Quantum Topol. 3 (2012), 1–137
DOI 10.4171/QT/26

Quantum Topology

Fivebranes and knots

Edward Witten[1]

Abstract. We develop an approach to Khovanov homology of knots via gauge theory (previous physics-based approaches involved other descriptions of the relevant spaces of BPS states). The starting point is a system of D3-branes ending on an NS5-brane with a nonzero theta-angle. On the one hand, this system can be related to a Chern–Simons gauge theory on the boundary of the D3-brane world-volume; on the other hand, it can be studied by standard techniques of S-duality and T-duality. Combining the two approaches leads to a new and manifestly invariant description of the Jones polynomial of knots, and its generalizations, and to a manifestly invariant description of Khovanov homology, in terms of certain elliptic partial differential equations in four and five dimensions.

Mathematics Subject Classification (2010). 57R56.

Keywords. Gauge theory, Khovanov homology, Jones polynomial, topological quantum field theory.

Contents

[1]Research supported in part by NSF Grant PHY-0969448.

1. Introduction

1.1. Knot polynomials. The Jones polynomial ([62] and [63]) associates to a knot K in Euclidean three-space $\mathbb{R}^3$ (or in a three-sphere S^3) a Laurent polynomial $\mathcal{J}(q;K)$ in a single variable q. The coefficients in this Laurent polynomial are integers. Some further details are explained below.

The Jones polynomial – and its many generalizations which are also Laurent polynomials with integer coefficients – can be constructed in a variety of ways from two-dimensional mathematical physics. The key ingredients include lattice statistical mechanics, Yang–Baxter equations, conformal field theory, and braid group representations; see [35], [68], [97], [85], [13], and [96]. These constructions are very efficient for computing the knot polynomials, demonstrating their topological invariance, and showing that they indeed are Laurent polynomials with integer coefficients.

However, such constructions do not make manifest the three-dimensional symmetry of the Jones polynomial. For this purpose, three-dimensional quantum gauge theory with a Chern–Simons action ([89], [88], and [24]) turns out to be useful. The Chern–Simons action for a gauge theory with gauge group[1] G and gauge field A on

[1] In this paper, G is always a compact Lie group, and all representations considered are finite-dimensional.

an oriented three-manifold W can be written

$$I = \frac{k}{4\pi}\int_W \mathrm{Tr}\Big(A\wedge \mathrm{d}A + \frac{2}{3}A\wedge A\wedge A\Big). \tag{1.1}$$

Here k is an integer for topological reasons; up to a choice of orientation, one may take k to be positive. In this theory, to an oriented embedded loop $K\subset W$ and a representation R of G, one can associate an observable, the trace of the holonomy or Wilson loop operator:

$$\mathcal{W}(K,R) = \mathrm{Tr}_R P\exp\oint_K A. \tag{1.2}$$

Reversing the orientation of K has the same effect as replacing R by its complex conjugate. It turns out [102] that the Jones polynomial and its generalizations can be computed as expectation values of Wilson loop operators, if we express the argument q of the knot polynomials in terms of the Chern–Simons level k by

$$q = \exp\left(2\pi i/(k+h)\right), \tag{1.3}$$

where h is the dual Coxeter number of G. For example, if we take $G = \mathrm{SU}(2)$, R to be the two-dimensional irreducible representation of SU(2), and $W = S^3$, then the expectation value of $\mathcal{W}(K,R)$ is equal to the Jones polynomial:

$$\mathcal{J}(q;K) = \langle \mathcal{W}(K,R)\rangle. \tag{1.4}$$

1.1.1. Some details. We will spell out a few details about the function $\mathcal{J}(q;K)$. First of all, the definition extends immediately to an oriented link, that is a union L of ν disjoint oriented embedded circles K_i. We label the K_i by representations R_i of G and set

$$J(q;K_i,R_i) = \Big\langle \prod_i \mathcal{W}(K_i,R_i)\Big\rangle. \tag{1.5}$$

For $G = \mathrm{SU}(2)$ and all R_i equal to the two-dimensional representation, this function is known as the Jones polynomial of the link L. We denote this special case as $\mathcal{J}(q;L)$.

In (1.4) and (1.5), the symbol $\langle\ \rangle$ refers to an expectation value, that is, a ratio of two path integrals

$$J(q;K_i,R_i) = \frac{\displaystyle\int DA\ \exp(iI)\prod_i \mathcal{W}(K_i,R_i)}{\displaystyle\int DA\ \exp(iI)}. \tag{1.6}$$

For $W = S^3$, the denominator is non-trivial (for example, it equals $\sqrt{2/(k+2)}\ \sin(\pi/(k+2))$ for $G = \mathrm{SU}(2)$) and it is necessary to divide by this factor to obtain a function $J(q;K_i,R_i)$ that has the simple properties we will explore in this paper.

However, in our framework, it will be more natural to study a path integral rather than a ratio of two path integrals. $J(q; K_i, R_i)$ can be expressed in this form by simply replacing $W = S^3$ with $W = \mathbb{R}^3$. The ratio in (1.6) is unaffected, but now the denominator (regularized by the procedure in the present paper to deal with the behavior at infinity) equals 1 and can be omitted. Taking $W = \mathbb{R}^3$ will also simplify the arguments in this paper by suppressing infrared fluctuations, in a sense that will be clear later, and in certain other technical details. Accordingly, though we will define an analog of Khovanov homology on any three-manifold, its relation to Chern–Simons theory is most simple for the case of links in $\mathbb{R}^3$.

We should warn the reader of a few differences between our conventions and the ones that are most common in the mathematical literature. First, a very basic case of a link is the empty link $\varnothing$ for which the number of embedded circles is $\nu = 0$. With our definition, $J(q; \varnothing) = 1$. In the mathematical literature, it is customary to normalize the Jones polynomial so that its value is 1 for the unknot K_0 rather than the empty link $\varnothing$, so the usual mathematical definition corresponds to what we would call $\widetilde{\mathcal{J}}(q; L) = \mathcal{J}(q; L)/\mathcal{J}(q; K_0)$. An analogous statement holds for the more general invariants $J(q; K_i, R_i)$.

The precise sense in which $J(q; K_i, R_i)$ is a Laurent polynomial is as follows. In general, depending on the representations R_i, $J(q; K_i, R_i))$ is either a Laurent polynomial in q, or $q^{1/2}$ times a Laurent polynomial in q. For example, the Jones polynomial is $q^{\nu/2}$ times a Laurent polynomial,

$$\mathcal{J}(q; L) = \sum_{n \in \mathbb{Z} + \nu/2} a_n q^n, \quad a_n \in \mathbb{Z}. \tag{1.7}$$

The coefficients a_n are integers and all but finitely many of them vanish. The half-integral powers are often suppressed by taking the basic variable to be not our q but $\tilde{q} = q^{1/2}$. In many ways, however, the variable q is more natural. For example, it will turn out to be the natural instanton counting factor in a dual gauge theory description. The fractional powers of q turn out to have a natural topological interpretation, and it seems unnecessary to suppress them. (In a sense, it is also ultimately fruitless to try to suppress them, since as will become clear, on a general three-manifold, we meet general fractional powers of q, not just half-integral powers.)

One further detail is that, as explained via gauge theory in [102], the invariants $J(q; K_i, R_i)$ are most naturally defined for framed links. (A framing of an embedded circle $K \subset W$ is a trivialization of the normal bundle to K in W.) Under a change in framing, $J(q; K_i, R_i)$ is multiplied by a certain (generically fractional) power of q. For links in S^3 or $\mathbb{R}^3$, one can suppress this phenomenon, since an embedded circle in S^3 has a distinguished framing (relative to which its self-linking number is zero). Standard formulas such as (1.7) implicitly refer to this standard framing. Similarly, the Chern–Simons path integral on a general three-manifold W depends naturally on a framing of W (a trivialization of its tangent bundle T) or more generally [6] on a two-framing (a trivialization of $T \oplus T$). A change of framing of W has the same

sort of effect as a change in framing of a link: it multiplies the path integral by a power of q. This power cancels out of the ratio (1.6), but when we assert that the denominator is 1 for $W = \mathbb{R}^3$, this statement refers to the path integral defined with the obvious framing associated to a Euclidean metric on $\mathbb{R}^3$.

1.1.2. What Chern–Simons theory doesn't explain. The Chern–Simons path integral gives a definition of the invariants $J(q; K_i, R_i)$ with manifest three-dimensional symmetry, provided that q is a root of unity of the particular form (1.3). Granted that $J(q; K_i, R_i)$ is a Laurent polynomial, it is determined by its behavior at these values of q. However, the gauge theory path integral does not shed much light on why these functions are Laurent polynomials. This is clearer in any of the definitions of the link invariants based on two-dimensional mathematical physics. The only known way to deduce that $J(q; K_i, R_i)$ is a Laurent polynomial starting from three-dimensional gauge theory is to first reduce to a two-dimensional description, for example via representations of braid groups, in which this fact is clear. The Chern–Simons path integral has been used directly [108] to explain the existence of an analytic continuation of Wilson loop expectation values to complex values of k, but not the fact that the result is a Laurent polynomial.

