Chern–Simons Theory in a Knotshell

David Grabovsky

September 14, 2022

Abstract

In this pedagogical review, we introduce and study a three-dimensional topological field theory called Chern–Simons theory. We begin with the phenomenology of the U(1) action: we discuss its classical equations of motion, its quantization, and its observables. We then pass to the nonabelian theory, which we treat in the path integral and canonical formalisms. We pay special attention to the framing of manifolds and knots, as well as to the geometrical structure of the quantum theory.

Contents

1 Introduction and Motivation

Introduction. Chern–Simons theory is an exercise in the simplicity, beauty, and weirdness of topology. It is an archetypical example of a topological field theory, a quantum field theory where the physical observables are topological invariants of the spacetime in which the theory lives. In particular, Witten showed in the late 1980s that nonlocal observables in Chern–Simons (CS) theory called Wilson loops, represented by knots in a 3-dimensional spacetime, compute certain invariants of those knots that generalize the celebrated Jones polynomial [1]. As part of this work, which earned him the Fields medal, Witten described how to use path integrals to give a physical interpretation to such invariants. He also quantized and solved the theory, and gave an account of its Hilbert space structure. His work highlighted deep connections between diverse areas of mathematics and physics, and set off a flurry of activity that greatly enriched and brought together both fields.

Motivation. CS theory appears in 2D conformal field theory, 3D quantum gravity, 4D gauge theory, condensed matter physics, geometry, topology, and more. The moral here is that the present review should have something useful in it for everyone.

- *In Yang—Mills theory,* the theta term responsible for solitons, instantons, monopoles, and anomalies is accompanied by a topological charge. The quantization of this charge is analogous to the quantization of CS theory, and many properties of fermions in four dimensions can be seen as echoes of CS physics in one dimension lower.

- *In 2-dimensional conformal field theory (CFT),* several surprising connections to CS theory manifest themselves in subtle ways. The behavior of fermions, the physics of Wess–Zumino–Witten models, representations of affine Kac–Moody algebras, and rational CFTs all admit interpretations through the 3-dimensional lens of CS theory.

- *In condensed matter theory,* topological insulators and the fractional quantum Hall effect are governed by low-energy effective actions with a CS term. Coupling matter to CS terms produces a host of novel topological phenomena and phases. In addition, the physics of Landau levels see application in the formal solution of abelian CS theory.

- *In differential geometry and topology,* one studies the curvature of various principal bundles and their (principal) connections using tools like characteristic classes and Chern–Weil theory. The CS form, its gauge-theoretic properties, and the invariants it begets undergird the subject and, in turn, inform the study of CS theory itself.

- *In knot theory,* Witten's work directly gave a physical and intrinsically 3-dimensional realization—and generalization—of many topological invariants, among them the Reidmeister and Ray–Singer torsions, the Jones polynomial, and Khovanov homology.
- *In quantum gravity,* Witten showed that 3-dimensional gravity could be recast as a CS theory with non-compact gauge group and solved exactly. Together with the advent of holographic duality, this development placed CS theory at the forefront of our modern understanding of low-dimensional quantum gravity.

Outline. In this review, we will describe some aspects of Witten's work. We will essentially develop CS theory twice; first in the abelian setting (§2) following Dunne [2], and then in general (§§3–4), following Witten's landmark paper [1]. The abelian case will serve as a warm-up: there we will discover key features of CS theory such as anyons, Wilson loops, linking numbers, its gauge behavior, and some subtleties of its quantization. This will form the core of our intuitive understanding of CS physics, and the formal development in the rest of the review will both rely on and flesh out these ideas. In §3, we will emphasize path integrals and the role played by the framing of knots. Careful consideration of this issue, which may initially seem like a pedantic nuisance, will allow us to explicitly evaluate the CS path integral at weak coupling. We will then describe the canonical quantization of the theory in §4, taking a holomorphic approach based on bundles over various moduli spaces. Here we will allow ourselves to be slightly more vague in order to get the main points across. We will largely avoid a discussion of knot polynomials, but we will briefly mention them in §5, together with a few other odds and ends that bring our discussion to a close.

Conventions. We begin in §2 by working in Lorentzian signature using the mostly minus convention. Starting with §3, however, however, we will switch to Euclidean signature, and will stay there for the remainder of the review. We use Greek indices $\mu, \nu, \rho \in \{0, 1, 2\}$ for spacetime and middle Latin indices $i, j, k \in \{1, 2\}$ for space. The Einstein summation convention is employed throughout, except when indicated otherwise. As a warning, the letter g will variously refer to the metric on a 3-manifold, an element of the gauge group, or the genus of a Riemann surface; its meaning should be inferred from the local context.

2 Abelian Chern–Simons Theory

The theory of Chern and Simons. Instead of diving headfirst into abstract formalism, we will begin with a simple but instructive example. Let the 3-dimensional Minkowski space $M = \mathbb{R}^{2,1}$ represent the universe. The abelian *Chern–Simons action* is constructed from the U(1) gauge field $A_\mu(x)$, which is analogous to the photon field in electrodynamics:

$$S_{\mathrm{CS}}[A] \equiv \frac{k}{4\pi} \int_M \mathrm{d}^3x \, \varepsilon^{\mu\nu\rho} A_\mu \partial_\nu A_\rho. \tag{2.1}$$

Here $\varepsilon^{\mu\nu\rho}$ is the totally antisymmetric symbol, $k \in \mathbb{R}$ is called the *Chern–Simons level*, and the factor of $\frac{1}{4\pi}$ is purely conventional. This action is a bit peculiar: it only makes sense in 3 dimensions, and is of first order in derivatives of A. To gain some intuition about its behavior, we will explore the classical and quantum consequences of the theory it defines.

We will follow Dunne's review [2], occasionally drawing on wisdom from Tong [3, 4].

2.1 The Classical Theory

The equations of motion. The equations of motion that follow from the action (2.1) are

$$\delta S_{\mathrm{CS}} = 0 \implies \frac{k}{4\pi}\varepsilon^{\mu\nu\rho}F_{\nu\rho} = 0 \implies F = 0, \qquad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{2.2}$$

This looks rather boring: there are no propagating local degrees of freedom, and the constant k does not affect the classical physics. We can remedy this situation by adding matter. For example, a Dirac fermion ψ couples to the CS term by means of a current:

$$\mathcal{L} = \frac{k}{4\pi}\varepsilon^{\mu\nu\rho}A_\mu\partial_\nu A_\rho + \overline{\psi}\big(i\partial\!\!\!/ - m\big)\psi - eA_\mu\overline{\psi}\gamma^\mu\psi = \mathcal{L}_{\mathrm{CS}} + \mathcal{L}_\psi + A_\mu J^\mu. \tag{2.3}$$

In terms of the matter current J, the resulting equations of motion for A are

$$\frac{k}{4\pi}\varepsilon^{\mu\nu\rho}F_{\nu\rho} = \frac{k}{2\pi}\varepsilon^{\mu\nu\rho}\partial_\nu A_\rho = J^\mu. \tag{2.4}$$

Geometrical aside. As we will discuss later, A is actually a connection on a principal U(1)-bundle over M, and the Lagrangian is a multiple of the *Chern–Simons 3-form* $A \wedge \mathrm{d}A$. The field strength $F = \mathrm{d}A$ is the curvature of this connection, so the equation of motion $F = 0$ describes flat connections. The presence of a matter current J modifies the equation of motion to $\frac{k}{4\pi} \star F = J$ (where $\star$ is the Hodge star), and the matter introduces curvature in the gauge bundle. Miraculously, this curvature can be measured using a topological invariant of the spacetime, constructed from the CS form. This fact is the beginning of Chern–Weil theory, which is (to a mathematician) the correct way to study the CS action [5].

Electric and magnetic fields. The field strength tensor $F_{\mu\nu}$ is populated by the electric and magnetic fields $\mathbf{E}$ and B, which are defined exactly as in the Maxwell theory:

$$E_i \equiv -\partial_i A_0 - \partial_0 A_i, \qquad B \equiv \varepsilon^{ij}\partial_i A_j. \tag{2.5}$$

If we write the matter current $J^\mu \equiv (\rho, \mathbf{J})$ in terms of the charge and current densities, it is straightforward to check that the CS equations of motion are given in components by

$$\rho = \frac{k}{2\pi}B, \qquad J^i = \frac{k}{2\pi}\varepsilon^{ij}E_j. \tag{2.6}$$

Evidently B is sourced by electric charges, while $\mathbf{E}$ is a consequence of current: this is the opposite of what happens in electromagnetism. The physical situation is depicted in Fig. 1 (left): moving charges generate in-plane $\mathbf{E}$ fields, while their B fields point in an imaginary "z" direction. (This picture is actually experimentally accurate in Hall physics.) Each *magnetic flux line* is attached to a source charge and pierces the spatial manifold.

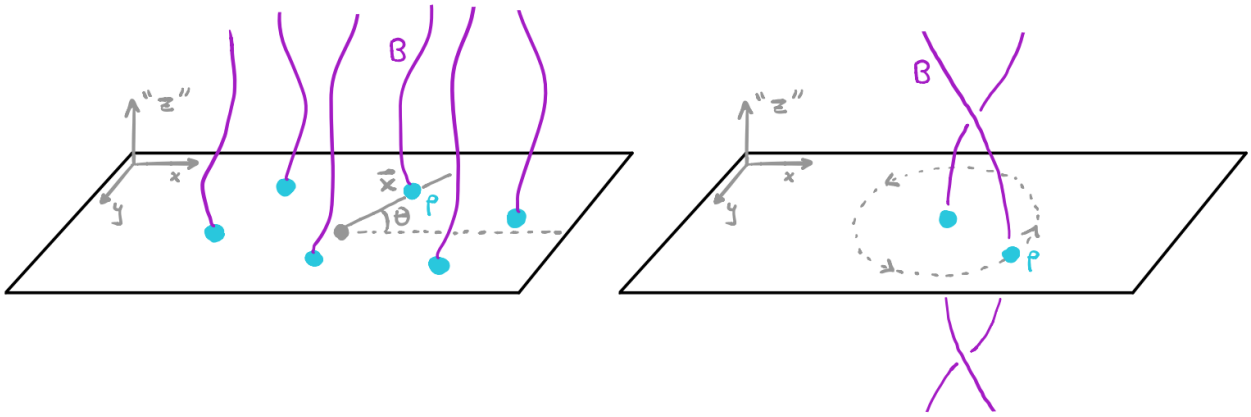

Figure 1: Left: the Chern-Simons universe. Right: an anyon!

A single test charge. To gain a more detailed understanding of the physics at work, let us solve the CS equations (2.6) with sources $\rho = \delta^{(2)}(\mathbf{x} - \mathbf{x}_a(t))$ and $\mathbf{J} = \mathbf{0}$. We will work in Coulomb gauge, where $A_0 = 0 = \partial_i A^i = \nabla \cdot \mathbf{A}$. We plug these sources into (2.6), write out the $\mathbf{E}$ and B fields in terms of $A_\mu = (0, A_i)$ as in (2.5), and use standard Green's function techniques (specifically, the identity $\nabla^2 \log|\mathbf{x}| = 2\pi\delta^{(2)}(\mathbf{x})$) to obtain the solution to (2.6):

$$A_i(\mathbf{x}, t) = \frac{1}{k}\varepsilon_{ij}\frac{x^j - x_a^j(t)}{|\mathbf{x} - \mathbf{x}_a|^2} = -\frac{1}{k}\partial_i\theta(\mathbf{x} - \mathbf{x}_a(t)), \qquad \theta(\mathbf{x}) \equiv \tan^{-1}\left(\frac{y}{x}\right) = \arg(\mathbf{x}). \tag{2.7}$$

Observe that A_i is a total derivative: it is a pure gauge configuration, and can be brought to zero by a gauge transformation that adds the total derivative of $\omega(x) \equiv \left(\frac{1}{k}\right)\theta(\mathbf{x} - \mathbf{x}_a(t))$. Hence $A(x) \equiv 0$, which makes $\mathbf{E} \equiv \mathbf{0}$ and $B = 0$ trivial. But the same gauge transformation also acts on the matter field ψ—which was responsible for the source charge—by giving it a nontrivial Aharanov-Bohm phase that depends on its angular position θ and on k:

$$\omega = \frac{1}{k}\theta \implies \begin{cases} A_i(x) \longrightarrow A'(x) = A_i(x) + \partial_i\omega(x) = 0, \\ \psi(x) \longrightarrow \psi'(x) = e^{i\omega(x)}\psi(x) = e^{i\theta/k}\psi(x). \end{cases} \tag{2.8}$$

Double exchange and anyons. Now consider two charges, as shown in Fig. 1 (right). One remains stationary, while the other moves around the first. We view this process as a double exchange: the moving charge trades places with the stationary one twice, once for

each π rotation. By (2.8), the phase picked up by the moving particle is given by

$$\Delta\theta = 2\pi \implies \psi(x) \longrightarrow \exp\left(\frac{2\pi i}{k}\right)\psi(x) = \exp\left(i \oint_{\theta=0}^{\theta=2\pi} \mathrm{d}x^i A_i\right)\psi(x) \neq \psi(x). \tag{2.9}$$

If these particles were bosons or fermions, ψ would return to itself under double exchange. But (2.9) shows that by tuning k, we can give ψ arbitrary statistics valued in U(1). So these particles are *anyons*, because they can have *any* statistics. (The quantity $e^{2\pi i/k}$ and exponentials of closed-loop integrals of A will return frequently throughout this review.)

Comments. The phase picked up by an anyon depends only on k and the topology of the path it takes. In our example, it winds once around a puncture poked in $\mathbb{R}^2$ by the stationary charge (which has infinite charge there). Thus CS theory comes equipped with a field-theoretic probe—the anyonic phase—that detects the *winding number* of a nontrivial loop in $\mathbb{R}^2 \setminus \{\mathbf{0}\} \simeq S^1 \times \mathbb{R}$. Another reflection of the same idea is found in the tangling of the magnetic flux lines in Fig. 1. As we will soon see, it is useful to think of these flux lines as loops—that is, *knots*—that close up at infinity. Aside from providing a picture of Gauß's law, this perspective reveals that CS theory is really computing the *linking number* of these B-field knots. The main point is that CS theory has no local degrees of freedom, yet it strongly affects far-separated particles and knows about infrared physics [3].

2.2 Path Integral Quantization

Gauge invariance? We have already seen the properties of CS theory under gauge transformations at work in our discussion of anyons. Now we turn to a simpler and more fundamental issue: we wish to show that the pure CS action (2.1) is *not* gauge invariant. This project gets off to a remarkably poor start: under a gauge transformation of the form $A_\mu \longrightarrow A'_\mu = A_\mu + \partial_\mu\omega$, a calculation shows that the action changes by a total derivative:

$$S_{\mathrm{CS}} \longrightarrow S'_{\mathrm{CS}} = S_{\mathrm{CS}} + \frac{k}{4\pi}\int_M \mathrm{d}^3x\, \varepsilon^{\mu\nu\rho}\partial_\mu(\omega\, \partial_\nu A_\rho). \tag{2.10}$$

If ω vanishes at infinity, then the boundary term is zero and S_{CS} is gauge invariant. Embarrassingly, even if ω does not vanish at infinity, the abelian CS action is *still* gauge invariant by Gauß's law, in the absence of magnetic monopoles.[1] We will therefore describe the simplest scenario in which the presence of magnetic flux ruins gauge invariance. In the end, a cleverly chosen gauge transformation will cause the action to transform as $S_{\mathrm{CS}} \longrightarrow S_{\mathrm{CS}} + 2\pi k$. As we will see, this implies that in the quantum theory, k must be an integer.

