformations, not the L^2 metric on the space of solutions. Figure 2 is a schematic
Exotic $\mathbb{R}^4$	The line of (8), the statements of Gompf and Taubes Known	Reproducing the proofs of Freedman and Taubes
Relation to physics	The words of the problem statement and of Witten's abstract Physics	Nothing is shown about the construction of quantum Yang–Mills or the mass gap

Sources and reproduction

Item	Kind	Source / tools
Quotations and line in § 03– § 06, the dimension in (6) (the index formula of § II)	Text (the relevant parts of § I– § III)	S. K. Donaldson, “An application of gauge theory to four-dimensional topology”, J. Differential Geom. 18 (1983) 279–315

Item	Kind	Source / tools
Quotations and examples in § 01– § 03 and § 07	Text (§ 0– § 2)	R. E. Gompf, “Three exotic R^4's and other anomalies”, J. Differential Geom. 18 (1983) 317–328
(a)–(c), R_r, uncountably many	Text (§ 1)	C. H. Taubes, “Gauge theory on asymptotically periodic 4-manifolds”, J. Differential Geom. 25 (1987) 363–430
§ 08	Text (§ 2, § 5)	A. Jaffe, E. Witten, Quantum Yang–Mills Theory, Clay Mathematics Institute
§ 08	Abstract	E. Witten, Math. Res. Lett. 1 (1994) 769–796, arXiv:hep-th/9411102
Original papers (via the texts above)	Bibliographic data only	Freedman, JDG 17 (1982) 357 / Taubes, JDG 17 (1982) 139 / Uhlenbeck, CMP 83 (1982) 11, 31 / Atiyah–Hitchin–Singer, Proc. R. Soc. A 362 (1978) 425 / Donaldson, JDG 26 (1987) 397 / Milnor, Ann. of Math. 64 (1956) 399
Milnor's (1956) example, the question of S^4, the version without the simply connected assumption (1987)	Secondary sources	Wikipedia “Exotic sphere”“Exotic R4”“Donaldson's theorem”
The Whitney trick, the construction of Casson handles, representing classes by surfaces, proofs of duality and of cobordism invariance, the upper bound on the cardinality of classes of manifolds, the analogy of § 08	Not checked (common account)	Text not checked
Properties of (3), Table 1, Table 2, Figure 1, Figure 2	Computed	Python 3.12, NumPy 1.26, SciPy 1.11 (lattices.py, gompf_check.py, moduli_disc.py). The formulas of the figures are checked with node

Next chapter to read: 3-07 Yang–Mills inside the Standard Model. For the physics of Seiberg–Witten, 17-11; for the "four" that matters in the problem, 0-04.

Yang–Mills inside the Standard Model — a product of three groups, with matter and the Higgs added

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

What this chapter shows — The gauge part of the Standard Model is the sum of three Yang–Mills fields, for $SU(3)\times SU(2)\times U(1)$. The charges satisfy an arithmetic called anomaly cancellation, and the value of the Higgs field gives mass to the W and Z while leaving the photon massless. We also sort out why the problem statement takes out pure Yang–Mills alone.

Prerequisite chapters — 3-02 (also using 3-01 and 1-04). Spin-1/2 fields are defined minimally in § 02.

Said first — This is a classical chapter. § 03 checks only the arithmetic of conditions from quantum theory, and the numbers in § 06 are estimates at the lowest order of perturbation theory (tree level). This chapter shows nothing about the mass gap.

Order of this chapter

1. Three groups — twelve gauge fields
2. Matter fields — one generation and the charges (Table 1)
3. Counting gives zero — anomalies (Table 2). Detour: SU(5)
4. The SU(3) of colour — quarks and colour singlets
5. The Higgs mechanism — an abelian model (Figure 1)
6. The electroweak $SU(2)\times U(1)$ — W, Z and the photon (Figure 2, Table 3)
7. Where mass comes from — the proton and the pion
8. The difference from pure Yang–Mills (Table 4)

01

Three groups — twelve gauge fields

The conventions are the same as in 3-02 § 02 ($\hbar = c = 1$, the metric $(-,+,+,+)$, Hermitian generators, $D_\mu = \partial_\mu - igA_\mu$).