1.2. Khovanov homology. Moreover, none of the constructions so far mentioned give a really good explanation of why the coefficients a_n of these Laurent polynomials are integers. This has been accomplished in Khovanov homology [69], in which the a_n are interpreted as the dimensions (in a $\mathbb{Z}_2$-graded sense) of finite-dimensional vector spaces. For motivation behind Khovanov homology see [23], [34], and [11], and for an introduction see [7]. In this theory, one associates to a link L in three-space a finite-dimensional vector space $\mathcal{k}(L)$, known as its Khovanov homology. The original construction was adapted to the Jones polynomial – or, if you like, to a link labeled by the two-dimensional representation of SU(2). $\mathcal{k}(L)$ is defined as the cohomology of a differential Q (a differential is simply a linear operator Q obeying $Q^2 = 0$) that acts on a larger vector space $\mathcal{h}(L)$. $\mathcal{k}(L)$ is natural and depends only on L, but there is much arbitrariness in the construction of $\mathcal{h}(L)$. $\mathcal{h}(L)$ is bigraded, with symmetry generators that we will call F and P. Q obeys $[\mathrm{F}, Q] = Q$, $[\mathrm{P}, Q] = 0$; these relations ensure that $\mathcal{k}(L)$ is bigraded,

$$\mathcal{k}(L) = \bigoplus_{m,n} \mathcal{k}^{m,n}(L), \tag{1.8}$$

where m, n are the eigenvalues of F, P. With the usual normalization, m and n take integer values. In our formulation in this paper, m is $\mathbb{Z}$-valued and n takes values in $\mathbb{Z} + \nu/2$ (where ν is the number of components of the link L) or more generally in a certain coset of $\mathbb{Z}$ in $\mathbb{R}$. Despite the nonintegrality of the eigenvalues of P, we will loosely refer to the group generated by F and P as $\mathrm{U}(1) \times \mathrm{U}(1)$ and the associated grading as a $\mathbb{Z} \times \mathbb{Z}$ grading. The relation between Khovanov homology and the Jones

polynomial is

$$\mathcal{J}(q;L) = \mathrm{Tr}_{\mathcal{k}(L)}\,(-1)^{\mathrm{F}} q^{\mathrm{P}}. \tag{1.9}$$

This formula makes manifest the fact that $\mathcal{J}(q;L)$ is a Laurent polynomial with integer coefficients. (The half-integral powers in $\mathcal{J}$ for a link with an odd number of components arise from the fact that, with our normalization, for a link in $\mathbb{R}^3$ with ν components labeled by the two-dimensional representation of SU(2), the eigenvalues of P lie in $\mathbb{Z}+\nu/2$. See Section 5.4.2.) We can describe (1.9) by saying that the Jones polynomial can be recovered from Khovanov homology by taking an equivariant index or Euler characteristic. Since F is $\mathbb{Z}$-valued but the right hand side of (1.9) only depends on the value of F mod 2, this formula also shows that Khovanov homology potentially contains more information than the Jones polynomial. It has turned out that the additional information is really essential.

The success in recovering the Jones polynomial from a homology theory raises the question of whether a similar construction is possible if components of L are labeled by arbitrary representations R_i of a compact Lie group G. In the literature, this has been accomplished for many classes of groups and representations. Here, we will make a general proposal.

From a physical point of view, a three-dimensional quantum field theory with loop operators will naturally assign a *number* – the value of the path integral – to a knot. To associate to a knot a *vector space* (its Khovanov homology) rather than a number, we want a four-dimensional quantum field theory with surface operators rather than loop or line operators. Thus,[2] introduce a fourth "time" dimension, parametrized by $\mathbb{R}$, and consider a four-dimensional topological field theory on $M = \mathbb{R}\times W$, with a surface operator on $\Sigma = \mathbb{R}\times K$; as before, K is a knot in a three-manifold W. The space of physical states in such a theory will be a vector space associated to the pair (W, K); this vector space will be bigraded – like the Khovanov homology of a knot in $W = S^3$ – if the four-dimensional theory has an appropriate $\mathrm{U}(1)\times\mathrm{U}(1)$-symmetry. What has just been described was part of the original motivation that led to Khovanov homology [23] and these matters have also been discussed from a physical point of view [51]. From the point of view of four-dimensional quantum field theory, the index formula (1.9) has a natural interpretation. Given a four-dimensional quantum field theory, one can reduce to a three-dimensional quantum field theory by compactifying on S^1. The partition function of a four-dimensional theory on a four-manifold of the form $M = S^1\times W$, where W is a three-manifold, will give a $\mathbb{Z}_2$-graded trace or index. (Here we assume that if surface operators are present, they are supported on $S^1\times K$, for some $K\subset W$, to be compatible with the product form of M.) In the reduction, if there is a conserved charge P that commutes with Q, one can make a twist by q^{P} (for some q) in going around the circle. The partition function of the reduced theory will then be an equivariant index as in (1.9).

In the mathematical literature, there actually is direct evidence that Khovanov

[2] The actual framework we develop later is more complicated than the idealized sketch offered here, mainly in the need to introduce a fifth dimension.

homology is part of a four-dimensional theory with surface operators. The main evidence comes from consideration of cobordism between knots. Here, we take $M = I \times S^3$, where $I = [0, 1]$ is the unit interval. In M, one considers an embedded two-manifold Σ whose restriction to one boundary $\{0\} \times S^3$ is a knot K, and whose restriction to the other boundary $\{1\} \times S^3$ is a knot K'. Physically, one would expect the path integral on M (with Σ understood as the support of a surface operator) to define a linear transformation from the space of physical states associated to the pair (S^3, K) to the corresponding space for (S^3, K'). Mathematically, it has been found that one can associate to such a cobordism a natural linear transformation Φ_Σ from the Khovanov homology of K to that of K':

$$\Phi_\Sigma \colon \mathit{k}(K) \longrightarrow \mathit{k}(K'). \tag{1.10}$$

If one glues together two knot cobordisms, the corresponding transition amplitudes multiply, just as one would expect physically.

The literature on Khovanov homology provides at least one more clue. In close parallel with the early mathematical constructions of the Jones polynomial and its cousins, mathematical constructions of Khovanov homology and its extensions are frequently based on familiar ingredients in mathematical physics. But these constructions do not make manifest the topological invariance of Khovanov homology, potentially creating an opportunity for physicists. Actually, a number of mathematical constructions of Khovanov homology are based on ways of associating a two-dimensional topological quantum field theory (or at least the category of branes in such a theory) to a two-sphere S^2 with marked points $p_i, i = 1, \dots, n$. A natural interpretation is that these constructions arise by specializing a four-dimensional quantum field theory to four-manifolds of the form $M = \Sigma \times S^2$, where Σ is a Riemann surface and surface operators are supported on the two-manifolds $\Sigma \times p_i$. In one construction, see [70] and [54], the effective theory on Σ seems to be a Landau–Ginzburg B-model (so that the branes are matrix factorizations); in a second construction [18], the effective theory is a B-model with target space a certain Kähler manifold; other approaches, see [90] and [64], are based on A-models. There have also been attempts, see [73] and [74], to make the three- or four-dimensional symmetry of Khovanov homology manifest by extracting it from a special case or analog of Donaldson–Floer theory in four dimensions. This of course is related to $\mathcal{N} = 2$ super-Yang–Mills theory in four dimensions.

1.3. Previous physics-based proposals. Actually, a proposal for a physical construction of Khovanov homology has been made some years ago. An initial clue [103] was that the knot invariants associated to Chern–Simons theory can be regarded as open-string analogs of the usual A-model invariants for closed strings. On the other hand, the topological A-model for either closed or open strings can be embedded in Type IIA superstring theory. For open strings, this embedding plus a hypothesis of a geometric transition in string theory has led to powerful results [46] about Chern–Simons theory. In addition, by considering the strong coupling limit of the Type IIA

model, in which the M-theory circle opens up, closed string A-model amplitudes (or Gromov–Witten invariants) can be fruitfully expressed in terms of Gopakumar–Vafa invariants [47]. The Gopakumar–Vafa invariants are simply the dimensions of certain spaces of BPS states of M-theory membranes, so they are automatically integers, unlike the A-model amplitudes themselves (which in general are rational numbers). Expressing the closed topological string amplitudes in terms of Gopakumar–Vafa invariants is powerful because purely numerical invariants (the Gromov–Witten invariants) are expressed in terms of vector spaces (the spaces of BPS states).

The Gopakumar–Vafa construction has an analog [82] for open strings, expressing A-model observables of open strings in terms of spaces of BPS states in the presence of certain branes. For further developments see [76], [87], and [75]; for a review of many of these topics see [78]. This approach has been extended into a proposal [53] to identify the Khovanov homology for a knot K with the space of BPS states – for an M-theory configuration that depends on the choice of K. A substantial amount of evidence for this proposal was given in [53], in part by using geometric transitions as a tool to compute the spaces of BPS states. Moreover, the proposal implied some new predictions concerning Khovanov homology and has led to a better understanding of some aspects of this subject [30]. The relevant brane constructions have been further studied in [26], [2], and [19]. For an extension of these ideas involving the topological vertex and the Nekrasov partition function for instantons, see [61] and [52].

A related road to a physical interpretation of Khovanov homology has appeared much more recently in a study of supersymmetric line operators in four-dimensional gauge theories with $\mathcal{N} = 2$ supersymmetry [38]. It was shown that such line operators form an "algebra," but with the structure constants being vector spaces rather than numbers. For the case that the four-dimensional theory is obtained by compactifying the six-dimensional $(0, 2)$ model on a Riemann surface C, as analyzed in most detail in [36], the algebra in question is closely related to the usual algebra of multiplication of Wilson loop operators in quantum Chern–Simons theory on C – except that the structure constants in the algebra are replaced by vector spaces. (One can recover the usual loop algebra of Chern–Simons theory by taking a supertrace, as in (1.9), to replace the vector spaces by numbers. This has been pointed out by the authors of [38].) These results should be related to a generalization of Khovanov homology for loops in the three-manifold $\mathbb{R} \times C$ – more precisely for product loops of the form $p \times \ell$, with p a point in $\mathbb{R}$ and ℓ a loop in C.