[1] As we will see, this is not the case for the nonabelian theory, which is more fragile: all "large" gauge transformations cause the action to transform nontrivially, with no need to add monopoles to the theory. The resulting transformation, $S_{\mathrm{CS}} \longrightarrow S_{\mathrm{CS}} + 2\pi k$, is the same as in the abelian theory with monopoles.

Euclidean acrobatics. The easiest way to break the gauge invariance of S_{CS} is to compactify the spatial manifold to S^2, and to pass to Euclidean signature via the Wick rotation $t \longrightarrow \tau = it$, whereby the time axis $\mathbb{R}$ becomes the thermal circle S^1. Crucially, the CS action is of first order in time derivatives and transforms to $S_{\mathrm{CS}} \longrightarrow S^{\mathrm{E}}_{\mathrm{CS}} = -iS_{\mathrm{CS}}$. Next, we choose the gauge transformation $\omega(\tau, \mathbf{x}) = \frac{2\pi\tau}{\beta}$, which winds once around the thermal circle and cannot be continuously deformed to the identity map. Finally, we introduce magnetic flux, i.e. a nonzero integral of $B = F_{12}$ over S^2. To do this, we imagine that $S^2 \subset \mathbb{R}^3$ encloses a monopole planted at the origin, and we measure the flux piercing its surface. Thinking fondly of Dirac, we recall that the compactness of U(1) quantizes such fluxes:

$$\int_{S^2} F_{12} = 2\pi n, \qquad n \in \mathbb{Z}. \tag{2.11}$$

At last, all of our ingredients come together: consider the gauge transformation (2.10) on $M = S^2 \times S^1$, and substitute the ω given above. An integration by parts reveals the boundary term to be proportional to the flux integral above; taking $n = 1$, we find $\delta S_{\mathrm{CS}} = 2\pi k$.

The partition function. We appear to have a big problem: S_{CS} is not gauge invariant! But the quantum theory can still be saved if its partition function $\mathcal{Z}$ is gauge invariant. Combining $\delta S_{\mathrm{CS}} = 2\pi k$ and $-S^{\mathrm{E}}_{\mathrm{CS}} = iS_{\mathrm{CS}}$, the path integral for $\mathcal{Z}$ transforms as follows:

$$\mathcal{Z} = \int \mathcal{D}A\, e^{-S^{\mathrm{E}}_{\mathrm{CS}}} \longrightarrow \int \mathcal{D}A\, e^{i(S_{\mathrm{CS}} + 2\pi k)} = \int \mathcal{D}A\, e^{2\pi i k} e^{-S^{\mathrm{E}}_{\mathrm{CS}}} = e^{2\pi i k}\mathcal{Z}. \tag{2.12}$$

Thanks to the fact that S_{CS}—like all topological terms—is imaginary in Euclidean signature, the path integral is gauge invariant whenever the CS level k is an integer.

Anyons as Wilson loops. Having used path integrals to quantize the CS level, we will now use them to rediscover anyons from a more formal perspective, following Polyakov [6]. We return to Lorentzian signature and consider the generating functional of pure CS theory, i.e. its partition function in the presence of a source J:

$$\mathcal{Z}[J] \equiv \frac{1}{\mathcal{Z}} \int \mathcal{D}A\, \exp\left[i \int_M \mathrm{d}^3x \left(\frac{k}{4\pi} \varepsilon^{\mu\nu\rho} A_\mu \partial_\nu A_\rho + A_\mu J^\mu \right) \right] \equiv \frac{1}{\mathcal{Z}} \int \mathcal{D}A\, e^{i\widetilde{S}_{\mathrm{CS}}[A,J]}. \tag{2.13}$$

To reproduce the physical situation in Fig. 1, consider two particles moving in closed loops γ_1 and γ_2 around each other in the plane. Model them by the source $J \equiv J_1 + J_2$, with

$$J^\mu_a \equiv \oint_{\gamma_a} \mathrm{d}x^\mu_a\, \delta^{(3)}(x - x_a(t)), \qquad a \in \{1, 2\}. \tag{2.14}$$

Thanks to these "point charge" delta functions, the source term in the action evaluates to

$$\int_M \mathrm{d}^3x\, A_\mu J^\mu = \oint_{\gamma_1} \mathrm{d}x_1^\mu\, A_\mu + \oint_{\gamma_2} \mathrm{d}x_2^\mu\, A_\mu. \tag{2.15}$$

We recognize the exponentials of these sources as nonlocal observables called *Wilson loops*, defined by $W_a \equiv \exp\Big[i \oint_{\gamma_a} \mathrm{d}x_a^\mu\, A_\mu\Big]$. (These are just the anyonic phass (2.9) in disguise!) In terms of W_1 and W_2, the path integral (2.13) becomes the expectation value of their product:

$$\begin{aligned}\mathcal{Z}[J] &= \frac{1}{\mathcal{Z}} \int \mathcal{D}A\, \exp\left[i \oint_{\gamma_1} \mathrm{d}x_1^\mu\, A_\mu\right] \exp\left[i \oint_{\gamma_2} \mathrm{d}x_2^\mu\, A_\mu\right] e^{iS_{\mathrm{CS}}[A]} = \\ &= \frac{1}{\mathcal{Z}} \int \mathcal{D}A\, W_1 W_2\, e^{iS_{\mathrm{CS}}} \equiv \big\langle W_1 W_2 \big\rangle. \end{aligned} \tag{2.16}$$

The linking number. Because the sourced CS action $\widetilde{S}_{\mathrm{CS}}[A, J]$ is quadratic in A, the path integral (2.13) is Gaussian. Thus it can be evaluated exactly (up to an irrelevant normalization constant) by substituting the classical solution A_μ^{cl} into (2.16). We have already found this solution in Coulomb gauge, but it is also useful to do it in Lorenz gauge, where[2]

$$A_\mu^{\mathrm{cl}}(x) = \frac{1}{2k} \int_M \mathrm{d}^3y\, \varepsilon_{\mu\nu\rho} \frac{\partial^\nu J^\rho(y)}{|x-y|} = \frac{1}{2k} \sum_{a=1}^{2} \oint_{\gamma_a} \mathrm{d}x_a^\nu\, \varepsilon_{\mu\nu\rho} \frac{(x-x_a)^\rho}{|x-x_a|^3}. \tag{2.17}$$

Here we have substituted J from (2.14) and integrated by parts. Next, we plug this solution into $\widetilde{S}_{\mathrm{CS}}$. Being quadratic in A, the sourced action consists of terms with two integrals over the loops γ_a; this fact captures the nonlocal nature of the interactions between the two (localized) particles. The terms that integrate twice over the same loop are divergent: they describe self-interactions, and are present even if J describes only a single particle.

After substituting (2.17) into (2.16), removing the self-interaction terms (or absorbing them into the normalization constant), and working through the algebra, we find

$$\mathcal{Z}[J] = \big\langle W_1 W_2 \big\rangle = \exp\Big(i\widetilde{S}_{\mathrm{CS}}[A^{\mathrm{cl}}]\Big) = \exp\left(\frac{i}{2k} \oint_{\gamma_1} \mathrm{d}x_1^\mu \oint_{\gamma_2} \mathrm{d}x_2^\nu\, \varepsilon_{\mu\nu\rho} \frac{(x_1-x_2)^\rho}{|x_1-x_2|^3}\right). \tag{2.18}$$

The integral in the exponent is related to a quantity called the *Gauß linking integral*:

$$\Phi[\gamma_1, \gamma_2] \equiv \frac{1}{4\pi} \oint_{\gamma_1} \mathrm{d}x_1^\mu \oint_{\gamma_2} \mathrm{d}x_2^\nu\, \varepsilon_{\mu\nu\rho} \frac{(x_1-x_2)^\rho}{|x_1-x_2|^3} \implies \big\langle W_1 W_2 \big\rangle = \exp\left(\frac{2\pi i}{k} \Phi[\gamma_1, \gamma_2]\right). \tag{2.19}$$

Gauß proved that $\Phi[\gamma_1, \gamma_2]$ is an integer, and that it is a topological invariant called the

[2]This is a very subtle step. In Lorenz gauge, the CS equations of motion (2.4) reduce to the standard equations of (Maxwell) electrostatics in flat 4-dimensional spacetime $\mathbb{R}^{3,1}$. In the notation of (2.17), the Greek indices μ, ν, ρ are the *spatial* indices of $\mathbb{R}^{3,1}$, and x, y, etc. are regarded as (spatial) vectors in $\mathbb{R}^3$. The quantity $|x-y|$ therefore instructs one to compute the Euclidean distance between x and y in $\mathbb{R}^3$.

linking number of γ_1 and γ_2, which counts the number of times that one curve winds around the other. Taking $\Phi = 1$ gives back the anyonic phase $\langle W_1 W_2 \rangle = e^{2\pi i/k}$ obtained in (2.9).

Interpretation. Fig. 1 provides two beautiful realizations of this mathematical story. First, the Gauß integral is familiar to students of electrodynamics: $\Phi[\gamma_1, \gamma_2]$ computes the magnetic flux passing through a single coil of wire due to a loop of current that passes through it [7]. Moreover, Φ remains the same when the coil of wire and current loop are exchanged: $\Phi[\gamma_1, \gamma_2] = \Phi[\gamma_2, \gamma_1]$. In CS theory, this is relevant because the magnetic flux lines extending from their source charges produce flux through each other in proportion to how tangled they are; this, in turn, detects the anyonic exchange phase of the charges themselves. Second, observe that the integrand of Φ is the Jacobian of the *Gauß map* $\Gamma\colon (x_1, x_2) \mapsto \frac{x_1 - x_2}{|x_1 - x_2|}$, which projects γ_1 and γ_2 onto the unit sphere $S^2 \subset \mathbb{R}^3$ and identifies their "crossing" points. The projection of the magnetic flux lines to the plane gives a visual representation of the Gauß map itself, and its determinant appears even in the classical solutions (2.7) and (2.17). So one should not be entirely surprised to see the linking integral show up.

The point here is that Wilson loops compute topological invariants of knots which are also physical observables. As we will soon see, CS theory is *really* good at doing this.

2.3 Canonical Quantization

Classical treatment. We now turn to the Hamiltonian formalism to see what it can teach us about the phase space of CS theory. To begin, the CS action (2.1) can be decomposed into its A_0 and A_i pieces and then integrated by parts:

$$S_{\mathrm{CS}} = \frac{k}{4\pi} \int_M \mathrm{d}^3x \left(\varepsilon^{ij} A_i \dot{A}_j + 2A_0 \varepsilon^{ij} \partial_i A_j \right) = \frac{k}{2\pi} \int_M \mathrm{d}^3x \left(\frac{1}{2} \varepsilon^{ij} A_i \dot{A}_j + A_0 B \right). \tag{2.20}$$

The field A_0 appears linearly and with no derivatives in S_{CS}. It is a cyclic coordinate, and is often called a non-dynamical or Lagrange multiplier field. Its equation of motion is

$$\frac{\partial S_{\mathrm{CS}}}{\partial A_0} = 0 \implies \frac{k}{2\pi} B = \frac{k}{2\pi} F_{12} = 0. \tag{2.21}$$

The condition $F_{12} = 0$ is the *Gauß law constraint*; it rules out magnetic monopoles.

If we impose it at the level of the action, we can get rid of the second term in (2.20):

$$S_{\mathrm{CS}} = \frac{k}{4\pi} \int_M \mathrm{d}^3x \, \varepsilon^{ij} A_i \dot{A}_j. \tag{2.22}$$

The Lagrangian $\varepsilon^{ij} A_i \dot{A}_j$ takes the canonical form $L = p_i \dot{x}^i - H$, with coordinates $x^i \propto \varepsilon^{ij} A_j$, momenta $p_i \propto A_i$, and (vanishing!) Hamiltonian $H = 0$. Thus $A_\mu = (0, A_1, A_2)$ contains both the dynamical fields and their conjugate momenta: it describes a point in the phase space of a theory governed by a vanishing Hamiltonian. Unfortunately, these degrees of

freedom are redundant, since A_μ transforms nontrivially under gauge transformations.

Topological aside: CS on the torus. In fact, a gauge-invariant parametrization of the CS phase space is provided by the topologically nontrivial Wilson loops in the theory. There are no such Wilson loops in pure CS theory on $\mathbb{R}^3$ or S^3, both of which are simply connected. So the CS phase space there is actually empty! By contrast, in §2.2 we introduced Wilson loops by hand via a source term. (One could alternatively argue that the infinite charge density of the sources introduced defects that changed the global topology.) In any case, our analysis of sources is equivalent to the study of pure CS theory on a manifold with two nontrivial homotopy classes of loops; for instance, on the torus $M = T^2 \times \mathbb{R}$, which has $\pi_1(M) = \pi_1(T^2) = \mathbb{Z} \oplus \mathbb{Z}$. As shown in Fig. 2, these generators give rise to the two Wilson loops discussed above. The W_i are both pure phases that take values in U(1), so the classical phase space for CS theory on the torus is the torus itself: $\mathcal{P} = \mathrm{U}(1) \times \mathrm{U}(1) = T^2$.

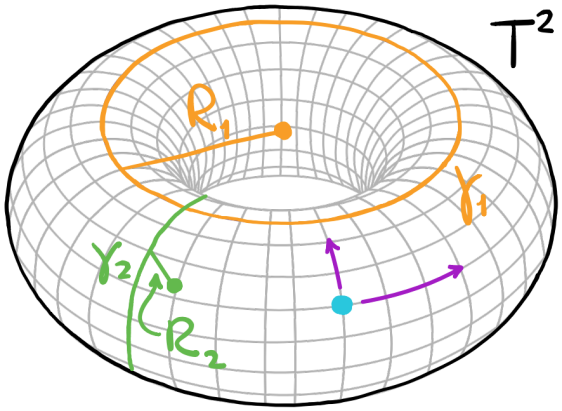

Figure 2: An anyon moving on a torus.

Canonical quantization. The canonical quantization of a gauge theory consists of three steps: (1) impose gauge invariance by fixing a gauge; (2) impose any non-dynamical constraints; and (3) promote the canonical Poisson brackets to commutators. The order of these steps does not affect the quantum theory they produce, but (depending on the theory) choosing the wrong order could lead to insurmountable difficulties. In the present case, we start from the action (2.20) and choose the gauge $A_0 = 0$, yielding (2.22). This step is identical to imposing the constraint $F_{12} = 0$ in the action, but one must remember that Gauß's law must be implemented throughout the phase space. At this stage, the *solutions* to the equations of motion are still unconstrained, i.e. they may have $F_{12} = \partial_1 A_2 - \partial_2 A_1 \neq 0$.