The Standard Model is a gauge theory with three groups side by side, and the PDG review defines it as "the renormalizable field theory with gauge group $G_{\rm SM} = \mathrm{SU}(3)_C \times \mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$" Known ($C$ is colour, L is left-handed, Y is hypercharge). Each factor carries a Yang–Mills field, and the number of fields is $8 + 3 + 1 = 12$: the gluons G^a_μ, W^i_μ and B_μ.

$$\mathcal{L}_{\rm gauge} = -\tfrac{1}{4} G^a_{\mu\nu} G^{a\,\mu\nu} - \tfrac{1}{4} W^i_{\mu\nu} W^{i\,\mu\nu} - \tfrac{1}{4} B_{\mu\nu} B^{\mu\nu} \quad (1)$$

The field strengths of G and W have commutator terms, in the form of (7) of 3-02. $B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$, like the electromagnetic field, has no self-interaction. The action on matter is collected into one covariant derivative.

$$D_\mu = \partial_\mu - i g_s\, G^a_\mu T^a - i g\, W^i_\mu \frac{\sigma^i}{2} - i g'\, Y\, B_\mu \quad (2)$$

T^a acts only on colour triplets and $\sigma^i/2$ (1-04) only on doublets; on singlets the term is dropped. Since the group is not simple, there is one coupling constant for each factor, three in all (1-03). Example. $g_s \approx 1.22$ is obtained from the strength of the strong coupling $\alpha_s = g_s^2/4\pi$ measured at the scale of the Z mass, $\alpha_s(m_Z) = 0.1180$ (the MS convention; the note to Table 3 in § 06). $g \approx 0.65$ and $g' \approx 0.35$ are tree-level estimates worked back from the W and Z masses in Table 3 of § 06 Computed.

02

Matter fields — one generation and $Q = T_3 + Y$

Quarks and leptons are spin-1/2 fields. We do not use what is inside the components (the Dirac matrices), only how they turn under the groups. What is needed is that such a field splits into a **left-handed** part ψ_L and a **right-handed** part ψ_R (for zero mass, the way the spin turns relative to the direction of motion), and that SU(2) acts only on the left-handed part Known. Spatial reflection $\mathbf{x} \mapsto -\mathbf{x}$ (parity) exchanges left-handed and right-handed, so a force acting only on the left-handed part is not invariant under this reflection. This is called **parity violation**.

Table 1 The fields of one generation ($(\mathbf{3},\mathbf{2})$ is a colour triplet and a doublet; for T_3, the upper component of the doublet is $+\frac{1}{2}$ and the lower is $-\frac{1}{2}$)

Field	Representation	Y	$Q = T_3 + Y$	Components

Field	Representation	Y	$Q = T_3 + Y$	Components
$Q_L = (u_L, d_L)$	$(\mathbf{3}, \mathbf{2})$	1/6	2/3, −1/3	6
u_R	$(\mathbf{3}, \mathbf{1})$	2/3	2/3	3
d_R	$(\mathbf{3}, \mathbf{1})$	−1/3	−1/3	3
$L_L = (\nu_L, e_L)$	$(\mathbf{1}, \mathbf{2})$	−1/2	0, −1	2
e_R	$(\mathbf{1}, \mathbf{1})$	−1	−1	1
Higgs Φ	$(\mathbf{1}, \mathbf{2})$	1/2	1, 0	2 (complex)

Computed Q was computed as exact fractions. Known (93.1) of the PDG review of grand unification writes the right-handed fields as their conjugates u^c, d^c, e^c (§03; colour becomes $\bar{\mathbf{3}}$ and the sign of Y flips), and attaches Y as a subscript, normalised to twice its value. For example, the u_R of Table 1 ($(\mathbf{3}, \mathbf{1})$, $Y = 2/3$) is the $(\bar{\mathbf{3}}, \mathbf{1})_{-4/3}$ of u^c. The Higgs is $(\mathbf{1}, \mathbf{2})_1$ in the PDG and corresponds to the $Y = 1/2$ of Table 1. Converted, they agree with Table 1.

$$u_L:\ \tfrac{1}{2} + \tfrac{1}{6} = \tfrac{2}{3}, \qquad e_L:\ -\tfrac{1}{2} - \tfrac{1}{2} = -1, \qquad \nu_L:\ \tfrac{1}{2} - \tfrac{1}{2} = 0 \ \text{(3)}$$

One generation has 15 components counting colour, and there are three generations (PDG) Known . The gauge principle does not decide which field goes into which representation (3-01), and the PDG too states that the “apparently rather complex group-theoretic data” of Table 1 are not explained.