1.4. The present paper. In this paper, we will re-examine the relation of Khovanov homology to the spaces of BPS states in M-theory, with three primary goals. One goal is to give a gauge theory definition of Khovanov homology (as opposed to a definition that requires a full knowledge of string/M-theory). String theory and branes will be used as clues, but the results can be expressed as a gauge theory construction. A second goal is to give a more transparent – or at least new – explanation in this context of the key property of Khovanov homology: the fact that a supertrace in the space of BPS states gives the path integral of Chern–Simons theory. The last goal is to develop an

effective framework to understand generalizations of Khovanov homology in which one varies the three-manifold W or the boundary conditions or other details. (This program is not actually achieved in the present paper.) Along the way, we will clarify some formal properties of Khovanov homology.

1.4.1. The basic idea. The basic idea behind this paper is simply explained. We would like to apply nonperturbative string theory or field theory dualities to three-dimensional Chern–Simons gauge theory, but there is no obvious way to do this directly. However, it is possible to express the path integral of Chern–Simons theory on a three-manifold W as a path integral of $\mathcal{N} = 4$ super Yang–Mills theory on a half-space $V = W \times \mathbb{R}_+$, where $\mathbb{R}_+$ is the ray or half-line $y \geq 0$. (Knots in W are represented by Wilson operators in the boundary of V.) Once this is done, one can apply standard gauge theory and string theory dualities to the $\mathcal{N} = 4$ path integral on the four-manifold V, leading to a description by a higher-dimensional theory with the desired properties.

The relation of the Chern–Simons path integral on W to the $\mathcal{N} = 4$ path integral on $W \times \mathbb{R}_+$ is one of the main results of [109] (and the basic idea is suggested in the conclusions of [108]). We will give an alternative explanation in this paper, partly to keep the paper self-contained, and partly to emphasize the aspects that we need. In general, in this correspondence, the $\mathcal{N} = 4$ path integral on $V = W \times \mathbb{R}_+$ depends on a boundary condition at $y \to \infty$, and the equivalent Chern–Simons path integral is not the usual one but is a path integral defined with an exotic integration cycle, in a sense described in [108]. However, for the case of links in $\mathbb{R}^3$ or S^3, there is essentially (up to a constant multiple) only one possible integration cycle and the path integral obtained this way is equivalent to the standard one. From the vantage point of the present paper, this is one of the reasons that Khovanov homology is simplest in the case of links in $\mathbb{R}^3$.

In order to relate the $\mathcal{N} = 4$ path integral on $V = W \times \mathbb{R}_+$ to a Chern–Simons path integral on W, we need to use the right boundary condition on the boundary of W. The requisite boundary condition is not exotic. It is simply the boundary condition of the D3–NS5 system of Type IIB superstring theory in the presence of a theta-angle. This boundary condition has been described in [39] and [40].

At this point, all we have done is to restate the problem of Chern–Simons theory in terms of an $\mathcal{N} = 4$ path integral on V. To get something like Khovanov homology, we want to re-express the $\mathcal{N} = 4$ path integral on V as a path integral of some other theory on $V \times S^1$. A path integral on $V \times S^1$ can be written as a trace (or, in the presence of fermions, as a $\mathbb{Z}_2$-graded trace) in a Hilbert space $\mathcal{H}$ associated to quantization on V. Suppose that the path integral on $V \times S^1$ is invariant under a supersymmetry generator Q that obeys $Q^2 = 0$. Then, by a standard argument, the $\mathbb{Z}_2$-graded trace in $\mathcal{H}$ reduces to a $\mathbb{Z}_2$-graded trace in $\mathcal{K}$, the cohomology of Q. (We will write $\mathcal{K}$ for cohomology spaces arising in quantum field theory and k for Khovanov homology; we make this distinction because we do not have a proof that these coincide even in situations where k has been defined.) Our strategy to get

a formula like (1.9) for the Jones polynomial is to first express the Jones polynomial as an $\mathcal{N} = 4$ path integral on $\mathbb{R}^3 \times \mathbb{R}_+$ – with knots represented by Wilson operators at the boundary – and then find a duality to re-express this as a path integral on $\mathbb{R}^3 \times \mathbb{R}_+ \times S^1$.

The most naive way to try to do this fails in an instructive way. We first embed the D3–NS5 system in Type IIB superstring theory on $\mathbb{R}^9 \times S^1$, where the S^1 direction is transverse to the branes. Compactifying one of the transverse directions on a circle does not affect anything that has been said so far. Then we perform a T-duality on the S^1. This replaces S^1 by a dual circle $\widetilde{S}^1$. At first sight, it seems that the T-dual of the D3–NS5 path integral will be a path integral on $\mathbb{R}^3 \times \mathbb{R}_+ \times \widetilde{S}^1$, leading in the desired fashion to a trace. However, in the presence of an NS5-brane wrapped on $\mathbb{R}^6 \times p \subset \mathbb{R}^9 \times S^1$ (here $\mathbb{R}^6$ is linearly embedded in $\mathbb{R}^9$ and p is a point in S^1), T-duality maps us not to $\mathbb{R}^9 \times \widetilde{S}^1$ but ([95] and [49]) to $\mathbb{R}^6 \times \mathrm{TN}$, where TN is a Taub-NUT space. TN is asymptotic at infinity to a twisted $\widetilde{S}^1$ bundle over $\mathbb{R}^3$, but crucially, $\widetilde{S}^1$ shrinks to a point in the interior of S^3. Because of this, the path integral in this T-dual description cannot be interpreted as a trace.

There is a simple way to avoid this difficulty. Before T-duality, we first perform S-duality. S-duality converts the D3–NS5 system to a D3–D5 system. (A system of D3-branes ending on a D5-brane has special properties that were investigated in [25], [81], [16], and [21], and interpreted in field theory language as a boundary condition in $\mathcal{N} = 4$ super Yang–Mills theory in [39].) We embed the D3–D5 system in $\mathbb{R}^9 \times S^1$ and now T-duality simply maps this to a D4–D6 system on $\mathbb{R}^9 \times \widetilde{S}^1$. Now the path integral can be straightforwardly interpreted as a trace and this leads to a formula like (1.9). What plays the role of $\mathcal{K}$ is the cohomology of a certain supercharge Q that is preserved by the construction. (The proper choice of Q depends on details that we have omitted here.) F corresponds to an R-symmetry of the brane configuration, and P is, from the point of view of the D4-brane gauge theory, the Yang–Mills instanton number integrated over $\mathbb{R}^3 \times \mathbb{R}_+$.

Most of these steps have analogs with $\mathbb{R}^3$ replaced by a more general three-manifold W, but in trying to formulate the resulting statements about Chern–Simons theory, one runs into infrared divergences and a need to understand how S-duality acts on the boundary conditions at $y = \infty$. The simplest case other than $\mathbb{R}^3$ is likely to be the case that W is obtained by omitting a point from a rational homology sphere. In this case, projecting the missing point to infinity and taking a metric on W that looks near infinity like the flat metric on $\mathbb{R}^3$, there are no infrared divergences and a close analog of Khovanov homology should exist. One will still have the problem of understanding the action of S-duality on the boundary conditions at $y = \infty$.

1.4.2. Organization of the paper. In Section 2, we describe in more detail, in the context of the D3–NS5 system, the relation of the Chern–Simons path integral in three dimensions to an $\mathcal{N} = 4$ path integral in four dimensions. Then we apply standard dualities to this situation, first S-duality in Section 3 followed by T-duality

(or in gauge theory simply the introduction of a fifth dimension) in Section 4. The first step leads to an essentially new description of knot invariants related to Chern–Simons theory, and the second leads to Khovanov homology. The two operations have different status. S-duality is natural purely as a field theory operation, but T-duality is not and leads to a description by a five-dimensional super Yang–Mills theory that is not ultraviolet complete.

A better and conceptually more satisfying formulation is to base our construction not on five-dimensional super Yang–Mills theory but on its familiar ultraviolet completion in the six-dimensional $(0, 2)$ model (for example, see [106] for a brief introduction). In Section 5, we proceed in this way: we begin with the $(0, 2)$ theory in six dimensions, and work our way down to five, four, and three dimensions. This gives the most economical and logically complete treatment of the topic, and it gives the clearest explanation of a number of questions. The top-down approach of Section 5 certainly could have been the starting point of the present paper. We have chosen instead a bottom-up presentation in which the relation to Chern–Simons theory is made as clear as possible at the outset.

In Section 6, we explore a second brane construction, which in some ways is closer to the setting of [53]. The starting point of the second construction is that Wilson operators of Chern–Simons theory can be expressed as codimension two monodromy defects. The two formulations – via Wilson operators or monodromy defects – are related to two different semiclassical limits of Chern–Simons theory. In one case, one takes the level k to be large while keeping fixed the representations R_i labeling the knots. This is the most direct framework for describing the Jones polynomial, Khovanov homology, and their generalizations. In the other type of semiclassical limit, the monodromies produced by the knots are kept fixed as k becomes large. This second limit is related to the volume conjecture of Chern–Simons theory, which has been reviewed with extensive references in [80] and explored physically in [50] and [108]. The formulation of Chern–Simons theory in terms of monodromy defects can be carried through all the dualities of the present paper, leading to descriptions based on codimension two defects in various dimensions, as we explain briefly in Section 6. This matter certainly merits much closer attention.