Here we will choose to quantize first, and then impose the constraint: this is a viable strategy because both the equations of motion and the constraint are linear in A. In §4, we will pursue the opposite strategy in a setting where the Gauß law is nonlinear. The canonical

Poisson brackets may be read off from (2.22), and the A_i promoted to operators:

$$\big\{A_i(\mathbf{x}), A_j(\mathbf{y})\big\} = \frac{2\pi}{k}\varepsilon_{ij}\delta^{(2)}(\mathbf{x}-\mathbf{y}) \;\rightsquigarrow\; \big[A_i(\mathbf{x}),\, A_j(\mathbf{y})\big] = \left(\frac{2\pi i}{k}\right)\varepsilon_{ij}\delta^{(2)}(\mathbf{x}-\mathbf{y}). \quad (2.23)$$

Solving the theory. We now seek to determine the ground state wave functional $\Psi_0[A(x)]$. This is rarely ever possible in field theory, but in $A_0 = 0$ gauge the CS action is quadratic—the theory is free—so we expect that the ground state will be Gaussian. For reasons motivated by Landau levels in condensed matter theory[3] (see [2]), we adopt complex coordinates:

$$z = x_1 + ix_2, \qquad A = A(z) = \sqrt{\frac{k}{2\pi}}(A_1 + iA_2). \quad (2.24)$$

We take the following ansatz for the ground state, hoping to determine it explicitly:

$$\Psi_0[A, \overline{A}] = \Psi[A(z)] \exp\left(-\frac{1}{2}\int |A|^2\right). \quad (2.25)$$

Now we must impose $F_{12} = 0$ as a functional equation on the Hilbert space. To do so, we first rewrite the commutator (2.23) in terms of the rescaled fields A and $\overline{A}$, which gives $[A(z), \overline{A}(w)] = \delta(z-w)$. We gain the right to view A as "x" and $i\overline{A}$ as "p," so in analogy to quantum mechanics we write $\overline{A} = -\frac{\delta}{\delta A}$. This makes $F_{12} = \partial_1 A_2 - \partial_2 A_1$ a differential operator that we can reorganize in terms of the (anti-)holomorphic derivatives $\partial_\pm = \partial_1 \mp i\partial_2$:

$$F_{12}\Psi_0[A, \overline{A}] = \left(\partial_- \frac{\delta}{\delta A} + \partial_+ A\right)\Psi_0[A, \overline{A}] = 0 \implies \left(\partial_- \frac{\delta}{\delta A} + \partial_+ A\right)\Psi[A(z)] = 0. \quad (2.26)$$

The factor $e^{-\frac{1}{2}\int |A|^2}$ drops out, and the resulting ODE for $\Psi[A(z)]$ has a Gaussian solution:

$$\Psi[A] \sim \exp\left[-\frac{1}{2}\int A\left(\frac{\partial_+}{\partial_-}\right)A\right] \implies \Psi_0[A(z)] = \exp\left(-\frac{1}{2}\int \left[A\left(\frac{\partial_+}{\partial_-}\right)A + |A|^2\right]\right). \quad (2.27)$$

The torus again. Earlier, we studied the classical phase space of pure CS theory on the torus. Now, with the commutator (2.23) in hand, we are ready to quantize the Wilson loops. We write $W_a = e^{iw_a}$, with $w_a = \oint_{\gamma_a} \mathrm{d}x_a^\mu A_\mu$. In $A_0 = 0$ gauge, we substitute the canonical commutator (2.23) and integrate to find $[w_1, w_2] = \frac{2\pi i}{k}$. The Baker–Campbell–Hausdorff formula yields the algebra obeyed by W_1 and W_2: $W_1 W_2 = e^{2\pi i/k} W_2 W_1$. The smallest nontrivial representation of this algebra has dimension $k \in \mathbb{Z}$, so fields in this representation produce a k-fold ground state degeneracy. On a genus-g surface, there are more Wilson loops; the correponding algebras are more complicated, and the ground state degeneracy grows to k^g. This degeneracy is one hallmark of a novel, *topological phase* of matter.

[3] Alternatively, the $A_0 = 0$ gauge reduces the physics to the plane $\mathbb{R}^2 = \mathbb{C}$, where complex coordinates are readily available. The discussion of holomorphic quantization in §4 will also echo and extend this idea.

2.4 Topological Bells and Whistles

An invariant perspective. Let us see what a coordinate-free formulation of CS theory can do for us. To begin, the CS action (2.1) can be recast in terms of differential forms:

$$S_{\mathrm{CS}} = \frac{k}{4\pi}\int_M A\wedge \mathrm{d}A, \qquad A = A_\mu \mathrm{d}x^\mu \in \Omega^1(M). \tag{2.28}$$

The CS 3-form $A\wedge \mathrm{d}A$ can be defined on any orientable 3-manifold M, without reference to a metric. Its integral is a topological invariant of M, and in particular it is invariant under arbitrary coordinate transformations. Just as in gravity, we regard coordinate transformations as gauge symmetries. One of these is the flow of time itself: the corresponding Noether charge, the Hamiltonian, must vanish identically, confirming our canonical analysis in §2.3. Moreover, the independence of S_{CS} from the metric causes the entire stress tensor to vanish:

$$T^{\mu\nu}_{\mathrm{CS}} = \frac{2}{\sqrt{|g|}}\frac{\delta \mathcal{L}_{\mathrm{CS}}}{\delta g_{\mu\nu}} = 0. \tag{2.29}$$

These features—diffeomorphism invariance and $H = 0$—are shared by both CS theory and gravity, and in fact [8] 3-dimensional gravity has a first-order formulation as a CS theory!

The θ term. One advantage of differential forms is that they clarify how S_{CS} is related to the *theta term* in 4D Yang–Mills theory. The main idea [3] is to view M as the boundary of some 4-manifold X, and then to use Stokes's theorem and $F = \mathrm{d}A$ to write

$$S^X_{\mathrm{CS}} = \frac{k}{4\pi}\int_{\partial X} A\wedge \mathrm{d}A = \frac{k}{4\pi}\int_X F\wedge F = \frac{k}{2\pi}\int_X \mathrm{d}^4x\, F^{MN}\widetilde{F}_{MN}. \tag{2.30}$$

Here we put a metric on X and wrote $F\wedge F$ in components using $\widetilde{F}^{MN} = \frac{1}{2}\varepsilon^{MNAB}F_{AB}$. (Equivalently, the integral of $F\wedge F$ is the Hodge inner product $\langle F, \star F\rangle$ of 2-forms.) So up to overall normalization, we find that the CS action is just a theta term.

Spin manifolds. From the 4D viewpoint, S_{CS} depends only on F and is manifestly gauge invariant. However, it now depends crucially on our choice of "bulk" spacetime X, and this cannot be. In order for the theory to be well defined, we require that for any two 4-manifolds X, X' with common boundary $\partial X = \partial X' = M$, we have $S^X_{\mathrm{CS}} - S^{X'}_{\mathrm{CS}} \in 2\pi\mathbb{Z}$ to preserve the partition function à la (2.12). One slick way to rephrase this criterion is to introduce the compact 4-manifold $Y = (X \cup X')/M$, defined by gluing X and X' together along their common boundary M in an orientation-reversing manner. Then our requirement reads

$$S^X_{\mathrm{CS}} - S^{X'}_{\mathrm{CS}} = \frac{k}{4\pi}\int_Y F\wedge F \overset{!}{=} 2\pi n, \qquad n\in\mathbb{Z}. \tag{2.31}$$

It turns out that this integrality condition holds only when Y is a *spin manifold*, which is a manifold that supports the existence of spinors. For instance, T^4, S^4, and $S^2 \times S^2$ are spin manifolds, but $\mathbb{CP}^2$ is not. Thus we have used geometry to solve a topological problem, and have learned that 3D CS theories are intimately tied to 4D gauge theories with fermions.

3 Framing and the Path Integral

Apologia for geometry. In the prevous section, we were deliberately naïve and chose to formulate the theory in an elementary but perhaps less insightful way. Now we will, in some sense, start over from scratch. Our exposition will be decidedly more formal, but the geometry developed here will always be guided by the intuition built up in the abelian theory. All of the phenomena we have encountered will now be fibered, bundled, outfitted with representations, and laid down abstractly. We will begin by formulating nonabelian the CS theory in full generality, give its behavior under gauge transformation, and describe its physical observables (Wilson loops) in terms of knots. We will then pass to the saddle-point evaluation of its partition function at large k, where the theory is weakly coupled. To regulate the rather subtle divergences that show up in the process, we will discuss the framing of oriented 3-manifolds and of knots. In the end, understanding the behavior of the path integral under change of framing will allow us to complete the calculations.

3.1 The Nonabelian Chern–Simons Action

Geometrical setting. Our story begins with an oriented 3-manifold M representing spacetime, and a compact, simple Lie group G—the gauge group—with Lie algebra $\mathfrak{g}$. As physicists, we sometimes confuse Lie groups with their Lie algebras. While we will attempt to distinguish G from $\mathfrak{g}$ here, we will inevitably be imprecise at times. To formalize the notion of a gauge transformation, fix a principal G-bundle[4] $E \longrightarrow M$. Sections of E, called *gauge transformations*, are maps $g\colon M \longrightarrow E$ which smoothly assign, to each $x \in M$, a point $g(x) = (x, g_x) \in E$, where $g_x \in G$. We often abuse notation and write $g(x) = g_x$, thinking of g as a map $M \longrightarrow G$ that produces a spacetime-dependent element of G. When g_x lies close to the identity of G, we consider instead infinitesimal (generators of) gauge transformations, and we view $g(x)$ as a $\mathfrak{g}$-valued zero-form, i.e. a spacetime-dependent Lie algebra generator.

Next, we introduce a principal connection on E. To a physicist,[5] this is nothing more than a one-form on M that takes values in $\mathfrak{g}$. We denote the connection by A; in local coordinates on M, we write $A = A_\mu(x)\mathrm{d}x^\mu$. Each $A_\mu(x)$ lies in the Lie algebra, so given a basis T^a of $\mathfrak{g}$, we may further decompose $A_\mu(x) = A^a_\mu(x)T^a$, with $a \in \{1, ..., d = \dim G\}$ the gauge index. The object $A_\mu(x)$, called the *gauge field*, generalizes the U(1) gauge field of §2. Now we can ask a natural question: given that E comes equipped with a G-action,

[4]For convenience, one may imagine that $E = M \times G$ is the trivial bundle, with $M = S^3$ and $G = \mathrm{SU}(N)$. Our formalism applies more generally, but this is a useful example to return to when the going gets tough.

[5]For a proper mathematical exposition, including wisdom on associated bundles, see [9] or [10].

what happens to the connection under this action? In other words: how does the gauge field change under gauge transformations? The answer is the following transformation law:

$$A_\mu \longrightarrow A'_\mu = g^{-1}A_\mu g + g^{-1}\partial_\mu g, \qquad g = g(x) \in E. \tag{3.1}$$

Covariant derivative and curvature. To obtain the infinitesimal version of this law, we introduce the *gauge covariant derivative* D. It acts on $\mathfrak{g}$-valued differential forms ω by $D_\mu\omega = \partial_\mu\omega + [A_\mu, \omega]$, where $[\,\cdot\,,\cdot\,]$ is the Lie bracket. Under a local gauge transformation $\varepsilon(x) \in \Omega^0(M, \mathfrak{g})$, the gauge field transforms by a covariant derivative: $A_\mu \longrightarrow A_\mu - D_\mu\varepsilon$.

One says that the ordinary derivative ∂_μ is "twisted" by the connection via $[A_\mu, \cdot\,]$. We can examine the amount of twisting introduced by A_μ by measuring the degree to which D_μ fails to commute with itself, as would be expected if the gauge bundle were flat. Therefore we define the curvature of the connection, also called the *gauge field strength*, by

$$F_{\mu\nu} \equiv [D_\mu, D_\nu] = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu] \iff F = \mathrm{d}A + A \wedge A. \tag{3.2}$$

If G is abelian (for instance, if $G = \mathrm{U}(1)$), then all commutators vanish. Then $D_\mu = \partial_\mu$, and $F_{\mu\nu}$ reduces to the usual Maxwell field strength tensor. However, the nonabelian field strength contains a term quadratic in A and is generally nonlinear. If $F = 0$, then we say that the connection A is *flat*. We shall soon see, taking inspiration from (2.2), that the space of flat connections on E determines the classical phase space of pure CS theory.

The Chern–Simons action. Starting from A, we define the *Chern–Simons action* by

$$\begin{aligned} S_{\mathrm{CS}}[A] &\equiv \frac{k}{4\pi}\int_M \mathrm{Tr}\left(A \wedge \mathrm{d}A + \frac{2}{3}A \wedge A \wedge A\right) = \\ &= \frac{k}{4\pi}\int_M \mathrm{d}^3x\, \varepsilon^{\mu\nu\rho}\,\mathrm{Tr}\left(A_\mu(\partial_\nu A_\rho - \partial_\rho A_\nu) + \frac{2}{3}A_\mu[A_\nu, A_\rho]\right). \end{aligned} \tag{3.3}$$

The $\mathfrak{g}$-valued 3-form $\omega_{\mathrm{CS}} \equiv A \wedge \mathrm{d}A + \frac{2}{3}A \wedge A \wedge A$ is the famous *Chern–Simons form*, and by Tr we mean a multiple of the Killing form on $\mathfrak{g}$ (we will fix the normalization shortly). To be more concrete, in the adjoint representation of $\mathfrak{g}$, ω_{CS} is just a matrix (of 3-forms), and the trace of this matrix is proportional to $\mathrm{Tr}(\omega)$. The equations of motion that follow from (3.3), obtained from the variation δS_{CS} in response to a field variation δA, are $\delta S_{\mathrm{CS}} = 0 \implies F^a_{\mu\nu} = 0$. As promised, this describes the space of flat connections on E.

Quantization of the level. Echoing §2.2, we ask whether the CS action (3.3) is invariant under the gauge transformation (3.1). A brief exercise in integration by parts shows that

$$\delta S_{\mathrm{CS}} = \frac{k}{4\pi}\int_M \mathrm{d}^3x\, \varepsilon^{\mu\nu\rho}\left(\partial_\mu\,\mathrm{Tr}\left[(\partial_\nu g)(g^{-1}A_\rho)\right] + \frac{1}{3}\,\mathrm{Tr}\left[(g^{-1}\partial_\mu g)(g^{-1}\partial_\nu g)(g^{-1}\partial_\rho g)\right]\right). \tag{3.4}$$

The first term is the familiar total divergence (2.10) in disguise, and (as in the abelian case) it vanishes once suitable boundary conditions are imposed. The second term is a novelty of the nonabelian theory, and is related to an integral called the *winding number* of g:

$$w(g) \equiv \frac{1}{24\pi^2} \int_M \mathrm{d}^3x\, \varepsilon^{\mu\nu\rho}\, \mathrm{Tr}\left[\left(g^{-1}\partial_\mu g\right)\left(g^{-1}\partial_\nu g\right)\left(g^{-1}\partial_\rho g\right)\right] = (\text{constant}) \cdot n, \quad n \in \mathbb{Z}. \tag{3.5}$$

That $w(g)$ is quantized is well known from the study of instantons (see [11] for a proof). Thus we fix the normalization of Tr so that $w(g)$ is precisely an integer. In terms of $w(g)$, the CS action changes by a constant under gauge transformations: $S_{\mathrm{CS}} \longrightarrow S_{\mathrm{CS}} + 2\pi k w(g)$. Transformations with $w(g) = 0$ leave S_{CS} invariant, but those with $w(g) \neq 0$ do not. Those $g(x)$ with nonzero winding number are called *large gauge transformations*; they "wrap" nontrivially around G. Such transformations cause the classical CS theory to be ill defined, but (by the argument in §2.2) the quantum theory is still consistent as long as the amplitude $e^{-S^{\mathrm{E}}_{\mathrm{CS}}} = e^{iS_{\mathrm{CS}}}$ is single-valued. This forces $2\pi k w(g)$ to be an integer, which requires $k \in \mathbb{Z}$.