03

Counting gives zero — anomaly cancellation

First we prepare one word. Taking the complex conjugate of each component of a representation $\psi \mapsto \Omega\psi$ gives $\psi^* \mapsto \Omega^*\psi^*$, which is again a representation; we call it the **conjugate representation**. The conjugate of the triplet is written $\bar{\mathbf{3}}$ (the antitriplet), and in SU(3) it is not equivalent to $\mathbf{3}$. In U(1), $e^{iY\alpha}$ becomes $e^{-iY\alpha}$, so the sign of Y flips Known . For the SU(2) doublet $\sigma^2\Omega^*\sigma^2 = \Omega$ (σ^2 is the second Pauli matrix, not a square), so the conjugate is a representation equivalent to the original and is not distinguished Computed .

The Y of Table 1 is strongly constrained. Even when the classical action is gauge invariant, invariance under transformations that turn left and right separately can break at the stage of quantisation; this is called a **gauge anomaly**. The abstract of Gross–Jackiw (1972) states that it “may destroy this gauge invariance in perturbation theory” Known . We do not treat the derivation, and use only the fact that the condition becomes arithmetic. Taking the conjugate of a spin-1/2 field turns right-handed into left-handed, so aligning the right-handed fields to left-handed by conjugation (the sign of Y flips, and u^c becomes $-2/3$), the condition is that all the following sums are 0. The brackets on the right show which combination of groups each anomaly condition belongs to.

$$\begin{aligned} \textstyle\sum_{\text{colour triplets and antitriplets}} Y &= 0 \quad (\mathrm{SU}(3)^2\,\mathrm{U}(1)) \\ \textstyle\sum_{\text{doublets}} Y &= 0 \quad (\mathrm{SU}(2)^2\,\mathrm{U}(1)) \\ \textstyle\sum Y^3 &= 0 \quad (\mathrm{U}(1)^3) \\ \textstyle\sum Y &= 0 \quad (\text{gravity}^2\,\mathrm{U}(1)) \\ \#\mathbf{3} - \#\bar{\mathbf{3}} &= 0 \quad (\mathrm{SU}(3)^3) \end{aligned} \tag{4}$$

The sums are weighted by the dimension of the representations of the groups not in the combination (in the sum over doublets Q_L is multiplied by the number of colours, 3; in the sum over triplets and antitriplets Q_L is multiplied by 2). $\sum Y^3$ and $\sum Y$ run over the 15 components counting colour and doublet components, and the latter is the condition for the combination with gravity, $\text{gravity}^2\,\mathrm{U}(1)$. The abstract of Bouchiat, Iliopoulos and Meyer (1972) states that in models with leptons and quarks "the troublesome anomalies cancel" Known . Example. Fixing the charges and changing only the number of colours N_c gives Table 2.

Table 2 The sums of (4) for one generation (charges fixed, only N_c changed)

N_c	$\mathrm{SU}(N_c)^2\mathrm{U}(1)$	$\mathrm{SU}(2)^2\mathrm{U}(1)$	$\mathrm{U}(1)^3$	$\text{gravity}^2\,\mathrm{U}(1)$
1	0	−1/3	1/2	0
2	0	−1/6	1/4	0
3	**0**	**0**	**0**	**0**

Computed Added as exact fractions. For general N_c, the $\mathrm{SU}(2)^2\mathrm{U}(1)$ column is $N_c/6 - 1/2$ and the $\mathrm{U}(1)^3$ column is $(3 - N_c)/4$. The $\mathrm{SU}(3)^3$ condition has no column in the table; it is $\#\mathbf{3} - \#\bar{\mathbf{3}} = 2 - 2 = 0$ for $N_c = 3$ (the two Q_L and the two u^c, d^c). Known SU(2) has another condition too: the abstract of Witten (1982) states that an SU(2) gauge theory with an odd number of left-handed doublets (and no other representations) is mathematically inconsistent. One generation has $3 + 1 = 4$ doublets, an even number.

For $\mathrm{U}(1)^3$, the quarks contribute $-3/4$ and the leptons $+3/4$; they cancel only when a whole generation is present, and only when there are 3 colours Computed . In pure Yang–Mills, with no spin-1/2 fields, this condition does not arise.

Detour: it fits into SU(5)

A subsection you may skip.

Georgi–Glashow (1974) "conjectured" that the three forces arise from a single interaction of SU(5) (one of the groups of the problem statement, 0-03). Placing SU(3) at the top left of 5×5 and SU(2) at the bottom right leaves a traceless $Y = \mathrm{diag}(-\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}, \frac{1}{2}, \frac{1}{2})$ that commutes with both. One generation splits exactly into $\bar{\mathbf{5}}$ (d^c, L) and $\mathbf{10}$ (Q, u^c, e^c). The $T\mathbf{3}$ of $Q = T\mathbf{3} + Y$ is also a traceless element of SU(5), so the trace of Q is 0 too. Written with $\bar{\mathbf{5}}$ it is $3Q(d^c) + Q(\nu) + Q(e) = 0$, and from $Q(\nu) = 0$ and $Q(e) = -1$ follows $Q(d^c) = \frac{1}{3}$, that is, $|Q(d)| = |Q(e)|/3$. That the

charge of the down quark is exactly a third of the electron's follows from the shape of the group Computed . Grand unification is a hypothesis.