We probably should mention here two important puzzles that we will *not* unravel. First, Khovanov homology is explicitly calculable for any given link in $\mathbb{R}^3$, though the requisite calculations may not be easy. Indeed, Khovanov homology was originally defined (see [7] for an accessible account) by an explicit algebraic recipe for computing it, though not one that makes topological invariance manifest. The description in the present paper has the opposite properties: topological invariance is manifest, but computability is not. It would be highly desirable to bridge the gap between the two types of knowledge by deducing a known definition of Khovanov homology from the quantum field theory construction studied here (or its close cousin studied earlier in [53]). To do this requires understanding concretely the solutions of the localization equations presented later; one must understand the four-dimensional version of the equations, presented in (2.56), to understand the Jones polynomial, and the five-

dimensional generalization, presented in (5.36), to understand Khovanov homology. Not much of this is done in the present paper; the only examples of actual solutions of the equations presented here are in Section 3.6. However, since the present paper was written, a reasonable understanding of the four-dimensional equations has been obtained in [41] and this indeed has given a concrete understanding of how the Jones polynomial emerges in the present framework. Some interesting special solutions of the four-dimensional equations have also been analyzed in [58].

Second, our approach here makes some things clearer than has been the case hitherto, but we fail to make contact with one important insight from [53]. We consider each gauge group as a problem in its own right, while in [53], the A theories were treated in a unified way, and this has been generalized to B, C, and D; see [92], [79], and [72].

1.4.3. Comparison to other work. Some relations of the present paper to other work, beyond what has already been cited, are as follows.

Geometric Langlands duality (for a review, see [33]) has a generalization, sometimes called quantum geometric Langlands in the mathematics literature, involving a parameter that was called Ψ in [67]. This generalization has been related to the theory of quantum groups [42], suggesting that geometric Langlands should be related to Chern–Simons theory. Indeed, we show in this paper that if formulated on a four-manifold V of boundary W, the four-dimensional topological field theory associated to geometric Langlands is related to Chern–Simons theory on W, with Ψ as essentially the Chern–Simons level. Khovanov homology has previously been defined [18] using moduli spaces of geometric Hecke transformations, which are vital in geometric Langlands and were interpreted via gauge theory in Sections 9 and 10 of [67].

On an abstract three-manifold W, Chern–Simons gauge theory only makes sense if the level k is an integer. But we show in the present paper that if W is the boundary of a given four-manifold V, and we are willing to accept an answer that depends on V, then a theory with many of the properties of Chern–Simons theory can be formulated as a function of a complex variable k. Moreover, the theory appears to be unitary in Lorentz signature if k is real. All this has a counterpart in contemporary developments in condensed matter physics. Topological insulators and superconductors – see for example [86] for a review – are materials of d dimensions (and therefore $d+1$ spacetime dimensions) that on their $(d-1)$-dimensional surface realize physical phenomena that could never occur in a purely $(d-1)$-dimensional material. The values of d that have been realized experimentally are 3 (a bulk material with a two-dimensional surface) and 2 (a thin film with a one-dimensional edge). The $d=3$ topological insulators are materials that ultimately prove to have a "forbidden" Chern–Simons coupling (for the ordinary electromagnetic field), somewhat like the system we study in the present paper for non-integer k.

Apart from papers already cited, a relation between four-dimensional $\mathcal{N}=4$ super Yang–Mills and three-dimensional Chern–Simons – or at least q-deformed

two-dimensional Yang–Mills – has been described in certain geometries in [1]. And a recent paper dealing with topics relatively close to that of the present paper is [27].

While the present paper was in gestation, it developed that the five-dimensional gauge theory equations that we present in (5.36) have been formulated independently by A. Haydys [57]. Haydys' point of view was roughly to study the A-model with target the moduli space of complex-valued flat connections on a three-manifold. He also presented the two reductions of the equations that are described in Section 5.3.1. Even more recently, the author has become aware of work by M. Kontsevich and Y. Soibelman that may have a bearing on the present topic.

Acknowledgments

I would like to thank A. Ashtekar, M. Aganagic, C. Beasley, S. Cherkis, D. Bar-Natan, R. Cohen, R. Dijkgraaf, D. Gaiotto, J. Gomis, S. Gukov, J. Heckman, L. Hollands, J. Kamnitzer, A. Kapustin, B. Kostant, S. Lewallen, R. Mazzeo, M. Mariño, G. Moore, L. Rozansky, Y. Tachikawa, C. Taubes, C. Vafa, and the members of the Stanford and IAS particle theory groups for their comments and I. Frenkel and D. Bar-Natan for having introduced me to the subject.

2. Chern–Simons from four dimensions

2.1. The D3–NS5 system with a theta-angle. As indicated in Section 1.4.1, our starting point is the D3–NS5 system of Type IIB superstring theory. The local picture is that in Minkowski spacetime $\mathbb{R}^{1,9}$, with coordinates $x^0, \dots, x^9$ (and metric signature $-++\cdots+$), we consider N D3-branes supported at $x^4 = x^5 = \cdots = x^9 = 0$. The D3-branes end on a single NS5-brane that is supported at $x^3 = x^7 = x^8 = x^9 = 0$. In the four-dimensional spacetime parametrized by $x^0, \dots, x^3$, the D3-brane world-volume spans the half-space $x^3 > 0$. The gauge theory of the D3-branes is a $\mathrm{U}(N)$ gauge theory with $\mathcal{N} = 4$ supersymmetry. In this gauge theory, the NS5-brane provides a half-BPS boundary condition, that is, a boundary condition that preserves half of the supersymmetry.

When the gauge theory theta-angle vanishes, this boundary condition is simply Neumann boundary conditions for gauge fields, extended to the rest of the vector multiplet in a supersymmetric fashion. However, the brane construction implies the existence of a more general half-BPS boundary condition even for $\theta \neq 0$. Indeed, Type IIB superstring theory has a complex coupling parameter $\tau = \theta/2\pi + i/g_s$ (θ is the expectation value of a Ramond–Ramond scalar and g_s is the string coupling constant), which in the gauge theory becomes $\tau = \theta/2\pi + 4\pi i/g_{\mathrm{YM}}^2$, with g_{YM} the gauge coupling constant and θ the gauge theory theta-angle. The D3–NS5 system is half-BPS for any value of τ, so from a gauge theory point of view, Neumann boundary conditions must have a half-BPS generalization for $\theta \neq 0$.

This generalization was described in Section 2 of [39] (a more roundabout construction was also presented in [40]). We will summarize the essential points here, referring for more details to [39]. Though the initial motivation is the D3–NS5 system, once the half-BPS boundary condition is expressed in field theory language, it makes sense for any gauge group G, and we will present it that way.

The R-symmetry group of $\mathcal{N} = 4$ super Yang–Mills theory is SO(6) (or actually its spin double cover), acting by rotation of the normal bundle to the D3-brane. The presence of the NS5-brane breaks SO(6) to SO(3) × SO(3), where one factor rotates x^4, x^5, x^6 and the second rotates x^7, x^8, x^9. In [39], the two SO(3)'s are denoted respectively as $\mathrm{SO}(3)_X$ and $\mathrm{SO}(3)_Y$ and the corresponding two sets of scalar fields on the D3-brane were denoted as $\vec{X}$ and $\vec{Y}$. The D3–NS5 boundary condition on $\vec{Y}$ is

$$\vec{Y}| = 0 \tag{2.1}$$

(for any field Φ, its restriction to $x^3 = 0$ will be denoted as $\Phi|$), irrespective of θ, but the other boundary conditions are more subtle.

It is useful to adopt a ten-dimensional notation[3] in which $\mathcal{N} = 4$ super Yang–Mills theory comes by dimensional reduction from ten dimensions and the supersymmetries of the D3-brane transform under SO(1, 9) as a spinor **16** of definite chirality; thus a generator ε of supersymmetry obeys

$$\Gamma_{012\ldots9}\varepsilon = \varepsilon, \tag{2.2}$$

where Γ_I, $I = 0, \ldots, 9$, are the SO(1, 9) gamma matrices. (As usual, a symbol such as $\Gamma_{I_1\ldots I_k}$ denotes the antisymmetrized product of the corresponding gamma matrices.) The D3–NS5 boundary condition is invariant under $\mathbb{U} = \mathrm{SO}(1,2) \times \mathrm{SO}(3)_X \times \mathrm{SO}(3)_Y$, where SO(1, 2) acts on the dimensions x^0, x^1, x^2 common to the two types of brane. Each factor in $\mathbb{U}$ has a two-dimensional representation that we denote as **2**, and the **16** transforms as two copies of the tensor product $(\mathbf{2}, \mathbf{2}, \mathbf{2})$. This tensor product, which we denote as V_8, is a real representation of $\mathbb{U}$ of dimension 8. The supersymmetries transform as $\mathbf{16} = \mathsf{V}_8 \otimes \mathsf{V}_2$, where V_2 is a two-dimensional real vector space. The natural operators that act on V_2 are the even elements of the SO(1, 9) Clifford algebra that commute with $\mathbb{U}$. They are generated by

$$B_0 = \Gamma_{456789}, \quad B_1 = \Gamma_{3456}, \quad B_2 = \Gamma_{3789}, \tag{2.3}$$

and in view of the algebraic relations they obey (such as $B_0^2 = -1$, $B_0B_1 + B_1B_0 = 0$, etc.), we can choose a basis for V_2 in which

$$B_0 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{2.4}$$

[3] We will attempt to follow conventions of [67]. In particular, adjoint-valued fields such as gauge fields are real and anti-hermitian. (This accounts for some minus signs in formulas such as (2.10).) We define the Levi-Civita tensor $\epsilon^{\mu\nu\alpha\beta}$ of $\mathbb{R}^{1,3}$ and the corresponding tensor $\epsilon^{\mu\nu\lambda}$ of the hyperplane $x^3 = 0$ as antisymmetric tensors obeying $\epsilon^{0123} = 1 = -\epsilon_{0123}$ and $\epsilon^{012} = 1 = -\epsilon_{012}$, respectively.