Topology of gauge transformations. The appearance of $w(g)$ may seem like a miracle, but it is really a consequence of the topology of G. Recall that gauge transformations are maps $g\colon M \longrightarrow G$ sending $x \mapsto g(x)$. Among these is the identity map $x \mapsto 1_G$, and one may ask whether all gauge transformations are homotopic (i.e. continuously deformable) to the identity. The answer is no: if (for instance) $M = S^3$, then the set of maps $g\colon S^3 \longrightarrow G$ is classified by the group $\pi_3(G)$, and it is a famous result of Bott [12] that $\pi_3(G) = \mathbb{Z}$ for every compact, simple Lie group G. Hence each map $g(x)$ is classified by an integer $w(g) = [g] \in \pi_3(G) = \mathbb{Z}$ called its *winding number*, and this integer is the integral (3.5).

Wilson loops and knots. Recall that CS theory is topological and admits no local gauge-invariant observables, as these would violate general covariance. Instead, it computes topological invariants via Wilson loops, which we construct in the nonabelian theory as follows. Fix an irreducible representation R of G, and let γ be an oriented, closed curve in M, called a *knot*. The connection A, viewed as a 1-form, may be be integrated over γ to yield an element of $\mathfrak{g}$. The (path-ordered) exponential of this integral gives an element of G, well defined up to conjugacy: this element is the *holonomy* of the connection around γ. The *Wilson loop* $W_R[\gamma]$ is then defined as the trace, in the representation R, of that holonomy:

$$W_R[\gamma] \equiv \mathrm{Tr}_R\left[\mathcal{P}\exp\left(\oint_\gamma \mathrm{d}x^\mu\, A_\mu\right)\right]. \tag{3.6}$$

Despite appearances, this definition never requires one to choose a metric on M, so $W_R[\gamma]$ is a topological invariant of γ. Next, consider r oriented and non-intersecting closed curves $\gamma_1, ..., \gamma_r$ in M. Their disjoint union, $L = \bigcup_{a=1}^r \gamma_a$, is called a *link*. Choose irreducible

representations R_a of G for each curve γ_a, and consider the following path integral:

$$\mathcal{Z}(M,L) \equiv \frac{1}{\mathcal{Z}} \int \mathcal{D}A\, e^{iS_{\rm CS}[A]} \prod_{a=1}^{r} W_{R_a}[\gamma_a] = \left\langle \prod_{a=1}^{r} W_{R_a}[\gamma_a] \right\rangle. \tag{3.7}$$

Here the normalization factor $\mathcal{Z} = \mathcal{Z}(M,\emptyset)$ is computed in the absence of Wilson loops. The expectation value $\mathcal{Z}(M,L)$ will be our main object of study: in fact, Witten showed that when $G = \mathrm{SU}(2)$ and $M = S^3$, $\mathcal{Z}(S^3,L)$ is precisely the Jones polynomial of the link L.

Comments. Observe that the orientation of any loop $\gamma \subset L$ gives the direction in which a particle, charged under the corresponding representation R, moves around that loop. Reversing the orientation of γ is equivalent to conjugating R, so taking both $\gamma \longrightarrow -\gamma$ and $R \longrightarrow \overline{R}$ leaves $\mathcal{Z}(M,L)$ invariant. Furthermore, reversing the orientation of all of the loops in L is equivalent to conjugating all of their representations. This operation, called *charge conjugation*, leaves the CS action invariant, and it also leaves $\mathcal{Z}(M,L)$ unchanged.

3.2 The Path Integral at Weak Coupling

Stationary phase. We are finally ready to consider the CS path integral in earnest. We begin in the weakly coupled limit, which corresponds to large k. (Indeed, anyonic phases and other observables all depend on $\frac{1}{k}$, and become small when k is large.) In the absence of Wilson loops, the CS path-integrand is rapidly oscillating because $S_{\rm CS}$ itself is large:

$$\mathcal{Z} = \int \mathcal{D}A\, e^{iS_{\rm CS}[A]} = \int \mathcal{D}A\, \exp\left[\frac{ik}{4\pi}\int_M \mathrm{Tr}\left(A\wedge \mathrm{d}A + \frac{2}{3}A\wedge A\wedge A\right)\right]. \tag{3.8}$$

In the weakly coupled limit, $\mathcal{Z}$ is dominated by contributions from the points of stationary phase. Such field configurations $A^{(\alpha)}$ are solutions to the classical equations of motion, i.e. flat connections. It will be convenient to evaluate $S_{\rm CS}[A^{(\alpha)}]$ by factoring out k to define the purely topological quantity $I[A^{(\alpha)}]$, called the *Chern–Simons invariant* of $A^{(\alpha)}$:

$$I[A^{(\alpha)}] \equiv \frac{1}{4\pi}\int_M \mathrm{Tr}\left(A^{(\alpha)}\wedge \mathrm{d}A^{(\alpha)} + \frac{2}{3}A^{(\alpha)}\wedge A^{(\alpha)}\wedge A^{(\alpha)}\right) \implies S_{\rm CS}[A^{(\alpha)}] = kI[A^{(\alpha)}]. \tag{3.9}$$

Denoting each saddle-point contribution by $\mu[A^{(\alpha)}]$, our first attempt to evaluate $\mathcal{Z}$ reads

$$\mathcal{Z} \approx \sum_\alpha \mu[A^{(\alpha)}] \approx \sum_\alpha e^{iS_{\rm CS}[A^{(\alpha)}]} = \sum_\alpha e^{ikI[A^{(\alpha)}]}. \tag{3.10}$$

Here the index α runs over the set $\mathcal{M}$, which is called the *moduli space of flat connections*. For the moment we assume that each $A^{(\alpha)}$ is isolated and that $\mathcal{M}$ is finite, but this is false in general. (As we will discuss in §4, $\mathcal{M}$ usually looks like a manifold with some singularities.) The result of this exercise is that by summing over contributions from all flat connections,

$\mathcal{Z}$ computes a topological invariant of M. In the rest of this section, we will consider small fluctuations about each $A^{(\alpha)}$ and carry out a one-loop version of the calculation above.

Fluctuations. To begin, we expand $A = A^{(\alpha)} + a$ around a "background" flat connection, substitute this decomposition into (3.8), and change variables in the path integral to the fluctuation field a. The action itself expands in powers of a, the lowest few terms being

$$S_{\rm CS}[A] = kI[A^{(\alpha)}] + \frac{k}{4\pi}\int_M \mathrm{Tr}(a\wedge Da) = kI[A^{(\alpha)}] + \frac{k}{4\pi}\int_M \mathrm{d}^3x\,\varepsilon^{\mu\nu\rho}\,\mathrm{Tr}(a_\mu D_\nu a_\rho). \quad (3.11)$$

Here and henceforth, it will be understood that $D_\mu = \partial_\mu + [A^{(\alpha)}_\mu, \cdot\,]$ is the covariant derivative with respect to the background flat connection. For each such flat connection $A^{(\alpha)}$, the corresponding saddle point contribution $\mu[A^{(\alpha)}]$ to the path integral (3.8) is

$$\mu[A^{(\alpha)}] = e^{ikI[A^{(\alpha)}]}\int \mathcal{D}a\,\exp\left[\frac{ik}{4\pi}\int_M \mathrm{Tr}(a\wedge Da)\right] \equiv e^{ikI[A^{(\alpha)}]}\int \mathcal{D}a\,e^{iS[a]}, \quad (3.12)$$

where we have ignored higher-order terms in the action. This path integral is Gaussian, but to evaluate it we must choose a gauge, and this requires us to put a metric on M. Nevertheless, we aim for a computation of $\mathcal{Z}$ in terms of purely topological invariants of M.

Gauge fixing. We choose the "Lorenz" gauge $D_\mu a^\mu = 0$. To implement it, we follow the Faddeev–Popov gauge fixing procedure. The construction introduces into the path integral an auxiliary Lagrange multiplier field ϕ, as well as the anticommuting but bosonic ghost fields c and $\bar{c}$. (The fields ϕ, $\bar{c}$, and c are all scalars, but following Witten–Bar-Natan [13], we view ϕ as a 3-form.) To calculate the one-loop contribution $\mu[A^{(\alpha)}]$, we augment the action $S[a]$ appearing in (3.12) by adding gauge fixing and ghost terms, and then integrate over all of the auxiliary fields. Towards the first step, we change the action to

$$\begin{aligned} S[a] \longrightarrow S[a] + S_{\rm gf}[\phi,a] + iS_{\rm gh}[\bar{c},c] &\equiv \int_M \mathrm{d}^3x\;\mathrm{Tr}\left(\frac{k}{4\pi}\varepsilon^{\mu\nu\rho}a_\mu D_\nu a_\rho + \phi D_\mu a^\mu + i\bar{c}D_\mu D^\mu c\right) = \\ &= \int_M \mathrm{Tr}\left(\frac{k}{4\pi}a\wedge Da + \phi\star D\star a + i\bar{c}D\star Dc\right). \end{aligned} \quad (3.13)$$

The one-loop contribution from each flat connection $A^{(\alpha)}$ is therefore

$$\mu[A^{(\alpha)}] = e^{ikI[A^{(\alpha)}]}\int \mathcal{D}[a,\phi,\bar{c},c]\exp\left[i\int_M \mathrm{Tr}\left(\frac{k}{4\pi}a\wedge Da + \phi\star D\star a + i\bar{c}D\star Dc\right)\right]. \quad (3.14)$$

One-loop determinants. The path integral (3.14) is Gaussian and falls apart into pieces that can be evaluated in terms of the determinants of certain operators. The ghost fields decouple from the rest of the action because they appear only in the kinetic term $\bar{c}D_\mu D^\mu c$.

The Grassmann integral over $\bar{c}$ and c yields the determinant of the kinetic operator:

$$\int \mathcal{D}\bar{c}\,\mathcal{D}c\,\exp\Big[-\int_M \mathrm{Tr}\left(\bar{c}D_\mu D^\mu c\right)\Big] = \det(\Delta), \qquad \Delta \equiv D_\mu D^\mu = \star D \star D. \tag{3.15}$$

The remainder of the path integral can be evaluated in terms of an operator called L_-, which is constructed as follows. Consider first the operator $L = \star D + D\star$, which resembles a Dirac operator[6] in the sense that its square contains the Laplacian (among other curvature terms). In 3 dimensions, L is self-adjoint and maps forms of even (resp. odd) degree to forms of even (resp. odd) degree, so we may define its restrictions L_+ and L_- to forms of even and odd degree, respectively. We then combine a and ϕ into an odd-form field $H = (a, \phi)$. It can be shown, after rescaling a and ϕ, that the non-ghost part of the action (3.13) is precisely the Hodge inner product $\frac{1}{2}\langle H, L_- H\rangle$. Its path integral is therefore a one-loop determinant:

$$\begin{aligned}\mu_0[A^{(\alpha)}] &\equiv \int \mathcal{D}\phi\,\mathcal{D}a\,\exp\left[i\int_M \mathrm{Tr}\left(\frac{k}{4\pi}a\wedge Da + \phi\star D\star a\right)\right] = \\ &= \int \mathcal{D}H\,\exp\left[\frac{i}{2}\int_M \mathrm{Tr}\left(H\wedge\star(L_- H)\right)\right] = \frac{1}{\sqrt{\det(L_-)}}.\end{aligned} \tag{3.16}$$

Bringing (3.14) and (3.15–3.16) together, we obtain the one-loop contribution from $A^{(\alpha)}$:

$$\mu[A^{(\alpha)}] = e^{ikI[A^{(\alpha)}]}\mu_0[A^{(\alpha)}]\det(\Delta) = e^{ikI[A^{(\alpha)}]}\left(\frac{\det(\Delta)}{\sqrt{\det(L_-)}}\right). \tag{3.17}$$

The Ray–Singer torsion. The result (3.17) is admittedly not very explicit. More worryingly, it depends on the metric of M. In spite of these concerns, it is a marvelous result of Schwarz [14] that the absolute value of the ratio $\det(\Delta)/\sqrt{\det(L_-)}$ is a topological invariant called the *Ray–Singer torsion* T_α of the flat connection $A^{(\alpha)}$. The phase of this ratio, however, is potentially troublesome. Denoting that phase by θ_α for the moment, we can write down the one-loop partition function (3.8) as a sum of its saddle-point contributions:

$$\mu[A^{(\alpha)}] = e^{ikI[A^{(\alpha)}]}\big(T_\alpha\, e^{i\theta_\alpha}\big) \implies \mathcal{Z} = \sum_\alpha e^{i\left(kI[A^{(\alpha)}]+\theta_\alpha\right)}\,T_\alpha. \tag{3.18}$$

Because the Laplacian Δ is positive and self-adjoint, we actually have $\det(\Delta) \in \mathbb{R}_+$. Thus it remains to study the phase of $\det(L_-)$; as we shall see, this turns out to be tricky.

The phase—I. Let us examine the path integral (3.16) more closely by passing to an eigenbasis of L_-. We call the eigenfunctions χ_j and their eigenvalues λ_j, so that $L_-\chi_j = \lambda_j\chi_j$.

[6] This is our second hint—after the comments in §2.4—that the geometry of spinors is close by.