04

The SU(3) of colour — quarks and colour singlets

The SU(3) part is quantum chromodynamics (QCD). The PDG calls QCD “the SU(3) component of the SU(3)×SU(2)×U(1) Standard Model”, and writes that quarks come in three colours in the fundamental representation and gluons are of 8 kinds in the adjoint representation Known .

$$\mathcal{L}_{\mathrm{QCD}} = \sum_q \bar{\psi}_q \left(i\gamma^\mu D_\mu - m_q \right) \psi_q - \tfrac{1}{4} G^a_{\mu\nu} G^{a\,\mu\nu}, \qquad D_\mu = \partial_\mu - i g_s G^a_\mu T^a \tag{5}$$

The sum runs over the 6 kinds of quark (u, d, s, c, b, t), γ^μ is a 4×4 matrix acting on spin, and $\bar{\psi} = \psi^\dagger \gamma^0$. It is the form of (9.1) of the PDG with the sign of the covariant derivative matched to 3-02, and the last term is pure Yang–Mills. The γ^μ term is left in the PDG's convention (metric $(+,-,-,-)$), and its content is not used in this chapter. The mass term $m_q \bar{\psi}\psi$ multiplies left and right; colour turns left and right the same way, so it can be written, but SU(2) turns only the left-handed part, so $\bar{u}_L u_R$ is not invariant, and in the Standard Model mass enters through the Higgs field (§ 07) Known .
Colour singlets can be made from two properties of SU(3).

$$\bar{\psi}_a \psi^a \quad (\Omega^\dagger \Omega = 1), \qquad \varepsilon_{abc}\, \psi^a \psi^b \psi^c \quad (\det \Omega = 1) \tag{6}$$

The former is the form of mesons, the latter of baryons (proton, neutron). Example. Under $\Omega = \mathrm{diag}(e^{i\alpha}, e^{i\beta}, e^{-i(\alpha+\beta)})$ the phase of $\psi^1 \psi^2 \psi^3$ does not change. The elements of the centre of 1-04, $\mathbb{Z}_3$, also give 1 when taken three times. Jaffe–Witten state confinement as the physical particle states being “SU(3)-invariant” Known (for the mechanism, 17-02).

05

The Higgs mechanism — mass grows in an abelian model

A mass term breaks gauge invariance (3-02 § 08), yet the W and Z are heavy particles. What bridged this gap was the idea of placing the minimum of the potential of a charged scalar field away from $\phi = 0$. We look at it in a U(1) model.

$$\mathcal{L} = -\tfrac{1}{4} F_{\mu\nu} F^{\mu\nu} - (D_\mu \phi)^* (D^\mu \phi) - \mu^2 |\phi|^2 - \lambda |\phi|^4, \qquad D_\mu = \partial_\mu - i q A_\mu \tag{7}$$

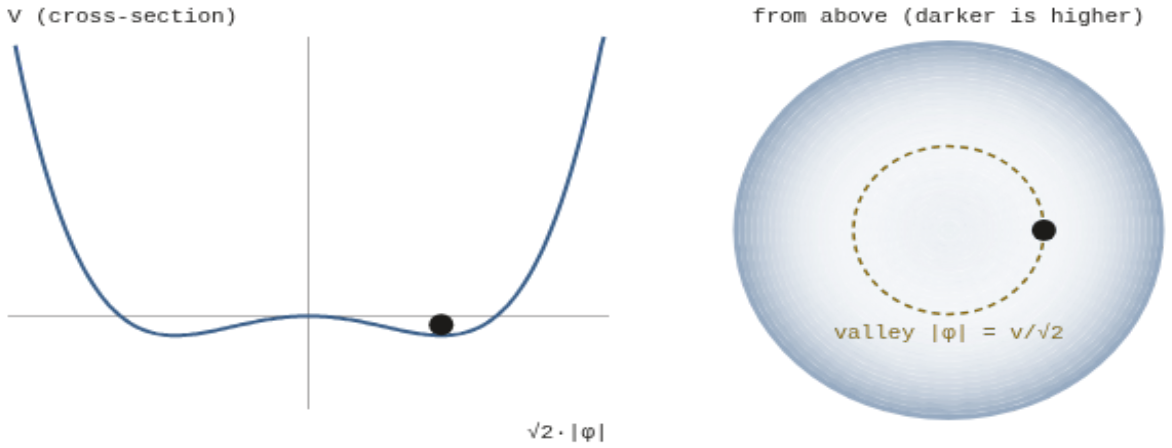
Let $\lambda > 0$. If $\mu^2 < 0$, the minima line up on the circle $|\phi| = v/\sqrt{2}$ ($v = \sqrt{-\mu^2/\lambda}$), and the gauge transformation $\phi \mapsto e^{i\alpha(x)}\phi$ moves along this circle. Turning the phase at each point and writing $\phi = (v + h)/\sqrt{2}$ (real),