The expression $(\varepsilon, \tilde{\varepsilon}) = \bar{\varepsilon}\Gamma_3\tilde{\varepsilon}$ defines an $\mathrm{SO}(1,2)\times\mathrm{SO}(6)$-invariant bilinear form on the **16** of $\mathrm{SO}(1,9)$; it factors as the tensor product of an antisymmetric $\mathbb{U}$-invariant form on V_8 and an antisymmetric form on V_2. If we write $\varepsilon_0 \in \mathsf{V}_2$ as a column vector $\binom{s}{t}$ and $\bar{\varepsilon}_0$ as the row vector $(t, -s)$, then we can write the antisymmetric inner product on V_2 as $\langle\varepsilon_0, \tilde{\varepsilon}_0\rangle = \bar{\varepsilon}_0\tilde{\epsilon}_0$.

In any half-BPS boundary condition that is $\mathbb{U}$-invariant, the unbroken supersymmetries must be precisely those of the form $\mathsf{V}_8 \otimes \varepsilon_0$, for some nonzero vector $\varepsilon_0 \in \mathsf{V}_2$. Since scaling of ε_0 is immaterial, the choice of ε_0 depends essentially on a single real parameter. We can take

$$\varepsilon_0 = \begin{pmatrix} -a \\ 1 \end{pmatrix}, \quad \bar{\varepsilon}_0 = (1\ a) \tag{2.5}$$

(we include the possibility $a = \infty$, which means that the bottom component of ε_0 vanishes). It is shown in [39] that for every $a \in \mathbb{R} \cup \infty$ there is a unique $\mathbb{U}$-invariant half-BPS boundary condition that preserves all of the gauge symmetry. The parameter a corresponds to the gauge theory theta-angle.[4]

Without repeating the full derivation, we will cite the results that we need. The fermion fields λ of $\mathcal{N} = 4$ super Yang–Mills are adjoint-valued fields that transform as the **16** of $\mathrm{SO}(1,9)$, like the supersymmetry generators. The boundary conditions they obey turn out to be

$$\lambda| \in \mathsf{V}_8 \otimes \vartheta, \tag{2.6}$$

where $\vartheta \in \mathsf{V}_2$ is

$$\vartheta = \begin{pmatrix} a \\ 1 \end{pmatrix}. \tag{2.7}$$

The boundary conditions on $\vec{X}$ at $x^3 = 0$ are

$$D_3 X_c - \frac{a}{1+a^2}\epsilon_{cde}[X_d, X_e] = 0, \tag{2.8}$$

and the boundary conditions on the gauge fields at $x^3 = 0$ are

$$F_{3\mu} + \frac{a}{1-a^2}\epsilon_{\mu\nu\lambda}F^{\nu\lambda} = 0. \tag{2.9}$$

At $a = 0$ and $a = \infty$, equations (2.8) and (2.9) reduce to the more obvious Neumann boundary conditions $D_3 X_a = F_{3\mu} = 0$ (the two choices actually correspond to the D3–NS5 and D3–$\overline{\text{NS5}}$ systems). The additional terms in the boundary conditions for generic a reflect boundary corrections to the familiar $\mathcal{N} = 4$ super Yang–Mills action in bulk. Let us first consider $\vec{X}$. The usual bulk action for $\vec{X}$ is in Lorentz signature

$$I_{\vec{X}} = \frac{1}{g_{\mathsf{YM}}^2}\int_{x^3\geq 0} \mathrm{d}^4x \sum_{\mu=0}^{3}\sum_{c=1}^{3} \mathrm{Tr}\, D_\mu X_c D^\mu X_c. \tag{2.10}$$

[4] In the context of the D3–NS5 system, θ is not really an angle as a shift $\theta \to \theta + 2\pi$ would convert the NS5-brane to a $(1,1)$ fivebrane. Accordingly, the following formulas have no periodicity.

Let us consider when happens when we vary $\vec{X}$. If we place no restriction on the value of δX_c at $x^3 = 0$, we will learn that to make the boundary term in the variation of $I_{\vec{X}}$ vanish, the boundary condition must be $D_3 X_c = 0$. Suppose, however, that there is an additional boundary coupling

$$\widetilde{I}_{\vec{X}} = \frac{2a}{3g_{\mathrm{YM}}^2(1+a^2)} \int_{x^3=0} \mathrm{d}^3x\, \epsilon^{cde} \mathrm{Tr} X_c [X_d, X_e]. \tag{2.11}$$

If we now vary $\widehat{I}_{\vec{X}} = I_{\vec{X}} + \widetilde{I}_{\vec{X}}$ with respect to $\vec{X}$, placing again no restriction on $\delta X_c|$, we find that setting the boundary variation of $\widehat{I}_{\vec{X}}$ to zero gives the boundary condition (2.8). So the boundary coupling (2.11) underlies the boundary condition (2.8).

The boundary coupling $\widetilde{I}_{\vec{X}}$ is unfamiliar, but it has a more familiar analog for gauge fields. The analog of (2.10) for the gauge field A, whose field strength we denote as $F_{\mu\nu}$, is

$$I_A = \frac{1}{2g_{\mathrm{YM}}^2} \int_{x^3>0} \mathrm{d}^4x \sum_{\mu,\nu=0}^{3} \mathrm{Tr}\, F_{\mu\nu} F^{\mu\nu}. \tag{2.12}$$

If we work just with this action, then setting its boundary variation to zero (with no restriction on $\delta A|$), we learn that the boundary condition on the gauge field must be $F_{3\mu}| = 0$. To arrive at (2.9), we need an additional term in the action. This extra term is the usual topological term of four-dimensional gauge theory

$$\widetilde{I}_A = -\frac{\theta}{32\pi^2} \int_{x^3\geq 0} \mathrm{d}^4x\, \epsilon^{\mu\nu\alpha\beta}\, \mathrm{Tr}\, F_{\mu\nu} F_{\alpha\beta}, \tag{2.13}$$

with

$$\frac{\theta}{2\pi} = \frac{2a}{1-a^2} \frac{4\pi}{g_{\mathrm{YM}}^2}. \tag{2.14}$$

Viewed as an equation for a with θ, g_{YM} fixed, (2.14) has two roots. The two roots correspond to half-BPS boundary conditions of the D3–NS5 and D3–$\overline{\mathrm{NS5}}$ systems, respectively.

Although written as a bulk integral, $\widetilde{I}_A$ has only a boundary variation, simply because on a manifold V without boundary, $\int_V \mathrm{Tr}\, F \wedge F$ is a topological invariant. In fact, we can almost write $\widetilde{I}_A$ as a boundary integral, the integral over the surface $x^3 = 0$ of the Chern–Simons form:

$$\widetilde{I}_A = -\frac{\theta}{8\pi^2} \int_{x^3=0} \mathrm{d}^3x\, \epsilon^{\mu\nu\lambda}\, \mathrm{Tr}\left(A_\mu \partial_\nu A_\lambda + \frac{2}{3} A_\mu A_\nu A_\lambda\right). \tag{2.15}$$

But there is a problem with this last formula: the Chern–Simons integral on a three-manifold is not quite gauge-invariant. The right hand side of (2.15) is gauge-invariant modulo an integer multiple of θ. Since the action of a quantum theory must be well defined modulo $2\pi\mathbb{Z}$, $\widetilde{I}_A$ would not make sense as the action of a purely

three-dimensional theory unless θ is an integer multiple of 2π. This case is not trivial, since in the presence of an NS5-brane, there is no symmetry of shifting θ by 2π; a shift $\theta \to \theta + 2\pi k$ would convert the NS5-brane to a $(1,k)$ fivebrane. However, we do not wish to be limited to the case $\theta \in 2\pi\mathbb{Z}$. The reason that we are not so restricted is that we are not doing gauge theory on an abstract three-manifold; rather, the three-manifold at $x^3 = 0$ on which we do the integral (2.15) is the boundary of a four-manifold $x^3 \geq 0$ on which the gauge theory is defined; the precise, gauge-invariant definition of $\widetilde{I}_A$ is the original four-dimensional integral (2.13). Still, it can be convenient to informally write $\widetilde{I}_A$ as a Chern–Simons integral (2.15), and we will sometimes do so.