We expand the field H, introduced above, as $H = \sum_j h_j \chi_j$, and obtain

$$\mu_0[A^{(\alpha)}] = \int \mathcal{D}H \, \exp\left[\frac{i}{2}\int_M \mathrm{Tr}\left(H \wedge \star(L_- H)\right)\right] = \prod_j \int_{-\infty}^{\infty} \frac{\mathrm{d}h_j}{\sqrt{2\pi}} \, e^{i\lambda_j h_j^2}, \tag{3.19}$$

where the repreated indices in the exponent are not summed. This change of variables produces wildly oscillatory integrals which reveal that the phase of $\det(L_-)$ is divergent. It can be regularized, however: in one dimension, the analogous integral is

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}x}{\sqrt{2\pi}} \, e^{i\lambda x^2} = \lim_{\varepsilon\to 0} \int_{-\infty}^{\infty} \frac{\mathrm{d}x}{\sqrt{2\pi}} \, e^{i\lambda x} e^{-\varepsilon x^2} = \left|\frac{1}{\sqrt{\lambda}}\right| \exp\left(\frac{i\pi}{4}\mathrm{sign}(\lambda)\right). \tag{3.20}$$

Taking the product in (3.19), we obtain both the magnitude and phase of $\det(L_-)$:

$$\mu_0[A^{(\alpha)}] = \prod_j \left|\frac{1}{\sqrt{\lambda_j}}\right| \exp\left(\frac{i\pi}{4}\mathrm{sign}(\lambda_j)\right) = \left|\frac{1}{\sqrt{\det(L_-)}}\right| \exp\left(\frac{i\pi}{4}\sum_j \mathrm{sign}(\lambda_j)\right). \tag{3.21}$$

Hence the phase of the determinant depends on the *signature* of L_-, the difference between its number of its positive eigenvalues and its number of negative eigenvalues. This number is ill defined, but it can be regularized (again!) by introducing the *eta invariant* [15, 16, 17]:

$$\mu_0[A^{(\alpha)}] = \frac{1}{\sqrt{\det(L_-)}} = \left|\frac{1}{\sqrt{\det(L_-)}}\right| e^{i\pi\eta[A^{(\alpha)}]/2}, \qquad \eta[A^{(\alpha)}] \equiv \frac{1}{2}\lim_{s\to 0}\sum_{\lambda\neq 0}\frac{\mathrm{sign}(\lambda)}{|\lambda|^s}. \tag{3.22}$$

The phase—II. Towards an explicit computation of the phase, the *Atiyah–Patodi–Singer index theorem* gives a relation between $\eta[A^{(\alpha)}]$ and other topological data of the gauge bundle:

$$\frac{1}{2}\Big(\eta[A^{(\alpha)}] - \eta[0]\Big) = \frac{c_2(G)}{2\pi} I[A^{(\alpha)}] \iff \theta_\alpha = \frac{\pi}{2}\eta[A^{(\alpha)}] = \frac{c_2(G)}{2} I[A^{(\alpha)}] + \frac{\pi}{2}\eta[0]. \tag{3.23}$$

Here, $I[A^{(\alpha)}]$ is the CS invariant defined in (3.9), $c_2(G)$ is the *quadratic Casimir* of the gauge group, and the mysterious term $\eta[0]$ is the eta invariant of the trivial connection. Putting (3.23) together with (3.18) yields the one-loop partition function:

$$\mathcal{Z} = e^{i\pi\eta[0]/2} \sum_\alpha \exp\left[i\Big(k + \frac{c_2(G)}{2}\Big) I[A^{(\alpha)}]\right] T_\alpha. \tag{3.24}$$

3.3 Regularization and Framing

Summary so far. We began in §3.1 by setting up the geometry of the gauge bundle, introducing the CS action (3.3), and discussing its quantization via large gauge transformations. In §3.2, we wrote down the partition function (3.8) of pure CS theory and attempted its saddle-point evaluation by expanding the action at large k around a flat connection, as

in (3.12). We then chose a metric, fixed the gauge, introduced ghosts, and evaluated the resulting Gaussian path integral in terms of the operators whose quadratic forms made up the gauge-fixed action. The result, (3.17), was a ratio of determinants whose absolute value T_α is a topological invariant. The phase was more subtle, and was eventually expressed in terms of the eta invariant (3.22). Unfortunately, $\eta[A^{(\alpha)}]$ is not a topological invariant, since it depends on L_- and hence on the metric of M. The index theorem gave us (3.24) and improved the situation by making $\eta[0]$ the only part of $\mathcal{Z}$ that is not a topological invariant.

Towards $\eta[0]$, the trivial connection $A^{(\alpha)} = 0$ reduces the covariant derivative to the ordinary one: $D_\mu = \partial_\mu$. Consequently, L_- falls apart into the direct sum of $d = \dim G$ identical copies of the "gravitational" operator $D_- = (\star\,\mathrm{d} + \mathrm{d}\,\star)_{\mathrm{odd}}$, so named because it depends only on the metric of M. We may therefore write $\eta[0] = d\eta_g$, where $\eta_g = \eta_g$ is the eta invariant of D_- at $A^{(\alpha)} = 0$. We would then hope to express the global phase $\Lambda \equiv \exp\bigl(\frac{id\pi}{2}\eta_g\bigr)$ in (3.24) as a topological invariant, but here we run out of luck.

Philosophy of failure. The crucial thing to appreciate is that although we started with a topological quantity, there was no way to regularize or evaluate it without introducing a metric. To the extent that general covariance is a "classical" symmetry of the CS action broken by any regularization of the quantum theory, we expect the theory to have an anomaly—and sure enough, we will soon see one called the *framing anomaly*. More immediately, however, we must tend to the fact that our final answer depends on the metric. This should not be surprising: the gauge fixing terms we added to the action in (3.13) were explicitly coupled to the metric, so our calculation can only be correct up to the effects of such "gravitational" terms in the action. Perhaps the surprising thing is that we almost managed to completely eliminate such effects altogether, except for the global phase factor $\Lambda = \exp\bigl(\frac{id\pi}{2}\eta_g\bigr)$.

To remedy the situation, we should add a compensating term to the action that precisely cancels the gravitational effects we introduced. This prescription is consistent with the ideas of renormalization: any two regularizations must differ by a local counterterm in the action. In our case, we require a counterterm $\mathcal{C}[g]$ constructed *only* from the metric g on M, so that adding $i\mathcal{C}[g]$ to the action will have no effect on $\mathcal{Z}$ other than to add $\mathcal{C}[g]$ to the phase $\frac{d\pi}{2}\eta_g$. We will agree that $\mathcal{C}[g]$ "precisely cancels" the gravitational effects of η_g if the sum $\frac{d\pi}{2}\eta_g + \mathcal{C}[g]$ is a topological invariant, since then it no longer depends on the metric.

The gravitational counterterm. It may be difficult to see what to add to the action. One natural guess is the *gravitational Chern–Simons action*, which is modeled on S_{CS}, but uses the *spin connection* ω on M in place of a principal G-connection A:

$$I[g] \equiv \frac{1}{4\pi}\int_M \mathrm{Tr}\biggl(\omega \wedge \mathrm{d}\omega + \frac{2}{3}\omega \wedge \omega \wedge \omega\biggr). \tag{3.25}$$

Crudely speaking, one may think of ω as the Levi–Civita connection on M. Actually, it is both more illuminating and more precise to regard ω as the Levi–Civita connection on the

spinor bundle of M. This construction immediately gets more concrete due to the happy 3-dimensional fact that Spin(3), the universal cover of SO(3), is isomorphic to SU(2). Thus ω is just an SU(2) gauge field, and its precise field configuration—and hence the value of $I[g]$—is uniquely fixed by the "gravitational" requirement that it be the Levi–Civita connection.

The term $I[g]$ is the crucial ingredient we need. By the Atiyah–Singer index theorem (again!), the combination $\Xi \equiv \frac{1}{2}\eta_g + \frac{1}{12}\frac{I[g]}{2\pi}$ of gravitational terms is a topological invariant. In accordance with the philosophy expounded above, we should replace $\frac{\eta_g}{2}$ with Ξ in the global phase Λ of the partition function. In other words, we should add the counterterm $\mathcal{C}[g] = \frac{\pi d}{12}\frac{I[g]}{2\pi} = \frac{dI[g]}{24}$ to the action, so that the problematic phase in (3.24) becomes

$$\Lambda = \exp\left(\frac{i\pi}{2}\eta[0]\right) = \exp\left(\frac{id\pi}{2}\eta_g\right) \rightsquigarrow \exp\left(i\pi d\left[\frac{\eta_g}{2} + \frac{1}{12}\frac{I[g]}{2\pi}\right]\right) = \exp(i\pi d\Xi). \quad (3.26)$$

At long last, the one-loop partition function is rendered completely topological:

$$\mathcal{Z} = \exp\left(i\pi d\left[\frac{\eta_g}{2} + \frac{I[g]}{24\pi}\right]\right)\sum_\alpha \exp\left[i\left(k + \frac{c_2(G)}{2}\right)I[A^{(\alpha)}]\right]T_\alpha. \quad (3.27)$$

Framing. The discussion above elided a crucial but technical point: $I[g]$ suffers from a geometrical ambiguity analogous to the failure of gauge invariance in S_{CS}. It is well known [18] that every oriented 3-manifold admits a trivialization of its tangent bundle: such a trivialization is called a *framing*. It is also well known [19] that every oriented 3-manifold admits a *spin structure*. (This assures us that the characterization of ω after (3.25) makes sense.) In fact, every choice of framing uniquely determines a spin structure; this choice influences the geometry of the spinor bundle and therefore affects its Levi–Civita connection. There is no canonical choice of framing for a generic 3-manifold, so the value of $I[g]$ is ill defined unless a framing is chosen. In other words: relative to one choice of spin structure, Levi–Civita connections on other spinor bundles over M will look as though they have torsion, but there is no canonical way to tell which of these connections is "truly" torsion-free.

In the absence of an unambiguous value for $I[g]$, the best we can do is to describe what happens to $I[g]$, and hence to $\mathcal{Z}$, under a change of framing. In fact, the value of $I[g]$ differs between any two framings by 2π times an integer s that measures the number of relative "twists" between them. (The transformation law $I[g] \longrightarrow I[g] + 2\pi s$ is directly analogous to the behavior $S_{\text{CS}} \longrightarrow S_{\text{CS}} + 2\pi kw$ of the CS action under large gauge transformations.) Upon shifting the framing by s units, (3.27) shows that the partition function changes by

$$I[g] \longrightarrow I[g] + 2\pi s \implies \mathcal{Z} \longrightarrow \exp\left(\frac{2\pi i s d}{24}\right)\mathcal{Z}. \quad (3.28)$$

3.4 Wilson Loops and Commentary

An abelian example. The incorporation of Wilson loops in nonabelian CS theory, i.e. the calculation of $\mathcal{Z}(M,L)$ via (3.7), is difficult, so we will postpone it. Instead, we revisit the abelian case discussed in §2.2, with $M = S^3$, $G = \mathrm{U}(1)$, and CS action (2.1). Choose r non-intersecting, oriented, closed curves $\gamma_1, ..., \gamma_r$ whose union form a link L, and integers $n_1, ..., n_r$ that play the role of irreducible representations of U(1). Consider the *Wilson link*

$$W[L] \equiv \prod_{a=1}^{r} W_a[\gamma_a] = \prod_{a=1}^{r} \exp\left(n_a \oint_{\gamma_a} \mathrm{d}x_a^\mu \, A_\mu \right). \tag{3.29}$$

(The path ordering symbol in (3.6) is not necessary here because the theory is abelian.) Generalizing our results in §2.2, the expectation value of $W[L]$ may be written

$$\begin{aligned} \big\langle W[L] \big\rangle \equiv \frac{1}{\mathcal{Z}} \int \mathcal{D}A \, W[L] \, e^{iS_{\mathrm{CS}}[A]} &= \exp\left(\frac{i}{2k} \sum_{a,b=1}^{r} n_a n_b \oint_{\gamma_a} \mathrm{d}x_a^\mu \oint_{\gamma_b} \mathrm{d}x_b^\nu \, \varepsilon_{\mu\nu\rho} \frac{(x_a - x_b)^\rho}{|x_a - x_b|^3} \right) = \\ &= \exp\left(\frac{2\pi i}{k} \sum_{a,b=1}^{r} n_a n_b \, \Phi[\gamma_a, \gamma_b] \right), \end{aligned} \tag{3.30}$$

where x_a^μ and x_b^ν are local coordinates for a region of S^3 containing L, and where $\Phi[\gamma_a, \gamma_b]$ is the usual Gauß linking number of γ_a and γ_b, defined by (2.19).

The self-linking number. If $a = b$, the *self-linking number* $\Phi_{\mathrm{s}}[\gamma_a] \equiv \Phi[\gamma_a, \gamma_a]$ diverges. We resolved this issue in §2.2 by simply ignoring such self-interaction terms. But there is a much more elegant way of regulating these infinities. Along each knot $\gamma \subset L$, we choose a vector field everywhere orthogonal to γ. Such a choice is called a *framing* of γ, echoing the framing of 3-manifolds we developed[7] in §3.3. We then "thicken" γ into a ribbon by extending or displacing γ slightly along this vector field, as shown in Fig. 3. The ribbon is bounded by γ and a new knot γ', and we define the self-linking number $\Phi_{\mathrm{s}}[\gamma]$ to be $\Phi[\gamma, \gamma']$.

Change of framing. This prescription clearly depends on the topological class of the vector field used to extend γ, and it generically makes $\Phi_{\mathrm{s}}[\gamma]$ nonzero. In S^3, every knot actually has a canonical framing obtained by requiring its self-linking number to vanish, and this is the one we chose in §2.2. But this is not so in general: knots in arbitrary 3-manifolds do not admit a canonical choice of framing. Just as we did for $\mathcal{Z}$, we will content ourselves here with a law that governs how $\Phi_{\mathrm{s}}[\gamma]$ and $\langle W[L] \rangle$ transform under a change of framing. Fortunately, the index theorem guarantees that the self-linking number differs between two different framings by an integer number t of twists, as shown in Fig. 3. When the framing

[7] This notion of framing is equivalent to a choice of trivialization for the tangent bundle $T\gamma$ of the knot, and therefore defines a framing (in our original sense of the word) of γ as a 1-manifold.

of a knot γ_a is shifted by t units, (3.30) tells us that $\langle W[L]\rangle$ transforms by

$$\Phi_{\rm s}[\gamma_a] \longrightarrow \Phi_{\rm s}[\gamma_a] + t \implies \big\langle W[L]\big\rangle \longrightarrow \exp\left(\frac{2\pi i t n_a^2}{k}\right)\big\langle W[L]\big\rangle. \tag{3.31}$$

Mathematically, (3.28) and (3.31) say that $\mathcal{Z}$ and $\langle W[L]\rangle$ are topological invariants of *framed* manifolds. Physically, they describe a framing anomaly, due to the gravitational coupling.

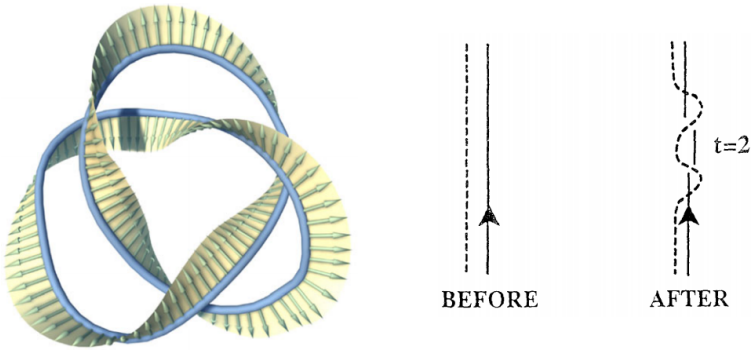

Figure 3: Left: a choice of framing for the trefoil knot is an orthogonal vector field that "frames" it [20]. Right: a change of framing can be viewed as a twist of γ' relative to γ [1].

Examples and caveats. Let us close with an assortment of comments, examples, and warnings that we failed to give in the course of developing the theory. We begin with some explicit partition functions. For $M = S^2 \times S^1$, one has $\mathcal{Z} = 1$ for any gauge group G. For $M = S^3$, the gauge group $G = \mathrm{U}(1)$ makes available the result (3.30); taking the empty link, the phase is zero and so $\mathcal{Z} = 1$. For $M = S^3$ and nonabelian G, the only flat connection is the trivial one, $A^{(\alpha)} = 0$. One might think to apply (3.27), but (as we discuss below) it actually breaks down here. The actual result and its large-k scaling for $G = \mathrm{SU}(2)$ are

$$\mathcal{Z} = \sqrt{\frac{2}{k+2}}\,\sin\left(\frac{\pi}{k+2}\right) \sim k^{-3/2}. \tag{3.32}$$

Notice that $c_2(\mathrm{SU}(2)) = 4$, so the factor $k+2$ that appears here is really the "renormalized" CS level $k + \frac{c_2(G)}{2}$ at one loop. Separately, it is somewhat amusing to note that the gauge group $G = \mathrm{SU}(5)$ has $d = \dim G = 24$, so that when this is substituted into (3.28), the dependence of $\mathcal{Z}$ on the framing of M disappears completely.