$$-(D_\mu\phi)^*(D^\mu\phi) = -\tfrac{1}{2}\,\partial_\mu h\,\partial^\mu h - \tfrac{1}{2}\,q^2(v+h)^2 A_\mu A^\mu \quad (8)$$

follows. The $h = 0$ part is exactly the mass term forbidden in (14) of 3-02, and the masses are as follows Known . The action stays invariant, and the solution (the position of the minimum) brings in the mass.

$$m_A = q\,v, \qquad m_h = \sqrt{2\lambda}\,v = \sqrt{-2\mu^2} \quad (9)$$

Figure 1 Changing μ^2 with the knob changes the potential (left: cross-section; right: from above) and v, m_A, m_h. "Play" shows motion along the valley (the direction removable by a gauge transformation) and the radial oscillation h (on devices set to reduce motion it advances one frame at a time). $\lambda = 1$, $q = 0.6$.



The five points of Figure 1 ($\lambda = 1$, $q = 0.6$)

μ^2	v	$m_A = qv$	Scalar mass	Minimum of V
+0.5	0	0	0.707, 0.707	0
0	0	0	0, 0	0
−0.25	0.500	0.300	$m_h = 0.707$	−0.0156
−0.5	0.707	0.424	$m_h = 1.000$	−0.0625
−1	1.000	0.600	$m_h = 1.414$	−0.2500

Computed The numbers in this figure and table were computed on this machine (by (9)). In the $\mu^2 \geq 0$ rows the scalar mass is that of the two real components of ϕ (both $\sqrt{\mu^2}$); in the $\mu^2 < 0$ rows it is that of the radial h.

Degrees of freedom and the Goldstone boson

Example. We count the degrees of freedom. A massless vector field has 2 polarisations, because of its four components, one is removed by gauge transformations and one by the constraint (2-03 § 04). With a mass term the one removable by gauge

transformations is gone and only the constraint remains, so $4-1=3$ Known . Then before, there are the massless A (2) and ϕ (2 real components); after, the massive A (3) and h (1); both make 4. The phase that disappeared has become the longitudinal oscillation of A.

A massless excitation that appears when a continuous symmetry is spontaneously broken is called a **Goldstone boson** Known . The motion along the valley in Figure 1 corresponds to it, and when there is a gauge field, this direction can be removed by a gauge transformation. Higgs (1964) wrote "the longitudinal degrees of freedom of these particles (which would be absent if their mass were zero) go over into the Goldstone bosons when the coupling tends to zero", and called it the "relativistic analog" of Anderson's (1963) remark on the plasmon in superconductors. In the same year Englert–Brout and Guralnik–Hagen–Kibble also published papers Known . Higgs also cautioned that "nothing is proved about the quantized theory".

"Spontaneous symmetry breaking" is a way of speaking in a gauge-fixed description. Elitzur (1975) argued that in lattice gauge theory without gauge fixing, local symmetries are not spontaneously broken Known (3-01 § 07).

06

The electroweak SU(2)×U(1) — W, Z and the photon

In the Standard Model the same thing is done with SU(2)×U(1) and the Higgs doublet Φ of Table 1. What leaves one of the minima of $V=\lambda(\Phi^\dagger\Phi-v^2/2)^2$, $\Phi_0=(0,\,v/\sqrt{2})$, unchanged is the U(1) of the charge Q, which is $T_3+Y=-\tfrac12+\tfrac12=0$ on the lower component. In Jaffe–Witten's words, the structure group is reduced "to a U(1) subgroup (diagonally embedded in SU(2) × U(1))" Known . Letting (2) act on Φ_0,

$$-(D_\mu\Phi_0)^\dagger(D^\mu\Phi_0)=-\frac{v^2}{8}\Big[g^2\big(W^1_\mu W^{1\mu}+W^2_\mu W^{2\mu}\big)+\big(gW^3_\mu-g'B_\mu\big)\big(gW^{3\mu}-g'B^\mu\big)\Big]$$ (1

follows. W^3 and B appear only in the combination $gW^3-g'B$, so we rotate the plane (W^3,B) by the **weak mixing angle** θ_W ($\tan\theta_W=g'/g$).