2.1.1. Wick rotation. So far, our formulas have been in Lorentz signature, to make contact with [39] and to emphasize the fact that, as long as the parameter a is real, our boundary condition is unitary and physically sensible. However, to make contact with topological field theory in the rest of this paper, it is helpful to write the formulas analogous to the above in Euclidean signature. A Wick rotation $x^0 \to -ix^0$ reverses the sign[5] of $\widetilde{I}_X$, and multiplies $\widetilde{I}_A$ by $-i$. So in Euclidean signature, combining the terms involving X and A, the boundary interactions of the D3–NS5 system are

$$I^* = \frac{1}{g_{\mathrm{YM}}^2}\int_{x^3=0} \mathrm{d}^3x\Big(-\frac{2a}{3(1+a^2)}\epsilon^{abc}\mathrm{Tr}\, X_a[X_b,X_c] + i\frac{2a}{1-a^2}\epsilon^{\mu\nu\lambda}\mathrm{Tr}\left(A_\mu\partial_\nu A_\lambda + \frac{2}{3}A_\mu A_\nu A_\lambda\right)\Big). \tag{2.16}$$

In a convenient notation in which $\mathcal{N} = 4$ super Yang–Mills is obtained by dimensional reduction from ten dimensions, with the ten dimensions labeled by $x^0,\dots,x^9$, the Euclidean signature version of the chirality condition for supersymmetry generators and fermions is

$$\Gamma_0\Gamma_1\dots\Gamma_9\varepsilon = -i\varepsilon, \quad \Gamma_0\Gamma_1\dots\Gamma_9\lambda = -i\lambda. \tag{2.17}$$

2.2. Comparison to topological field theory. So far we have emphasized the half-BPS nature of the boundary condition of interest. We will also need to understand this boundary condition from the vantage point of topological field theory. The background necessary for this analysis can be found in Section 3 of [67], to which

[5] $\widetilde{I}_X$ is free of derivatives and is a contribution to the potential energy $\mathbb{V}$ of the theory. As usual, $\mathbb{V}$ appears in the Lorentz signature action with a minus sign and in the Euclidean signature action with a plus sign. Concretely, a contribution $\Delta I_L = -\int \mathrm{d}t\, \mathbb{V}$ to the Lorentz signature action I_L leads in the path integral to a factor $\exp(i\,\Delta I_L) = \exp(-i\int \mathrm{d}t\, \mathbb{V})$. After Wick rotation $t \to -it$, this becomes $\exp(-\int \mathrm{d}t\, \mathbb{V})$, which is interpreted as a factor in $\exp(-I_E)$, where I_E is the Euclidean action. So the contribution to I_E is $+\int \mathrm{d}t\, \mathbb{V}$. In the case of the Chern–Simons function, as it is a topological invariant, it is not affected directly by the Wick rotation. The coefficient with which it appears in the action acquires a factor of $-i$ under Wick rotation purely because of the convention that the integrand of the path integral is $\exp(iI_L)$ in Lorentz signature and $\exp(-I_E)$ in Euclidean signature.

we refer for detail (some aspects were treated originally in [111]). Here we will just summarize some necessary facts.

2.2.1. Twisting. The basic idea is to construct a four-dimensional topological field theory by twisting of $\mathcal{N} = 4$ super Yang–Mills theory. Postponing the consideration of possible boundary conditions, we consider $\mathcal{N} = 4$ super Yang–Mills realized on a system of D3-branes parametrized by $x^0, \dots, x^3$. The usual rotation group (in Euclidean signature) is SO(4), rotating these coordinates, while the normal directions $x^4, \dots, x^9$ are rotated by the SO(6) group of R-symmetries. To define a topological field theory, one defines a group $\mathrm{SO}'(4)$ that acts by rotating $x^0, \dots, x^3$ in the usual way, while simultaneously rotating four normal coordinates $x^4, \dots, x^7$. We pick a supersymmetry generator ε that is $\mathrm{SO}'(4)$-invariant, meaning that it obeys

$$(\Gamma_{\mu\nu} + \Gamma_{4+\mu,4+\nu})\varepsilon = 0, \quad \mu, \nu = 0, \dots, 3. \tag{2.18}$$

Denoting as Q the supersymmetry generated by such an ε, arguments of a standard type show that upon restricting to Q-invariant operators and states, one obtains a four-dimensional topological field theory.

From the point of view of $\mathrm{SO}'(4)$-symmetry, four of the adjoint-valued scalar fields of $\mathcal{N} = 4$ super Yang–Mills theory are reinterpreted as an adjoint-valued one-form $\varphi = \sum_{\mu=0}^{3} \varphi_\mu \, \mathrm{d}x^\mu$, while the other two combine to an adjoint-valued complex scalar field σ. $\mathrm{SO}'(4)$ commutes with a group $\mathrm{SO}(2) \cong \mathrm{U}(1)$ of R-symmetries that rotates x^8 and x^9. We normalize its generator F so that σ has charge 2.

This decomposition of the R-symmetry group and of the scalar fields of $\mathcal{N} = 4$ super Yang–Mills theory differs from that made in Section 2.1. In that discussion, the x^μ, $\mu = 0, \dots, 3$, were split in tangential coordinates with $\mu \leq 2$ and a normal coordinate x^3. In matching the two descriptions, we identify the tangential part of φ, that is $\vec{\varphi} = \sum_{\mu=0}^{2} \varphi_\mu \, \mathrm{d}x^\mu$, with $\vec{X}$, and we identify the normal part φ_3 with a component of $\vec{Y}$, say Y_1. (We also set $\sigma = Y_2 - iY_3$.) The boundary couplings (2.16) become in this notation

$$\begin{aligned} I^* = \frac{1}{g_{\mathrm{YM}}^2} \int_{x^3=0} \mathrm{d}^3 x \, \epsilon^{\mu\nu\lambda} \mathrm{Tr} \Big(& -\frac{4a}{3(1+a^2)} \varphi_\mu \varphi_\nu \varphi_\lambda \\ & + i \frac{2a}{1-a^2} \Big(A_\mu \partial_\nu A_\lambda + \frac{2}{3} A_\mu A_\nu A_\lambda \Big) \Big). \end{aligned} \tag{2.19}$$

2.2.2. Comparing the two descriptions. However, rewriting (2.16) in topological field theory notation is only a reasonable thing to do if the boundary condition that leads to (2.16) actually preserves the symmetry of the topological field theory. So let us explain why this is true.

First of all, the condition (2.18) for $\mathrm{SO}'(4)$-invariance of the supersymmetry generator actually has a two-dimensional space of solutions. It is possible to pick a basis

of solutions ε_ℓ, ε_r that are chiral in the four-dimensional sense,

$$\Gamma_{0123}\varepsilon_\ell = -\varepsilon_\ell, \quad \Gamma_{0123}\varepsilon_r = \varepsilon_r. \tag{2.20}$$

It is possible to normalize ε_ℓ and ε_r so that,[6] for $\mu = 0, 1, 2$, or 3,

$$\Gamma_{\mu,4+\mu}\varepsilon_\ell = -\varepsilon_r, \quad \Gamma_{\mu,4+\mu}\varepsilon_r = \varepsilon_\ell. \tag{2.21}$$

In constructing a topological field theory, we may take the supersymmetry generator ε to be an arbitrary linear combination of ε_ℓ and ε_r. Up to an inessential scaling, we take

$$\varepsilon = \varepsilon_\ell + t\,\varepsilon_r. \tag{2.22}$$

(We allow $t = \infty$, which corresponds up to scaling to $\varepsilon = \varepsilon_r$.)

So we get a family of topological field theories parametrized by a complex variable t. Now we can make contact with the D3–NS5 system. From (2.17), (2.20), and (2.3), we have

$$B_0\varepsilon_\ell = i\,\varepsilon_\ell, \quad B_0\varepsilon_r = -i\,\varepsilon_r. \tag{2.23}$$

Using also (2.21) and (2.18), one can show, with some gamma matrix algebra, that

$$B_1\varepsilon_\ell = -\varepsilon_r, \quad B_1\varepsilon_r = -\varepsilon_\ell. \tag{2.24}$$

It follows that

$$\Big(1 + i\,\frac{1-t^2}{1+t^2}B_0 + \frac{2t}{1+t^2}B_1\Big)(\varepsilon_\ell + t\,\varepsilon_r) = 0. \tag{2.25}$$

On the other hand, with the help of (2.4), we see that the object ε_0 defined in (2.5) obeys the same equation

$$\Big(1 + i\,\frac{1-t^2}{1+t^2}B_0 + \frac{2t}{1+t^2}B_1\Big)\varepsilon_0 = 0 \tag{2.26}$$

if and only if the parameter a used in describing the D3–NS5 system is related to the parameter t of the topological field theory by

$$a = i\,\frac{1-it}{1+it}. \tag{2.27}$$

The half-BPS boundary condition of the D3–NS5 system preserves every supersymmetry with a generator $\varepsilon = \eta \otimes \varepsilon_0$, with $\eta \in \mathsf{V}_8$. So in particular, once we impose the relation (2.27) between the parameters, this boundary condition preserves the supersymmetry generator of the twisted topological field theory. Substituting (2.27) in (2.14) and solving for t^2, we get the surprisingly simple result

$$t^2 = \frac{\bar\tau}{\tau}. \tag{2.28}$$

[6] In the following formulas, there is no sum over μ; a covariant version reads $(\Gamma_\mu\Gamma_{4+\nu} + \Gamma_\nu\Gamma_{4+\mu})\varepsilon_\ell = -2g_{\mu\nu}\varepsilon_r$, $(\Gamma_\mu\Gamma_{4+\nu} + \Gamma_\mu\Gamma_{4+\mu})\varepsilon_r = 2g_{\mu\nu}\varepsilon_l$.