Finally, we note an important caveat in §3.2: the determinants of the operators Δ and L_- must be nonzero in order for any of the consequences of (3.17) to make sense. In fact, these operators are nonsingular if and only if $A^{(\alpha)}$ trivializes all of the the de Rham cohomology groups $H^n_{\rm dR}(M, E)$. These objects arise from the observation that on a flat bundle, $[D_\mu, D_\nu] = F_{\mu\nu} = 0$ implies that $D \circ D = 0$. The covariant derivative is therefore a coboundary operator twisted by the flat connection $A^{(\alpha)}$, and we define the cohomology by $H^n_{\rm dR}(M, E) = \big(\ker D\big|_{\Omega^n(M,E)}\big)\big/\big(\operatorname{im} D\big|_{\Omega^{n-1}(M,E)}\big)$. The easiest way to make our results fail

is to find a situation where $H^0(M, E) \neq 0$. If this happens, the operators Δ and L_- will have zero eigenvalues, so the auxiliary fields ϕ, $\bar{c}$, and c will have zero modes and the gauge fixing becomes more complicated. This is what happens when $M = S^3$, where the unique flat connection is trivial. To wit, we find that $H^0(M, E) = \ker \mathrm{d}\big|_{C^\infty(M,E)} = E \neq 0$ is the set of constant maps $M \longrightarrow E$. So despite its "triviality," the trivial connection causes big problems. Finally, we observe that if $H^1_{\mathrm{dR}}(M, E) \neq 0$, then flat connections are not isolated or finite in number, but instead lie in a larger moduli space of flat connections.

4 Hilbert Space Structure

Into the fray. In this section, we turn away from path integrals and take up the ambitious goal of solving Chern–Simons theory exactly. To "solve" the theory will mean to describe, as precisely as possible, its Hilbert space and the structure of its physical observables. Our strategy, following Witten's insight [1], will be to (1) cut the 3-manifold M on which the theory lives into simpler pieces, (2) solve the theory on each piece by canonical quantization, (3) include Wilson loops, and then (4) glue the pieces back together. The rest of the present section will be devoted to fleshing out some of the details of this plan. The ideas developed in §2.3 should be kept close at hand throughout the rather abstract discussion that follows. Nevertheless, we will be rather impressionistic, and will leave out some technical details.

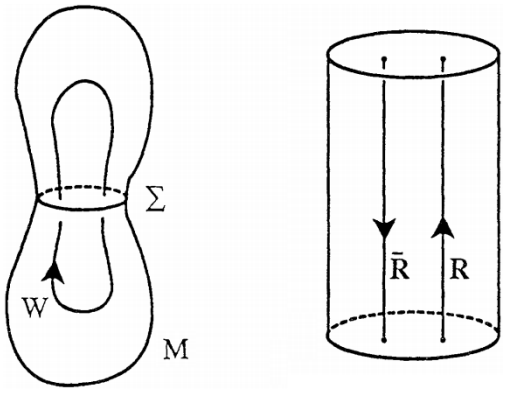

Figure 4: Left: a 3-manifold M containing a Wilson loop W, cut along a Riemann surface Σ. Right: a collar neighborhood of the cut. The surface Σ has two marked points representing charges, where the outgoing and ingoing parts of W insert representations R and $\overline{R}$ [1].

Synopsis. Every oriented 3-manifold may be cut into two disjoint pieces along a 2D surface Σ, which—as an oriented 2-manifold—may be given a complex structure and thus becomes a Riemann surface. (The fact that Σ does not have a *canonical* complex structure will become important to us later.) Near the cut, M has the topology $\Sigma \times \mathbb{R}$; the picture is that of slicing a block of swiss cheese. We thus consider CS theory on manifolds of this topology, and by treating $\mathbb{R}$ as the time direction, we will obtain a Hilbert space $\mathcal{H}_\Sigma$ associated to Σ. One surprise we will meet along the way is that the Gauß law constraint, which is now nonlinear,

renders $\mathcal{H}_\Sigma$ finite-dimensional. Morally speaking, this happens because $\mathcal{H}_\Sigma$ is obtained from the classical phase space of the theory, which is parametrized by finitely many phases due to the nontrivial Wilson loops allowed by the topology of Σ. We can also include sources, as before, by putting knots in M. Some of these knots may pass through Σ, producing the picture of Fig. 4. In this case, Σ comes to us with "marked" points p_i where it is pierced, each attached to the representation R_i given to the corresponding Wilson loop. The construction of $\mathcal{H}_\Sigma$ from this data is more involved, so we will only sketch how it is done.

4.1 Holomorphic Quantization

The canonical formalism. We begin with the CS action (3.3) on $M = \Sigma \times \mathbb{R}$. In the gauge $A_0 = 0$, our results are almost identical to what we found in §2.3, the only difference being the presence of group-theoretic indices. The action reduces to

$$S_{\text{CS}} = \frac{k}{4\pi} \int_{\mathbb{R}} \mathrm{d}t \int_\Sigma \mathrm{d}^2x\, \varepsilon^{ij} \operatorname{Tr}\Big(A_i \dot{A}_j\Big). \tag{4.1}$$

As before, the Hamiltonian vanishes, the "coordinates" are A_i, and their "momenta" are $\varepsilon^{ij} A_j$. So as before, the gauge field is canonically conjugate to itself. The canonical Poisson brackets obeyed by the spatial components of $A_\mu = (0, A_i^a T^a)$ are

$$\big\{A_i^a(\mathbf{x}), A_j^b(\mathbf{y})\big\} = \frac{2\pi}{k} \varepsilon_{ij} \delta^{ab} \delta^{(2)}(\mathbf{x} - \mathbf{y}). \tag{4.2}$$

This relation defines a classical phase space, but it is too large: the classical CS theory is also constrained by the nonabelian Gauß law, which restricts us to the flat connections:

$$\varepsilon^{ij} F_{ij}^a = 0 \iff F = 0. \tag{4.3}$$

Plan of attack. Since $F = \mathrm{d}A + A \wedge A$ is quadratic in the gauge field, the constraint (4.3) re-introduces nonlinearity into the eminently free theory defined by (4.1–4.2). It would be virtually impossible to follow the the "quantize, then constrain" procedure laid out in §2.3: implementing (4.3) at the level of wave functionals, as in (2.26), would lead to a badly nonlinear PDE constraining the physically allowed states. Instead, we will "constrain, then quantize" by implementing (4.3) at the classical level. The resulting phase space $\mathcal{M}$ is called the *moduli space of flat connections*; it consists of gauge equivalence classes of flat principal G-connections A on the gauge bundle $E \longrightarrow \Sigma$. It was studied by Atiyah and Bott [21], who showed that $\mathcal{M}$ is a compact manifold-like object with mild singularities, inherits a symplectic structure from the unconstrained phase space of (4.2), and has (real) dimension $d(2g-2)$, where $d = \dim G$ and g is the genus of Σ. Now, since $\mathcal{M}$ is compact, it has finite (symplectic) volume. In accordance with Heisenberg's uncertainty principle, quantization typically produces one quantum degree of freedom per unit volume in phase space. So we see that quantization of $\mathcal{M}$ should yield a finite-dimensional Hilbert space.

Geometric quantization. In quantum mechanics [22], one usually begins with a phase space $\mathcal{P}$ parametrized by coordinates q^i and momenta p^i, and then defines the Hilbert space $\mathcal{H}$ as the space of complex-valued, square-integrable functions of the q^i. The salient features of this construction are that it requires a decomposition of the phase space variables into coordinates and momenta, and that it "chooses" half of these coordinates. (Such a choice is called a *polarization*.) We can recast this construction in geometric terms: one attaches at every point of $\mathcal{P}$ a copy of the complex plane, and then defines a wave function ψ as any section of the resulting $\mathbb{C}$-bundle (We have not yet specified the bundle, but we will do so shortly.) that obeys two conditions. First, ψ is square-integrable; and second, it depends only on the q^i and not on the p^i, so that $\frac{\partial \psi}{\partial p^i} = 0$. This viewpoint is admittedly a bit awkward to spell out—but purposefully so, as we shall now watch it get dramatically more powerful.

Kähler quantization. A *Kähler manifold* is a manifold that is at once symplectic, complex, and Riemannian, with all three structures compatible.[8] If the phase space $\mathcal{P}$ happens to be Kähler, then we can coordinatize it by new variables $z^i = q^i + ip^i$ and $\overline{z}^i = q^i - ip^i$. The $\mathbb{C}$-bundle above becomes a complex line bundle L over $\mathcal{P}$, and we can choose our wave functions by admitting those sections $\psi \colon \mathcal{P} \longrightarrow L$ which depend on the z^i but not on $\overline{z}^i$. Thus we define the Hilbert space to be the space of *holomorphic sections* of L. Such sections have $\frac{\partial \psi}{\partial \overline{z}^i} = 0$, so they are annihilated by the $\overline{\partial}$ *operator* on L. As for "square integrability," one needs an inner product. This is obtained by endowing L with a *hermitian metric*, which allows one to compute inner products like $\langle \psi | \psi \rangle$. The hermitian metric determines a unique connection (in fact, the Levi–Civita connection!) on L, called the *Chern connection* $\mathcal{A}$, which is compatible with the Kähler structure on $\mathcal{P}$ in two ways. First, the covariant derivative $\mathcal{D}$ built from $\mathcal{A}$ is identical to the $\overline{\partial}$ operator, so physical wave functions are covariantly constant: $\mathcal{D}\psi = 0$. And second, the curvature Ω of $\mathcal{A}$ satisfies $\Omega = -i\omega$, where ω is (!) the symplectic form on $\mathcal{P}$. And now, a miracle: the *first Chern class* of a line bundle, $c_1(L) = [\frac{i}{2\pi}\Omega]$, is a topological invariant that characterizes L uniquely. Thus as soon as the symplectic structure on $\mathcal{P}$ is given, the necessary line bundle is uniquely determined.

So, to summarize: if $\mathcal{P}$ is a Kähler manifold with symplectic form ω, consider the line bundle L with first Chern class $c_1(L) = [\frac{\omega}{2\pi}]$, henceforth called the *prequantum line bundle* of $\mathcal{P}$. Endow L with the Chern connection, which determines a hermitian inner product on L and has curvature $\Omega = -i\omega$. Then the Hilbert space $\mathcal{H}$ associated to $\mathcal{P}$ is the space of holomorphic sections of L; and if $\mathcal{P}$ is compact, then $\mathcal{H}$ will be finite-dimensional.

The Hilbert space. In the case at hand, the phase space $\mathcal{P} = \mathcal{M}$ does not carry a natural complex or Kähler structure. However, it can be shown that upon choosing a complex structure J on the surface Σ, the moduli space $\mathcal{M} = \mathcal{M}_J$ also inherits complex and Kähler structures. Roughly, this proceeds by complexifying the gauge group G to $G_{\mathbb{C}}$, and then

[8] It is, in some sense, a phase space (since it is symplectic) that looks like a curvy Hilbert space (since it is complex and has a tangent-space inner product). More prosaically, a Kähler manifold is a fancy donut.

(echoing the holomorphic quantization above) choosing the "slice" through $G_{\mathbb{C}}$ that forces connections on the complexified gauge bundle to vary holomorphically. Once this choice is made, the symplectic form ω inherited by $\mathcal{M}_J$ from (4.2) defines the unique line bundle L with first Chern class $c_1(L) = [\frac{\omega}{2\pi}]$. The Chern connection on L is then used to construct a covariant derivative $\mathcal{D}$ that agrees with the $\overline{\partial}$ operator on L, and this latter operator defines the holomorphic sections which—at long last—associate to Σ a Hilbert space $\mathcal{H}_\Sigma^{(J)}$. This discussion is admittedly rather abstract, but we will see some examples in §4.3.

A slightly more concrete description of L is available when $G = \mathrm{SU}(N)$ and $k = 1$: the basic idea is to construct L from a family of operators arising from the $\overline{\partial}$ operator on Σ itself. Recall that in $A_0 = 0$ gauge, the gauge bundle $E \longrightarrow M$ may be considered as a bundle over Σ. Once we fix a complex structure on Σ, complexify the gauge group to $G_{\mathbb{C}}$, and choose a representation of $G_{\mathbb{C}}$ (say, the fundamental), E attains the structure of a holomorphic vector bundle over Σ. Each point in $\mathcal{M}_J$ is a flat connection on E, and the covariant derivative built from this connection retains some information about the complex structure on E. In fact, we can use it to "twist" the $\overline{\partial}$ operator that already lives on Σ thanks to its own complex structure. In this way, $\mathcal{M}_J$ parametrizes a family of twisted $\overline{\partial}$ operators on E.

The determinant bundle. There is now a nice way to associate a (complex) line to each twisted $\overline{\partial}$ operator—that is, to each point of $\mathcal{M}_J$—and thereby obtain the prequantum line bundle. The construction was introduced by Quillen [23] and essentially proceeds by "taking the determinant" of $\overline{\partial}$. One thinks of the determinant as a stand-in for the operation of taking the top exterior power of E, which gives a one-dimensional space regarded as a line. More precisely, one attaches to each $\overline{\partial}$ operator the line

$$\mathcal{L}_{\overline{\partial}} = \bigwedge^n \big(\ker \overline{\partial}\big)^* \otimes \bigwedge^n \big(\operatorname{coker} \overline{\partial}\big). \tag{4.4}$$

Finally, gluing these lines together defines the *determinant line bundle* L over $\mathcal{M}_J$.

It can be checked that, when $k = 1$, this is the correct line bundle. The Dirac determinant provides a hermitian metric on L and determines the Chern connection, which enables one to compute its curvature and first Chern class. This was done by Quillen, who showed that they both agree with the symplectic form (4.2) via $c_1(L) = [\frac{\omega}{2\pi}]$ and $\Omega = -i\omega$. When $k \neq 1$, the determinant bundle L gets the symplectic form wrong by a factor of k; in general, the correct line bundle at level k is $L^{\otimes k}$. And when $G \neq \mathrm{SU}(N)$, then it is some (potentially large) tensor power of $L^{\otimes k}$ that appears instead, thanks to the Kodaira embedding theorem.

The modular functor. Let us turn to what is by now a recurring theme: there is no canonical choice of complex structure on Σ. In fact, such complex structures J vary within their own *moduli space of Riemann surfaces*, which we denote $\mathcal{M}_\Sigma$. Now, on physical grounds the Hilbert space $\mathcal{H}_\Sigma^{(J)}$ constructed above *cannot* depend on the complex structure J. This is because $\mathcal{H}_\Sigma^{(J)}$ is the solution to a problem whose formulation depends on an oriented surface Σ, but not on its complex structure. Changing J should not affect the Hilbert space of a

theory that has no right to "know" about it. We can therefore think of the spaces $\mathcal{H}_\Sigma^{(J)}$ as the fibers of a vector bundle over the moduli space $\mathcal{M}_\Sigma$, whose points are the complex structures J on Σ. Because the fibers are all the same, this bundle should admit a flat connection that canonically identifies them. This allows us to write $\mathcal{H}_\Sigma$ in place of $\mathcal{H}_\Sigma^{(J)}$, and we obtain a map sending the surface Σ to its CS Hilbert space $\mathcal{H}_\Sigma$. These flat bundles on moduli space were studied by Segal [24], who called the map $\Sigma \mapsto \mathcal{H}_\Sigma$ the *modular functor*.