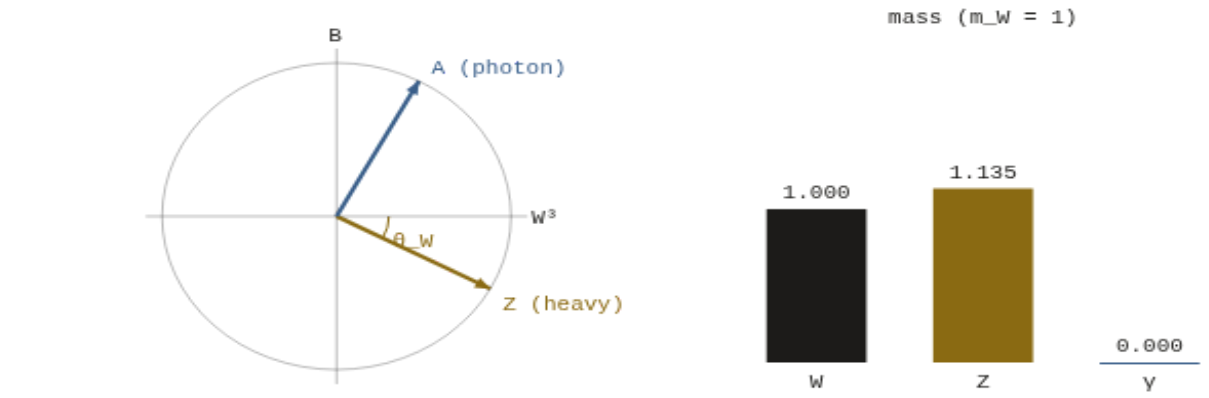
$$Z_\mu=\cos\theta_W\,W^3_\mu-\sin\theta_W\,B_\mu,\qquad A_\mu=\sin\theta_W\,W^3_\mu+\cos\theta_W\,B_\mu$$ (11)

$$m_W=\frac{gv}{2},\qquad m_Z=\frac{m_W}{\cos\theta_W},\qquad m_\gamma=0,\qquad e=g\sin\theta_W$$ (12)

$W^\pm=(W^1\mp iW^2)/\sqrt{2}$. This is the same as (10.3)(10.5) of the PDG, and the PDG cautions that these are masses "at tree level, i.e., to lowest order in perturbation theory" Known . Computed Building the mass matrix of 4×4 from Pauli matrices, the

eigenvalues agreed with (12) for three choices of (g, g', v) (relative differences below 10^{-15}, of the order of rounding error).

Figure 2 Changing $g'/g = \tan\theta_W$ with the knob rotates the directions of Z and the photon A in the (W^3, B) plane and changes $m_Z/m_W = 1/\cos\theta_W$. "Measured" returns it to 0.536.



g'/g = 0.536 θ_W = 28.19° sin²θ_W = 0.2232 m_Z/m_W = 1.1346 e/g = 0.4724

The five points of Figure 2

g'/g	θ_W	$\sin^2\theta_W$	m_Z/m_W	e/g
0	0.00°	0.0000	1.0000	0.0000
0.25	14.04°	0.0588	1.0308	0.2425
0.536	28.19°	0.2232	1.1346	0.4724
1	45.00°	0.5000	1.4142	0.7071
1.5	56.31°	0.6923	1.8028	0.8321

Computed The numbers in this figure and table were computed on this machine (by (12)). For $g'/g = 0$ the photon is B itself. 0.536 is the value from the measured values of Table 3.

Example. Working (12) back from the PDG 2024 values gives Table 3.

Table 3 Tree-level estimates (input: $G_F = 1.1663788 \times 10^{-5}\ \text{GeV}^{-2}$, $m_W = 80.3692$, $m_Z = 91.1880$, $m_H = 125.20$ GeV)

Quantity	Formula	Value
v	$(\sqrt{2}G_F)^{-1/2}$	246.22 GeV
$g,\ g'$	$2m_W/v,\ g\tan\theta_W$	0.6528, 0.3499
$\sin^2\theta_W$	$1 - m_W^2/m_Z^2$	0.22321
$1/\alpha$	$4\pi/e^2$	132.1
λ	$m_H^2/(2v^2)$	0.1293

Computed In the PDG, the on-shell $\sin^2\theta_W = 0.22348$ and $1/\alpha$ is 137.036 at low energy, and 127.930 at m_Z (MS). On-shell and MS are names for ways of including quantum corrections (renormalisation conventions), and quantities with the same name differ slightly in value (4-04). The λ of Table 3 is the value in the $V = \lambda(\Phi^\dagger\Phi - v^2/2)^2$ convention of the text. The PDG's λ becomes $m_H/v = 0.5085$ with the normalisation $m_H = \lambda v$ ((10.5a)), and the two are related by $\lambda = \lambda_{\text{PDG}}^2/2$. m_W is the average excluding CDF (2022).