The operation $t \to -t$ corresponds to $a \to -1/a$ and to exchange of the D3–NS5 and D3–$\overline{\text{NS5}}$ systems.[7]

With the aid of (2.27), the boundary couplings (2.19) can be rewritten

$$I^* = \frac{1}{g_{\text{YM}}^2}\int_{x^3=0} \mathrm{d}^3x\, \epsilon^{\mu\nu\lambda}\mathrm{Tr}\Big(-\frac{t+t^{-1}}{3}\varphi_\mu\varphi_\nu\varphi_\lambda + \frac{t+t^{-1}}{t-t^{-1}}\Big(A_\mu\partial_\nu A_\lambda + \frac{2}{3}A_\mu A_\nu A_\lambda\Big)\Big). \tag{2.29}$$

2.2.3. Global formulation and brane construction. The topological field theory under discussion can be defined on any (oriented) four-manifold, possibly with boundary. One can motivate how to do this by generalizing the D3–NS5 system beyond the special geometry that we have considered so far.

Introducing a slightly new nomenclature for a reason that will soon be clear, let V_0 be an arbitrary oriented four-manifold, and consider Type IIB superstring theory on $T^*V_0 \times \mathbb{R}^2$. For the moment, suppose that T^*V_0 admits a complete Calabi–Yau metric. Consider N D3-branes wrapped on $V_0 \times \{0\} \subset T^*V_0 \times \mathbb{R}^2$, where 0 is a point in $\mathbb{R}^2$ (the "origin"). This system is topologically twisted in precisely the way described in Section 2.2.1. Type IIB superstring theory on $T^*V_0 \times \mathbb{R}^2$ has four unbroken supersymmetries, of which two are preserved by the D3-branes wrapped on V_0. The two unbroken supersymmetries precisely correspond to the $\mathrm{SO}'(4)$-invariant supersymmetries with generators ε_ℓ and ε_r, as described above. This approach to realizing topologically twisted gauge theories via branes was described in [12]. The basic idea is that the twisting of the normal bundle to $V_0 \subset T^*V_0$ leads to the R-symmetry twist that is used in defining a topological field theory.

The above remarks are unaffected by possible presence of a Type IIB theta-angle – which becomes the theta-angle of the gauge theory along the D3-branes. Now suppose we are given an oriented three-manifold $W \subset V_0$, such that $T^*W \subset T^*V_0$ is a supersymmetric cycle (a complex submanifold). Then we can wrap an NS5-brane on $T^*W \times \{0\} \subset T^*V_0 \times \mathbb{R}^2$. The NS5-brane preserves half the supersymmetry of Type IIB on $T^*V_0 \times \mathbb{R}^2$ (that is, in the absence of D3-branes, two supercharges are conserved, while if one includes D3-branes, there is one conserved supersymmetry). Moreover, such a W, being oriented and of codimension 1 in V_0, may potentially divide V_0 into two pieces. Assuming this is the case, either one of the pieces, say V, is a four-manifold of boundary W. Now instead of D3-branes wrapped on V_0, we can consider D3-branes wrapped on V and ending on the NS5-brane. The support of the D3-branes is thus $V \times \{0\} \subset T^*V_0 \times \mathbb{R}^2$. With both types of brane present, there is now only one conserved supercharge; its generator is a linear combination of ε_ℓ and ε_r, depending on the theta-angle and other parameters.

[7] As long as the gauge theory parameters g_{YM} and θ are real, $\bar\tau$ is the complex conjugate of τ, so (2.28) implies that t has modulus 1, and (2.27) then implies that a is real. When we get to topological field theory, we may choose to analytically continue τ and $\bar\tau$ to independent complex variables, whereupon t no longer has modulus 1 and a becomes complex.

The geometry assumed above is rather special. For example, a complete Calabi–Yau metric on T^*V_0 exists if V_0 is S^4 or $S^2 \times S^2$, but not for most V_0. Actually, the above construction can be generalized by replacing T^*V_0 by any Calabi–Yau four-fold X that admits V_0 as a special Lagrangian four-cycle; similarly, T^*W can be replaced by any divisor in X. Moreover, we really only care about V, not V_0. So many cases can be realized, but we probably do not have enough freedom to accommodate an arbitrary W and V. Similarly, the brane construction naturally has a D3-brane gauge group $\mathrm{U}(N)$, and, though one could accommodate orthogonal or symplectic gauge groups by adding an orientifold plane to the construction, this construction does not naturally lead to exceptional gauge groups.

From our point of view, the most obvious merit of the brane construction is motivational. It presumably does not literally work, globally, for all oriented four-manifolds V with arbitrary boundary W; nor does it work for all gauge groups. But the brane construction suggests a purely field theoretic construction that does work in general. The R-symmetry twist that was sketched in Section 2.2.1 (and was described in far more detail in Section 3 of [67]) preserves two supercharges when the theory is formulated on an arbitrary four-manifold V; one linear combination of these two supercharges is preserved when V has a boundary W, with a boundary condition that is modeled locally on the D3–NS5 system. All these statements can be verified by infinitesimal calculations on V and W, and the fact that they work in the brane construction is enough to ensure that, as field theoretic statements, they work in general.

Apart from encouragement, what else do we gain from the brane construction? One answer is that ultimately, we will have to understand the behavior under certain nonperturbative dualities. For this, the brane construction provides invaluable insight. A second answer is that to understand Khovanov homology, we will have to ultimately go to five dimensions, where Yang–Mills quantum field theory is not ultraviolet-complete. The most rigorous and general formulation of our construction will ultimately be given in purely field theoretic terms, but the field theory required is the six-dimensional $(0, 2)$ theory (from which five-dimensional super Yang–Mills theory can be derived), whose existence and properties are known only from its multiple relations to string theory, M-theory, and branes. So the insights that come from brane constructions are again essential.

2.2.4. Wilson loops. $\mathcal{N} = 4$ super Yang–Mills theory in four dimensions admits 1/16-BPS Wilson loop operators [112]. They are constructed as follows. The supersymmetry transformation law for the bosonic fields of this theory is

$$\delta A_I = i\bar{\varepsilon}\Gamma_I\lambda = -i\bar{\lambda}\Gamma_I\varepsilon, \quad I = 0, \dots, 9. \tag{2.30}$$

Here we use a ten-dimensional notation; for $I \leq 3$, A_I is a component of a gauge field, and for $I \geq 4$, it is a scalar field. By twisting, we have converted four of the scalar fields to a one-form φ. Usually, we use Greek letters $\mu, \nu, \dots$ for four-dimensional

indices, so we write $A = \sum_{\mu=0}^{3} A_\mu \mathrm{d}x^\mu$, $\varphi = \sum_{\mu=0}^{3} \varphi_\mu \, \mathrm{d}x^\mu = \sum_{\mu=0}^{3} A_{4+\mu} \mathrm{d}x^\mu$.

Suppose that ε is such that

$$(\Gamma_\mu + i\Gamma_{4+\mu})\varepsilon = 0, \quad \mu = 0, \dots, 3. \tag{2.31}$$

Clearly, in this case, Wilson operators of the form

$$\mathrm{Tr}_R \, P \exp \oint_K (A + i\varphi) \tag{2.32}$$

are invariant, for an arbitrary embedded loop K in spacetime and any representation R of the gauge group. Similarly, if

$$(\Gamma_\mu - i\Gamma_{4+\mu})\varepsilon = 0, \quad \mu = 0, \dots, 3, \tag{2.33}$$

then there are supersymmetric Wilson operators of the form

$$\mathrm{Tr}_R \, P \exp \oint_K (A - i\varphi). \tag{2.34}$$

As explained in [67], the supersymmetry generator $\varepsilon = \varepsilon_\ell + t\varepsilon_r$ of interest here obeys (2.31) or (2.33) precisely for $t = i$ or $t = -i$. Therefore, in general, supersymmetric Wilson operators appear in this family of topological field theories precisely at those values of t. The occurrence of supersymmetric Wilson operators at $t = \pm i$ is actually important in geometric Langlands, and played a major role in [67]. But in the present paper, we are interested in other values of t.

Therefore, we do not have supersymmetric Wilson operators – except at the boundary of V. For a Wilson operator supported entirely at the boundary of V, we can use the boundary conditions obeyed by λ, as well as the conditions obeyed by ε, to establish supersymmetry. We will explore the conditions that on the boundary of V

$$0 = \delta(A_\mu + w\varphi_\mu) = -i\bar{\lambda}(\Gamma_\mu + w\Gamma_{4+\mu})\varepsilon, \quad \mu = 0, 1, 2. \tag{2.35}$$

The reason that we impose this condition only for $\mu < 3$ is that the goal is to construct Wilson operators that are supersymmetric only on the boundary of V, at $x^3 = 0$. In (2.35), w is a complex number, to be determined. If (2.35) holds, then upon setting

$$\mathcal{A}_w = A + w\varphi, \tag{2.36}$$

we can construct supersymmetric Wilson operators

$$\mathrm{Tr}_R \, P \exp \oint_K \mathcal{A}_w, \tag{2.37}$$

for any knot K in the boundary of V.

A preliminary reduction is that $\bar{\lambda}(\Gamma_\mu + w\Gamma_{4+\mu})\varepsilon = \bar{\lambda}\Gamma_\mu(1 + w\Gamma_{\mu,4+\mu})\varepsilon = \bar{\lambda}\Gamma_\mu(1 + iwB_0B_1)\varepsilon$. In the second step, we used the fact that $\Gamma_{\mu,\mu+4}\varepsilon = iB_0B_1\varepsilon$.