4.2 Quantization with Sources

Inserting Wilson loops. We are now ready to do something of real, physical interest: we shall describe the Hilbert space structure of CS theory in the presence of matter. Equivalently, as we saw in §2.2, we can couple $S_{\rm CS}$ to sources by considering the theory in the presence of Wilson loops. This is illustrated by the Fig. 4, which suggests that the imprint left by a Wilson loop on Σ is a set of static, nonabelian charges. In $A_0 = 0$ gauge, the Gauß law of CS theory in the presence of r such charges at points $p_1, ..., p_r \in \Sigma$ becomes

$$\frac{k}{4\pi}\varepsilon^{ij}F_{ij}^a = \sum_{s=1}^{r}\delta^{(2)}(x-p_s)T_s^a. \tag{4.5}$$

Here the $T_s^a \in \mathfrak{g}$ are Lie algebra generators assigned to each charge. In the quantum theory, these generators are to be considered in the appropriate representations R_s.

How *not* to quantize. The quantization of (4.5) seems like a daunting task. The naïve approach, which is to quantize the Poisson brackets (4.2) and then impose the sourced Gauß law at the quantum level, is even more impractical now than it was in the source-free case. To impose the constraint at the classical level seems like the only way out, but even this runs into immediate difficulties: for one thing, the connections defined by (4.5) are definitely not flat. It is no longer obvious how to put a symplectic structure on the space of classical solutions, and more generally the failure of flatness causes large swaths of our geometrical machinery to collapse. Even more worryingly, the generators T^a that appear in (4.5) do not commute. The "connections" they define therefore cannot be ordinary principal connections, but must instead be some bizarre noncommutative objects. These musings serve as a natural starting point for noncommutative geometry and the theory of quantum groups [25], both of which have natural formulations and solutions to the problem posed above. We will not entertain this approach, however. Instead, we take a more cavalier perspective: observe that the connections defined by (4.5) really *are* flat, except at the marked points $p_1, ..., p_r$, where they have nonabelian delta-function "defects." It stands to reason that once we learn how to deal with these static charges, we might somehow modify or augment the moduli space of flat connections to include their effects. For this, we need to take a slight detour.

Borel–Weil–Bott. There is an astonishingly simple idea at the heart of physics: given any classical system with symmetry group G, the corresponding quantum Hilbert space should be a unitary irreducible representation R of G.[9] In fact, the correspondence goes both ways: every such R is the Hilbert space that quantizes some phase space whose symplectic form ω_R is invariant under G. Let us give a very rough sketch of how this works. The key insight, due to Kirillov [26], is that all of the ingredients for the quantization can be found "natively" in G. Let T be a maximal torus in G, and consider the quotient $\mathcal{P} = G/T$. This space, called a *flag manifold*, admits a symplectic structure ω_R for each unitary irreducible representation R of G. It can be checked that ω_R is G-invariant, so $\mathcal{P}$ really is the phase space of our classical system. We then proceed with Kähler quantization. First, we complexify G to put a complex structure on $\mathcal{P} = G/T$, which thereby becomes Kähler. Next, we consider the prequantum line bundle L of $\mathcal{P}$ by choosing the unique one with $c_1(L) = [\frac{\omega_R}{2\pi}]$. And finally, we take holomorphic sections of L to obtain the Hilbert space $\mathcal{H}$. In this setting, the Borel–Weil–Bott theorem guarantees [27] that $\mathcal{H}$ is actually isomorphic to R itself. And conversely, every unitary irreducible representation R of G comes from quantizing G/T.

Quantization at last. Now we shall use the Borel–Weil–Bott mechanism to outmaneuver the problems laid out above. The key is to think of the T_s^a appearing in (4.5) as *quantum* objects: not only do they not commute, but they should genuinely be regarded as operators in the representations R_s attached to each marked point on Σ. To implement (4.5) classically, we must "de-quantize" the T_s^a, as follows. At each marked point, we place a copy of the flag manifold G/T, and give it the symplectic structure ω_{R_s} corresponding to the representation R_s that lives there. The Borel–Weil–Bott theorem allows us to view R_s as the Hilbert space that quantizes the phase space $(G/T, \omega_{R_s})$, so we can replace the quantum operator T_s^a with the unique phase-space function on G/T that maps to T_s^a under quantization. (Recall that quantization sends phase-space functions to quantum observables, and each T_s^a—being anti-hermitian—is exactly i times an observable.) This replacement renders the constraint (4.5) completely classical; its effect is to augment the moduli space of flat connections by several copies of G/T. More precisely, our new phase space $\widetilde{\mathcal{M}}$ consists of flat principal G-connections on Σ that suffer a reduction of structure group to T at the marked points. (This reduction implements the additional G-symmetry carried by the Wilson loops.) Finally, $\widetilde{\mathcal{M}}$ can be given a Kähler structure and quantized in exactly the same way as before.

4.3 Example: Genus Zero

The Riemann sphere. The discussion so far has been rather dense, so let us describe a special case more concretely. We will take up the case where Σ has genus zero: it has the topology of S^2, potentially with marked points $p_1, ..., p_r$ that can be thought of as punctures.

[9] Rotational invariance manifests spin via representations of SU(2) by the Pauli matrices; translational invariance gives the familiar Hilbert space $L^2(\mathbb{R}^n)$; Poincaré invariance yields the QFT Fock space; etc.

By the Riemann uniformization theorem, (the unmarked) Σ has a unique complex structure, which we adopt without question, in which it is biholomorphic to the Riemann sphere $\mathbb{P}^1$.

Let us first study the Hilbert space of pure CS theory on $\mathbb{P}^1$ without any Wilson loops. By the hairy ball theorem (or by Poincaré–Hopf), the 2-sphere admits no nonwhere-vanishing vector fields. This implies that there are no flat connections on $\mathbb{P}^1$, nor on any principal G-bundle over it. Therefore the moduli space of flat connections in genus zero is empty. The quantization is "trivial," in the sense that the empty set is automatically Kähler, and any line bundle over the empty set is also empty. To find the space of holomorphic sections, one must consider maps $\mathcal{M} = \emptyset \longrightarrow \emptyset = L$. Now, it is a fact of elementary set theory that there is a unique function from the empty set to any other set, so the empty line bundle actually has a single section. This section is trivially holomorphic, so we find that the Hilbert space in genus zero is one-dimensional and consists of a single physical state. In fact, this is precisely the state whose Gaussian wave functional (2.27) we obtained in the U(1) theory.

Remarks on marking. When marked points are introduced, the analysis above breaks down: this is because the r-punctured Riemann sphere allows flat connections for $r \geq 1$, or equivalently because the augmented moduli space $\widetilde{\mathcal{M}}$ is no longer empty. Naïvely, one might try to obtain the Hilbert space of the theory by just tensoring together the Hilbert spaces R_s of each source charge. But this tensor product flouts the conservation of charge, which intuitively requires the "total charge" on Σ to be zero. One might think to force each representation R to come paired with its conjugate $\overline{R}$, as Fig. 4 might suggest. But this condition is too strict: there may be other ways for all of the representations R_s to combine to "cancel out" the charge. Thus we require only that the R_s must collectively couple to the trivial representation. To that end, we expand the tensor product of the R_s as a direct sum of all of the irreducibles of G, among them the trivial representation. This is called *fusion*:

$$\bigotimes_{s=1}^{r} R_s = \left(\bigoplus_{i=1}^{n} \mathbf{1} \right) \oplus \Big(\cdots \Big) \equiv \mathcal{H} \oplus \Big(\cdots \Big). \tag{4.6}$$

The subspace $\mathcal{H} = \mathbf{1}^{\oplus n}$ of fixed points of G is precisely the sector of $\otimes_{s=1}^{r} R_s$ with "zero charge," and is the correct Hilbert space for the Riemann sphere with marked points.

One, two, three. Let us apply the formula above to explicitly compute the CS Hilbert spaces of CS theory on the Riemann sphere with only a few marked points.

For one marked point with representation R, the decomposition above is just $R = R$. If R is the trivial representation, then $\mathcal{H} = \mathbf{1}$ is one-dimensional. If $R \neq \mathbf{1}$, then there is no way to satisfy charge conservation, so the Hilbert space is zero-dimensional (i.e. trivial).

For two marked points with representations R_1 and R_2, the only way to satisfy charge conservation is to require $R_2 = \overline{R}_1$. That is, $R_1 \otimes R_2$ contains one copy of $\mathbf{1}$ if $R_2 = \overline{R}_1$, in which case once again $\mathcal{H} = \mathbf{1}$ is one-dimensional. And if $R_2 \neq R_1$, then $\mathcal{H}$ is trivial.

For three marked points with representations R_i, R_j, and R_k, things are more interesting.

The decomposition of $R_i \otimes R_j \otimes R_k$ contains a number N_{ijk} of $\mathbf{1}$s; this number gives the dimension of the Hilbert space $\mathcal{H} = \mathbf{1}^{\oplus N_{ijk}}$ of CS theory on the thrice-marked $\mathbb{P}^1$. A formula for N_{ijk} was given by Verlinde [28] and studied extensively by Moore and Seiberg [29, 30].

4.4 Outlook: Surgery and Sources

Gluing. In our discussion of canonical quantization, we left out the final step of gluing together our results on different pieces of the original 3-manifold M. To get a flavor of how this works, suppose that M can be split into two disjoint pieces, M_1 and M_2, by cutting along an embedded 2-sphere. Both M_1 and M_2 have an S^2 boundary, and therefore (in the absence of Wilson loops) both have 1-dimensional CS Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ living at their boundaries. The S^2 boundaries have opposite orientations, so $\mathcal{H}_1$ and $\mathcal{H}_2$ are canonically dual. Suitable boundary conditions can be chosen so that the path integral on M_1 "prepares" a state $|\psi\rangle \in \mathcal{H}_1$, and similarly the path integral on M_2 prepares $|\chi\rangle \in \mathcal{H}_2$. Since M is the connected sum of M_1 and M_2, the partition function of CS theory on M is the inner product $\mathcal{Z}(M) = \langle\chi|\psi\rangle$. Next, we carry out the same procedure on S^3, which splits along its equator into two disjoint 3-balls B_1 and B_2. We find that $\mathcal{Z}(S^3) = \langle v'|v\rangle$, where $|v\rangle \in \mathcal{H}_1$ and $|v'\rangle \in \mathcal{H}_2$ are the states prepared by the CS path integrals on B_1 and B_2, respectively. But now—and this is the crucial point—since $\mathcal{H}_1$ and $\mathcal{H}_2$ are 1-dimensional, $|v\rangle$ and $|\psi\rangle$ must be scalar multiples of each other, and so must $|v'\rangle$ and $|\chi\rangle$, in such a way that

$$\langle\chi|\psi\rangle \cdot \langle v'|v\rangle = \langle\chi|v\rangle \cdot \langle v'|\psi\rangle \iff \mathcal{Z}(M)\,\mathcal{Z}(S^3) = \mathcal{Z}(M_1)\,\mathcal{Z}(M_2). \tag{4.7}$$

This formula can be understood pictorially by Fig. 5, which shows that S^3 can be taken apart and used to "cap off" the boundaries that are generated when M is sliced open.

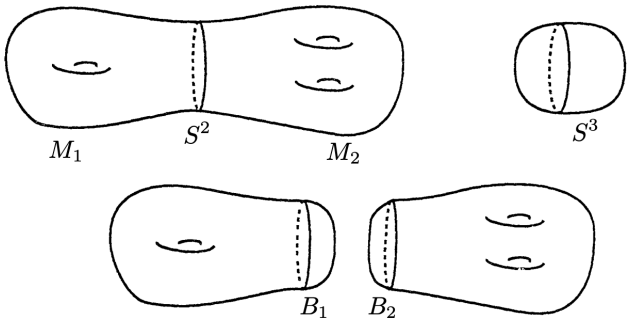

Figure 5: Disassembling $M = M_1 \# M_2$ along S^2, using the caps of S^3 to round off the pieces.

Surgery. One can generalize the preceding discussion to include Wilson loops; this enlarges the Hilbert space dimensions and makes the linear algebra more involved—one meets *skein relations* and so on—but the general story is much the same. A byproduct of this exercise is that one learns how to explicitly compute expectation values of Wilson loops in S^3. (This is a charge we fell short of answering in §3.4.) With S^3 expectation values in hand, one dreams of

computing expectation values on other 3-manifolds $M \neq S^3$. This is where *surgery* comes in: upon specifying a knot γ in an arbitrary 3-manifold M, one can thicken γ into a tube T with the topology of a solid torus, excise T from M, apply a cleverly chosen diffeomorphism to the torus boundary $\partial T = T^2$, and then glue the pieces back together to obtain a new 3-manifold $\widetilde{M}$. It is then a foundational result of 3-dimensional topology that every 3-manifold can be obtained from any other by a finite number of surgeries on embedded knots. In particular, if one pays close attention to what surgery does to the expectation values of Wilson loops in S^3, the aforementioned dream will be realized. This is precisely what Witten does.

In fact, the argument is nearly identical to the one above. When the solid torus T is removed from $M = S^3$, both T and the knot complement $N = S^3 \setminus T$ have boundaries that harbor canonically dual Hilbert spaces $\mathcal{H}_{T^2}$ and $\mathcal{H}_{\partial N} = \mathcal{H}^*_{T^2}$. The path integrals on T and N yield vectors $|\psi\rangle \in \mathcal{H}_{T^2}$ and $|\chi\rangle \in \mathcal{H}_{\partial N}$, respectively, and the diffeomorphism on the torus boundary $\partial T = T^2$ induces a linear transformation K on $\mathcal{H}_{T^2}$ that sends $|\psi\rangle \mapsto K\,|\psi\rangle$. The path integral on the new manifold $\widetilde{M}$ is therefore related to $\mathcal{Z}(M) = \langle\chi|\psi\rangle$ by the operator K: $\mathcal{Z}(\widetilde{M}) = \langle\chi|K|\psi\rangle$. Witten then describes the CS Hilbert space in genus one based on the work of Verlinde [28], performs a modular transformation on T^2 to obtain S^3 from $S^2 \times S^1$ from surgery, and uses the genus-one Hilbert space to deduce the formula (3.32) for the CS partition function on S^3, after which the whole procedure is repeated with sources.

5 Tying Up Loose Ends

A unified view. In this review, we have tried to untangle some of the topology and geometry lurking behind Chern–Simons theory. We began our journey in §2 with an invitation to the U(1) theory, perhaps from the perspective of an experimentalist encountering its bizarre physics for the first time. This phenomenological excursion through the abelian theory prepared us for a more formal treatment of the nonabelian theory by path integrals (in §3) and by canonical quantization (in §4). Along the way, we deliberately refrained from discussing many aspects of the broader physical and mathematical world in which CS theory is situated. Below, we will give one last overview of the narrow route we have taken through the theory; we will then conclude by giving brief glimpses into several parts of that broader world.

5.1 Broad Recapitulation

The abelian theory. In §2, we discovered that CS theory is topological. Its simplest physical observables are anyonic exchange phases that measure the winding of magnetic flux lines attached to electric charges that roam the plane. We soon realized these phases as expectation values of Wilson loops, which are most naturally treated using path integrals. The pure CS path integral showed us that, in the presence of magnetic flux, the CS level k must be quantized as an integer to preserve the gauge invariance of the quantum theory.