The 132.1 for $1/\alpha$ is 3.6 % smaller than the low-energy 137.036. The v of Table 3 was taken from G_F (a constant fixed by the muon lifetime), so the quantum correction Δr linking G_F and α is not included. Squaring (10.22b) of the PDG electroweak review gives $m_W^2 \sin^2\theta_W\,(1-\Delta r) = \pi\alpha/(\sqrt{2}\,G_F)$ ($\sin^2\theta_W$ is on-shell), and the PDG gives $\Delta r = 0.03685$ Known . The tree-level $1/\alpha$ corresponds to $(1-\Delta r)/\alpha$, and $137.036 \times (1 - 0.03685) = 131.99$ agrees with the 132.1 of Table 3 to 0.09 % Computed (quantum corrections in 4-04). The 127.930 at MS is a quantity with a different convention, listed only for reference. The degrees of freedom are 12 both before and after (from the 8 of W^i, B and the 4 of the doublet, to the 9 of the heavy $W^\pm$, Z, the 2 of the photon and the 1 of H) Computed .

Weinberg (1967) wrote that the masses of the W and Z come “not from a mass term put in at the beginning”, and in 2012 ATLAS and CMS reported the observation of the Higgs particle Known .

07

Where mass comes from — the proton and the pion

What v gives directly is the masses of the W and Z, and the masses of quarks and leptons through the **Yukawa couplings** Known . A Yukawa coupling is a term $\bar\psi_L \Phi \psi_R$ that is a product of the Higgs field and spin-1/2 fields; for up-type quarks $\tilde\Phi = i\sigma^2\Phi^*$ is used (a doublet of $Y = -\frac{1}{2}$; the conjugate of § 03). In every term the sum of Y is 0 (for example $\bar Q_L \Phi d_R$ is $-\frac{1}{6} + \frac{1}{2} - \frac{1}{3} = 0$), so it is gauge invariant Computed . Setting Φ to Φ_0 gives a mass term proportional to $\bar\psi_L \psi_R$. But this hardly explains the mass of the proton. The sum of the masses of the core u, u, d is about 1% of the proton's.

$$\frac{2m_u + m_d}{m_p} = \frac{2 \times 2.16 + 4.70}{938.272} \approx 0.0096 \quad (13)$$

Computed Values from the PDG 2024 Summary Tables (m_u, m_d is MS, 2 GeV). Quark masses depend on the definition, so (13) is a guide to orders of magnitude. The rest is thought to come from the dynamics of QCD, whose scale is the Λ produced by quantisation (4-07) Physics .

The pion is light, about 15% of the proton ($m_{\pi^\pm} = 139.57$ MeV) Computed . According to the PDG, in the limit where the masses of the three light quarks are set to 0, the flavour symmetry that turns left and right separately (chiral symmetry) breaks, and “eight massless Goldstone bosons, the π’s, K’s, and η” appear Physics . That QCD is thought to have no gap. This is the “chiral symmetry breaking” that Jaffe–Witten list as the third property of QCD Known . Pure Yang–Mills has no quarks, and its lightest states are thought to be glueballs (5-08) Physics . The gap Δ of the problem statement is a scale

that should arise from the Yang–Mills field alone, relying on neither quarks nor the Higgs field (4-11).

08

The difference from pure Yang–Mills — what the problem statement leaves out

The problem statement concerns pure Yang–Mills theory for any compact simple group (0-03). Table 4 shows what to remove from the Standard Model to get there.