Meanwhile, the CS path integral with sources revealed that anyonic phases really cap-

ture topological invariants—linking numbers—of the knots traced out by the Wilson loops attached to the sources. After this, we treated the U(1) theory in the canonical formalism, where we found that the Hamiltonian vanishes identically, and that the components of the CS gauge field are canonically conjugate to each other. We solved the theory exactly in $\mathbb{R}^2$, finding just a single physical state, discussed its finite-dimensional algebra of observables on T^2, and concluded with several fun but important topological remarks.

Path integrals. After setting up the machinery of nonabelian gauge theory, we introduced and quantized the nonabelian CS theory by examining the failure of its gauge invariance under gauge transformations with nonzero winding number. We then attempted to evaluate the pure CS path integral at weak coupling. The process yielded up a host of topological invariants of the underlying 3-manifold M, but nevertheless seemed to depend on its metric. We managed to remove this dependence, but at the cost of a counterterm that depends on the framing of M. In the end, we found an explicit, topological formula for the partition function of pure CS theory, together with a law describing its behavior under change of framing. The path integral with Wilson loops met a similar topological fate: to cure the divergences associated with the self-linking number of a Wilson loop, we had to choose a framing for the knot. In close analogy to the source-free theory, we obtained a formula describing the transformation of Wilson loop expectation values under change of framing.

Canonical quantization. To perform a canonical analysis of nonabelian CS theory, we chopped a generic 3-manifold M into pieces that locally resemble $\Sigma \times \mathbb{R}$, with Σ a Riemann surface. The classical phase space of CS theory on Σ is the moduli space of flat connections on Σ, and to quantize it we developed the powerful technique of Kähler quantization. This technique gave us a description of the pure CS Hilbert space as the space of holomorphic sections of a certain line bundle over $\mathcal{M}$. We also considered the inclusion of Wilson loops, which pierce Σ and leave nonabelian charges on its surface at distinguished or marked points, echoing the magnetic flux lines of the abelian theory. These loops initially seemed to present unresolvable problems, even for the classical description of the theory. However, by appealing to the Borel–Weil–Bott theorem and the philosophy of quantizing systems with symmetries, we were able to account for the effects of nonabelian charges by tweaking our moduli space. In the specific case where Σ is a surface of genus zero, much more concrete results were available, including a complete description of the Hilbert space and its dimension.

5.2 Extensions and Connections

The Jones polynomial. One of the principal contributions of Witten's seminal paper [1] was to understand an important knot invariant called the *Jones polynomial* in terms of Wilson loops in CS theory. To formulate some of his results, we begin by observing that the quantity $e^{2\pi i/k}$ has been ubiquitous throughout this review. We give it the name q, and we claim that it is the right variable (from the standpoint of knot theory) in which to write

down our final results. For complex values of k, the variable $q = e^{2\pi i/k}$ is closely related to the *nome* $q = e^{2\pi i\tau}$ that appears in the theory of elliptic functions.

It is no accident that both elliptic functions and the Jones polynomial V_q exhibit special properties when $k = 1/\tau$, the inverse CS coupling, assumes integer values: this is precisely when the results of quantum CS theory become available. Actually, this is really only correct "at tree level," as it were: the correct variable, for $G = \mathrm{SU}(N)$, is really $q = \exp\bigl(\frac{2\pi i}{k+N}\bigr)$. In terms of the nome, we can finally write down the expectation values of some Wilson loops in the nonabelian CS theory. For example, a single unknotted Wilson loop $W[\bigcirc]$ in S^3, in the fundamental representation of $\mathrm{SU}(N)$, has vacuum expectation value

$$\bigl\langle W[\bigcirc]\bigr\rangle = \frac{q^{N/2} - q^{-N/2}}{q^{1/2} - q^{-1/2}} \overset{N=2}{\rightsquigarrow} q^{1/2} + q^{-1/2} = \frac{\sin\left[2\pi/(k+2)\right]}{\sin\left[\pi/(k+2)\right]}. \tag{5.1}$$

This expression gives the correct (trivial!) Jones polynomial of the unknot after substitution into the famous *skein relation*, also derived by Witten using Hilbert space methods:

$$-qV_q(\bigcirc) + \bigl(q^{1/2} - q^{-1/2}\bigr)\bigl\langle W\bigr\rangle + q^{-1}V_q(\bigcirc) = 0 \implies V_q(\bigcirc) = 1. \tag{5.2}$$

In fact, some elementary properties of the Jones polynomial were already apparent from the discussion in §4.4. For instance, the factorization result (4.7) may be written

$$\frac{\mathcal{Z}(M)}{\mathcal{Z}(S^3)} = \frac{\mathcal{Z}(M_1)}{\mathcal{Z}(S^3)} \cdot \frac{\mathcal{Z}(M_2)}{\mathcal{Z}(S^3)}. \tag{5.3}$$

When Wilson loops are included, the ratios that appear above become Jones polynomials; the formula then expresses their multiplicative behavior under connected sums.

Relation to 2D CFT. Thus far, we have deliberately avoided discussing the deep connections between CS theory and 2-dimensional conformal field theory. A full treatment of these connections is beyond the scope of this review, but let us collect here a few suggestive remarks. Our first hint comes from our path-integral calculation of the partition function of pure CS theory in §3.2. The main difference between the results at leading order (3.10) and at one loop (3.27), aside from an increase in complexity, is that the relative phases making up the partition function are shifted, as if we had replaced k by $k + \frac{c_2(G)}{2}$:

$$\exp\bigl(ikI[A^{(\alpha)}]\bigr) \rightsquigarrow \exp\left[i\left(k + \frac{c_2(G)}{2}\right)I[A^{(\alpha)}]\right]. \tag{5.4}$$

This is evocative of a similar phenomenon in 2D CFT, where many semiclassical results become exact if the parameters in the approximate formulæ are shifted. In fact, when $G = \mathrm{SU}(N)$, k is related to the *central charge* c of a certain 2D CFT by the formula $c = \frac{kd}{k+N}$, where $d = \dim G = N^2 - 1$. When the CS coupling is weak, we have $k \longrightarrow \infty$ and thus $c \longrightarrow d$. It is then no surprise (from the CFT standpoint) that many results

of CS theory derived at large k, like the partition function's framing transformation law (3.28), have generalizations obtained by replacing d by c. A similar story plays out for the framing transformation law (3.31) of Wilson loops; this time, the factor $\frac{n_a^2}{2k}$ is replaced by the conformal weight h of a certain primary field in the CFT. More generally, the concept of framing in CS theory has a direct analog in the monodromies of the corresponding CFT.

Holographic duality. By now it should be clear that CS theory in 3 dimensions is intimately related to conformal field theory in 2 dimensions. We have a great deal of evidence: for one thing, the analogies described above are tantalizing in their own right. Furthermore, in §4 we caught a glimpse of how the Hilbert spaces and partition functions of CS theory are obtained by slicing 3-manifolds along Riemann surfaces and paying careful attention to the physics on these lower-dimensional boundaries. The constructions we met there—moduli spaces of Riemann surfaces and of flat connections, holomorphic vector bundles, monodromies, the fusion of representations—are all important ingredients in 2D CFT. Among these, perhaps the most important is Segal's modular functor [24]. In Segal's original formulation, a Riemann surface Σ defines the spacetime on which a 2D CFT lives; to this spacetime is assigned the vector space $\mathcal{H}_\Sigma$ of solutions to the conformal Ward identities for descendants of the identity operator. Segal's construction of this "space of conformal blocks" is completely identical to our construction of the CS Hilbert space on $\Sigma \times \mathbb{R}$, and this firmly establishes the equivalence of the CS and CFT theories. Indeed, it can be checked that our results for the dimensions of CS Hilbert spaces with sources agree with the number of conformal blocks in 2D CFT on the corresponding marked Riemann surfaces.

The picture whose vague outlines are now coming into view is that of a hologram: CS theory on a 3-manifold M is equivalent, or dual, to a particular 2D CFT living on a Riemann surface Σ that can be regarded as the "boundary" of M. (In our discussion of gluing and surgery, this interpretation was imposed forcibly.) The corresponding CFT is called the *Wess–Zumino–Witten* (WZW) model, and its action closely resembles the boundary term (3.4) in the variation of the CS action. In this setting, the marked points where CS Wilson lines pierce Σ represent operator insertions of primary fields in the WZW model that transform in the representations attached to the Wilson lines. The CS–WZW correspondence is an actual theorem, and can be viewed as a proto-example of the holographic principle.

3D gravity. One important element that keeps CS–WZW from being a real example of holography is the absence of gravitational physics on the CS side of the correspondence. Nevertheless, Witten showed [31] that 3-dimensional gravity *is* a Chern–Simons theory!

The Einstein field equations of general relativity in $2+1$ dimensions do not admit gravitational wave solutions. The theory has, therefore, no local propagating degrees of freedom; for this reason, it is purely topological. The Einstein–Hilbert action that governs this theory has a first-order formulation in terms of vielbein fields and spin connections, and Witten showed that it can be recast as a CS action of the form (3.3). The "gauge fields" in 3D gravity take values in the Lie algebra of the Poincaré group $\mathrm{ISO}(2,1)$ or some other noncompact variant

thereof, depending on the sign of the cosmological constant. (For de Sitter backgrounds, the group is $SO(3,1)$; for anti-de Sitter, it is $SO(2,2)$.) Since the gauge group is no longer compact, much of the theory of compact Lie groups is no longer applicable, and the theory becomes more difficult to interpret. In spite of this difficulty, Witten argued [32] that 3D gravity admits a well defined quantization, analogous to the quantization of CS theory that we have seen here. This claim establishes the CS–WZW correspondence as a true incarnation of the AdS_3/CFT_2 duality. Witten also argued that the theory is exactly solvable, both classically and quantum-mechanically. Even so, much remains to be understood about this duality and its implications for quantum gravity.

References

[1] E. Witten, "Quantum Field Theory and the Jones Polynomial," Commun. Math. Phys. **121**, 351–399 (1989) doi:10.1007/BF01217730

[2] G. V. Dunne, "Aspects of Chern-Simons theory," [arXiv:hep-th/9902115 [hep-th]]

[3] D. Tong, "Gauge Theory," ch. 8, `damtp.cam.ac.uk/user/tong/gaugetheory/gt.pdf`

[4] D. Tong, "Lectures on the Quantum Hall Effect," ch. 5, [arXiv:1606.06687 [hep-th]]

[5] S.-S. Chern and J. Simons, "Characteristic forms and geometric invariants," Ann. Math. **99** (1), 48–69 (1974) doi:10.2307/1971013

[6] A. M. Polyakov, "Fermi–Bose Transmutations Induced by Gauge Fields," Mod. Phys. Lett. A **3**, 325 (1988) doi:10.1142/S0217732388000398

[7] R. L. Ricca and B. Nipoti, "Gauss' Linking Number Revisited," J. Knot Theory Ramif., **20** (10), 1325–1343 (2011) doi:10.1142/S0218216511009261

[8] A. Achucarro and P. K. Townsend, "A Chern-Simons Action for Three-Dimensional anti-De Sitter Supergravity Theories," Phys. Lett. B **180**, 89 (1986) doi:10.1016/0370-2693(86)90140-1

[9] N. Steenrod. *The Topology of Fibre Bundles (PMS-14)*, Princeton University Press (1951) JSTOR, `www.jstor.org/stable/j.ctt1bpm9t5`

[10] M. Nakahara, "Geometry, Topology and Physics," ch. 9–10

[11] S. Coleman, "Aspects of Symmetry: Selected Erice Lectures," §7.3.2 doi:10.1017/CBO9780511565045

[12] R. Bott, "An Application of the Morse Theory to the Topology of Lie Groups," Bulletin de la S.M.F. **84**, 251–281 (1956), doi:10.24033/bsmf.1472, `http://www.numdam.org/articles/10.24033/bsmf.1472`

[13] D. Bar-Natan and E. Witten, "Perturbative expansion of Chern-Simons theory with noncompact gauge group," Commun. Math. Phys. **141**, 423–440 (1991) doi:10.1007/BF02101513

[14] A. S. Schwarz, "The Partition Function of Degenerate Quadratic Functional and Ray-Singer Invariants," Lett. Math. Phys. **2**, 247–252 (1978) doi:10.1007/BF00406412

[15] M. F. Atiyah, V. K. Patodi, and I. M. Singer, "Spectral asymmetry and Riemannian Geometry 1," Math. Proc. Cambridge Phil. Soc. **77**, 43 (1975) doi:10.1017/S0305004100049410

[16] M. F. Atiyah, V. K. Patodi, and I. M. Singer, "Spectral asymmetry and Riemannian geometry 2," Math. Proc. Cambridge Phil. Soc. **78**, 405 (1976) doi:10.1017/S0305004100051872

[17] M. F. Atiyah, V. K. Patodi, and I. M. Singer, "Spectral asymmetry and Riemannian geometry 3," Math. Proc. Cambridge Phil. Soc. **79**, 71–99 (1976) doi:10.1017/S0305004100052105

[18] P. E. Parker, "On Some Theorems of Geroch and Stiefel," J. Math. Phys. **25**, 597 (1984) doi:10.1063/1.526209

[19] J. W. Milnor and J. D. Stasheff, *Characteristic Classes* (AM-76) (1974) doi:10.1515/9781400881826

[20] M. Elhamdadi, M. Hajij, and K. Istvan, "Framed Knots," Math. Intell. **42**, 7–22 (2020) `https://link.springer.com/content/pdf/10.1007/s00283-020-09990-0.pdf`

[21] M. F. Atiyah and R. Bott, "The Yang-Mills equations over Riemann surfaces," Phil. Trans. Roy. Soc. Lond. A **308**, 523–615 (1982)

[22] N. Woodhouse, "Geometric Quantization."

[23] D. Quillen, "Determinants of Cauchy–Riemann operators over a Riemann surface," Fuct. Anal. Appl. **19**, 31 (1986)

[24] G. Segal, "Conformal field theory," Oxford preprint; and lecture at the IAMP Congress, Swansea, July, 1988

[25] A. Connes and M. Marcolli, "A Walk in the noncommutative garden," [arXiv:math/0601054 [math.QA]].

[26] A. A. Kirillov, *Lectures on the Orbit Method*, American Mathematical Soc., **64** (2004) doi:10.1090/gsm/064

[27] C. Beasley, "Localization for Wilson Loops in Chern-Simons Theory," Adv. Theor. Math. Phys. **17**, no.1, 1-240 (2013) doi:10.4310/ATMP.2013.v17.n1.a1 [arXiv:0911.2687 [hep-th]].

[28] E. P. Verlinde, "Fusion Rules and Modular Transformations in 2D Conformal Field Theory," Nucl. Phys. B **300**, 360-376 (1988) doi:10.1016/0550-3213(88)90603-7

[29] G. W. Moore and N. Seiberg, "Classical and Quantum Conformal Field Theory," Commun. Math. Phys. **123**, 177 (1989) doi:10.1007/BF01238857

[30] G. W. Moore and N. Seiberg, "Polynomial Equations for Rational Conformal Field Theories," Phys. Lett. B **212**, 451-460 (1988) doi:10.1016/0370-2693(88)91796-0

[31] E. Witten, "(2+1)-Dimensional Gravity as an Exactly Soluble System," Nucl. Phys. B **311**, 46 (1988) doi:10.1016/0550-3213(88)90143-5

[32] E. Witten, "Three-Dimensional Gravity Revisited," [arXiv:0706.3359 [hep-th]].