Table 4 The Standard Model, QCD and pure Yang–Mills

Item	Standard Model	QCD	Pure Yang–Mills
Group	$SU(3)\times SU(2)\times U(1)$	$SU(3)$	Compact simple group
Matter fields	Quarks, leptons, Higgs	6 kinds of quark	None
Parameters	19 (as the PDG counts)	g_s, 6 m_q, θ	g, θ (3-05)
Classical mass scale	μ^2 (v) and the Yukawa couplings	m_q	None (3-02 § 07)
Massless particles	Photon	π and others in the limit of massless light quarks Physics	Expected to be none Physics
Treatment in the problem statement	Outside the scope	Extension problem	The main problem

There are three differences. **The group**: it is not simple. The abelian $U(1)_Y$ part is not asymptotically free (the property that the coupling weakens at short distances) (Coleman–Gross, 4-06 § 04) Known , and at one loop the coupling grows at short distances Physics . **Massless particles**: the photon remains, so the Standard Model as a whole has no gap, and a gap can be asked about only for the strong-force part. **Where mass comes from**: in the Higgs mechanism the scale μ^2 is in the classical action and visible from the solution, but the classical action of pure Yang–Mills has no scale, and the gap is thought to arise from quantisation Physics . Jaffe–Witten too, in § 2, give the Higgs mechanism as an example of "classical properties", and the mass gap and confinement as examples of "quantum properties that do not have classical analogs" Known .

What was left out remains in extension problems. Jaffe–Witten list "other four-dimensional gauge theories (incorporating additional fields that preserve asymptotic freedom)" (0-03), and at the time of writing the problem description they call the two-dimensional abelian Higgs model of Brydges–Fröhlich–Seiler "the only complete example of an interacting gauge theory satisfying the axioms" Known . A special limit of the four-dimensional Higgs model is in 7-10, and a picture of where the groups come from is in 17-07.

What this chapter can say

Matter	What is said	What is not said
(1)–(6), Table 1, Table 2	The standard definitions, the arithmetic of charges and sums Known Computed	Why the representations are these, the derivation of the anomaly conditions, confinement
(7)–(12), figures, Table 3	How mass appears in classical fields, tree-level estimates, comparison of $1/\alpha$ using the PDG's Δr Known Computed	A theorem on the Higgs mechanism in quantum theory, values including quantum corrections
Table 4, the SU(5) detour	The PDG's counting, and the arithmetic of assigning representations Known Computed	That grand unification is right, the origin of the 19
(13), § 07	Computation of ratios and the PDG's account Physics	Proofs of these from QCD
The mass gap	—	That pure Yang–Mills has a gap (open; this site shows nothing) Physics

Sources and reproduction

Item	Kind	Source / tools
Definition of the Standard Model, representations, SU(5), the 19, electroweak and QCD formulas and values, quark masses, Goldstone bosons	Reviews and Summary Tables (text checked)	S. Navas et al. (Particle Data Group), Phys. Rev. D 110 (2024) 030001: § 9 QCD, § 10 Electroweak Model, § 60 Quark Masses, § 93 Grand Unified Theories, Summary Tables
Quotations of Jaffe–Witten	Primary source (text checked)	A. Jaffe, E. Witten, “Quantum Yang–Mills Theory”, the official problem description of the Clay Mathematics Institute, § 1, § 2, § 5, § 6
Quotations of Higgs and Weinberg	Primary source (text checked)	P. W. Higgs, Phys. Rev. Lett. 13 (1964) 508 / S. Weinberg, Phys. Rev. Lett. 19 (1967) 1264
Other quotations and summaries	Primary source (abstract checked)	Englert, Brout, PRL 13 (1964) 321 / Bouchiat, Iliopoulos, Meyer, PLB 38 (1972) 519 / Gross, Jackiw, PRD 6 (1972) 477 / Georgi, Glashow, PRL 32 (1974) 438 / Elitzur, PRD 12 (1975) 3978 (same as 3-01) / Witten, “An SU(2) anomaly”, PLB 117 (1982) 324 / ATLAS, PLB 716 (2012) 1 / CMS, PLB 716 (2012) 30
Anderson	Primary source (abstract checked; the wording in the text follows the text of Higgs 1964)	P. W. Anderson, Phys. Rev. 130 (1963) 439

Item	Kind	Source / tools
Brydges–Fröhlich–Seiler	Bibliographic data only (via reference [7] of Jaffe–Witten § 6.4)	“On the construction of quantized gauge fields” I: Ann. Phys. 121 (1979) 227 / II: CMP 71 (1980) 159 / III: CMP 79 (1981) 353
Guralnik–Hagen–Kibble	Bibliographic data only	PRL 13 (1964) 585
Tables 1–3, values in the figures and tables, eigenvalues of the mass matrix, (13), the comparison of Δr, checks of conjugation and Yukawa terms	Computed on this machine	Python 3.12, NumPy 1.26, fractions (`sm_tables.py`, `sm_checks.py`). The figures do (9)(12) in plain JavaScript

Next chapter to read: 4-01 Canonical quantisation and the Hamiltonian.