This follows from (2.21), (2.24), (2.23), and the fact that ε is a linear combination of ε_ℓ and ε_r. So we need to explore the vanishing of

$$\bar\lambda\Gamma_\mu(1+iwB_0B_1)\varepsilon. \tag{2.38}$$

The expression $(\lambda,\varepsilon)=\bar\lambda\Gamma_\mu\varepsilon$, for any μ, gives a symmetric bilinear form on the **16** of SO(1, 9). As before, we decompose $\mathbf{16}=\mathsf{V}_8\otimes\mathsf{V}_2$. For $\mu\le 2$, $\bar\lambda\Gamma_\mu\varepsilon$ is the tensor product of a symmetric bilinear form on V_8 (transforming as $(\mathbf{3},\mathbf{1},\mathbf{1})$ under $\mathrm{SO}(1,2)\times\mathrm{SO}(3)_X\times\mathrm{SO}(3)_Y$) with a symmetric bilinear form on V_2. If we represent $\vartheta,\varepsilon_0\in\mathsf{V}_2$ as two-component column vectors, then the form on V_2 can be written as $\vartheta^T\varepsilon_0$. The fermion boundary condition of the D3–NS5 system says that λ, on the boundary, is the tensor product of some vector in V_8 with $\vartheta\in\mathsf{V}_2$ (where ϑ was defined in (2.7)), and similarly the generator ε of any unbroken supersymmetry of the D3–NS5 boundary condition, including the one of topological interest, is the tensor product of some vector in V_8 with ε_0 (defined in (2.5)). So to justify the definition (2.37) of supersymmetric Wilson loops, we require

$$\vartheta^T(1+iwB_0B_1)\varepsilon_0=0. \tag{2.39}$$

With the definitions of ϑ and ε_0 and the formulas (2.4) for B_0 and B_1, it is straightforward to compute that (2.39) is obeyed precisely if

$$w=i\frac{a^2-1}{a^2+1}=\frac{t-t^{-1}}{2}, \tag{2.40}$$

where in the last step, we used the relation (2.27). For real θ and g_{YM}, a is always real (by virtue of (2.14)), so the first formula in (2.40) shows that w is always imaginary. With the help of (2.28), we find

$$w=\mp i\frac{\mathrm{Im}\,\tau}{|\tau|}, \tag{2.41}$$

with the signs corresponding to $t=\pm|\tau|/\tau$.

The action I of $\mathcal{N}=4$ super Yang–Mills theory on a four-manifold V is the sum of a term proportional to $1/g_{\mathsf{YM}}^2$, which contains the kinetic energy for all fields, and a term proportional to θ:

$$I=\frac{1}{g_{\mathsf{YM}}^2}\int_V \mathrm{d}^4x\sqrt{g}\mathscr{L}_{\mathrm{kin}}+i\frac{\theta}{32\pi^2}\int_V \mathrm{d}^4x\,\epsilon^{\mu\nu\alpha\beta}\,\mathrm{Tr}\,F_{\mu\nu}F_{\alpha\beta}. \tag{2.42}$$

Here, for later reference, the part of $\mathscr{L}_{\mathrm{kin}}$ that involves A,φ only is (in Euclidean signature)

$$\mathscr{L}_{\mathrm{kin}}^{A,\varphi}=-\mathrm{Tr}\Big(\frac{1}{2}F_{\mu\nu}F^{\mu\nu}+D_\mu\varphi_\nu D^\mu\varphi^\nu+R_{\mu\nu}\varphi^\mu\varphi^\nu+\frac{1}{2}[\varphi_\mu,\varphi_\nu]^2\Big). \tag{2.43}$$

($R_{\mu\nu}$ is the Ricci tensor of V; when V is not Ricci-flat, the indicated term proportional to $R_{\mu\nu}$ is needed for Q-invariance.)

Let us first consider the case that V has no boundary. Both terms on the right hand side of (2.42) are Q-invariant. The θ term is Q-invariant because, more generally, it is a topological invariant, unchanged in any continuous deformations. It represents a nonzero class in the cohomology of Q (unless $t = \pm i$, as discussed momentarily). One might suspect that the integral of $\mathcal{L}_{\rm kin}$ would vanish in the cohomology of Q, as happens in many twisted topological field theories, but this is actually not the case. Instead, as shown in [67], the first term on the right of (2.42) is equivalent mod $\{Q, \dots\}$ to a multiple of the second term. The precise relation is

$$I = \{Q, \dots\} + \frac{2\pi i \Psi}{32\pi^2} \int_V \mathrm{d}^4 x \, \epsilon^{\mu\nu\alpha\beta} \, \mathrm{Tr} \, F_{\mu\nu} F_{\alpha\beta}, \tag{2.44}$$

where

$$\Psi = \frac{\theta}{2\pi} + \frac{4\pi i}{g_{\rm YM}^2} \frac{t - t^{-1}}{t + t^{-1}} \tag{2.45}$$

was called in [67] *the canonical parameter*.

Before twisting, $\mathcal{N} = 4$ super Yang–Mills theory in four dimensions depends on a complex parameter $\tau = \theta/2\pi + 4\pi i/g_{\rm YM}^2$, which is valued in the upper half-plane. Upon twisting, an additional complex parameter t appears in the choice of the topological supercharge. It was shown in [67] that the topological field theory obtained in this way depends on the two parameters τ and t only via their combination Ψ. A sketch of this argument is as follows. For the special cases $t = \pm i$, which correspond to $\Psi = \infty$, one shows directly that both terms on the right of (2.42) are of the form $\{Q, \dots\}$, so the parameter τ is irrelevant if $\Psi = \infty$. (The case $\Psi = \infty$ is important for geometric Langlands, but not for the present paper.) For $t \neq \pm i$, it is shown in [67] that by including auxiliary fields and making a local redefinition of the fermion fields, one can make the Q-transformation laws of all fields independent of t. After one thus eliminates the dependence of the theory on t that is hidden in the definition of Q, equation (2.44) shows that for fixed Ψ, t appears only in a term $\{Q, \dots\}$ and thus is irrelevant for the topological field theory.

In [67], the transformation of t under electric-magnetic duality was determined. It was shown that under a general S-duality transformation

$$\tau \longrightarrow \frac{a\tau + b}{c\tau + d}, \tag{2.46}$$

t transforms by

$$t \longrightarrow \frac{c\tau + d}{|c\tau + d|} t \tag{2.47}$$

and that Ψ transforms just as τ does:

$$\Psi \longrightarrow \frac{a\Psi + b}{c\Psi + d}. \tag{2.48}$$

(Unlike τ, Ψ is not restricted to take values in the upper half plane.) The formula (2.45) for Ψ holds for all τ, t. Imposing the relations (2.14), (2.27) that are natural in studying the D3–NS5 system, we can derive several interesting alternative formulas. Eliminating t in favor of $g_{\rm YM}$ and θ, we find

$$\Psi = \frac{|\tau|^2}{{\rm Re}\,\tau}, \tag{2.49}$$

showing that Ψ is always real for the D3–NS5 system with physical values of the parameters (real $g_{\rm YM}$ and θ). Alternatively, eliminating θ in favor of $g_{\rm YM}$ and t, we get

$$\Psi = \frac{4\pi i}{g_{\rm YM}^2}\Big(\frac{t-t^{-1}}{t+t^{-1}} - \frac{t+t^{-1}}{t-t^{-1}}\Big). \tag{2.50}$$

Now let us discuss what happens when V has a boundary. $\int_V {\rm d}^4x\,\epsilon^{\mu\nu\alpha\beta}{\rm Tr}\,F_{\mu\nu}F_{\alpha\beta}$ is no longer Q-invariant, but varies by a boundary term. It is convenient to replace this integral by a multiple of the Chern–Simons function. We define the Chern–Simons function ${\rm CS}(\mathcal{A})$, for any connection $\mathcal{A}$, possibly complex-valued, by

$${\rm CS}(\mathcal{A}) = \frac{1}{4\pi}\int_{\partial V}{\rm d}^3x\,\epsilon^{\mu\nu\lambda}{\rm Tr}\,\Big(\mathcal{A}_\mu\partial_\nu\mathcal{A}_\lambda + \frac{2}{3}\mathcal{A}_\mu\mathcal{A}_\nu\mathcal{A}_\lambda\Big). \tag{2.51}$$

In terms of this function, we can make the following substitution on the right hand side of (2.44):

$$\frac{2\pi i\Psi}{32\pi^2}\int_V {\rm d}^4x\,\epsilon^{\mu\nu\alpha\beta}\,{\rm Tr}\,F_{\mu\nu}F_{\alpha\beta} \to i\Psi\,{\rm CS}(A). \tag{2.52}$$

As it was explained in the context of (2.15), relation (2.52) must be treated with care, since ${\rm CS}(A)$ is not quite gauge-invariant (but only invariant under topologically trivial gauge transformations), and the equality suggested in (2.52) really holds only modulo an integer multiple of $2\pi i\Psi$. The substitution (2.52) is a convenient shorthand, which can be used in computing the variation of the integral on the left under a small change in the connection, such as that generated by Q. For future reference, writing h for the dual Coxeter number of G, we can write a formula equivalent to (2.51) in terms of a trace ${\rm Tr}_{\rm ad}$ in the adjoint representation of G:

$${\rm CS}(\mathcal{A}) = \frac{1}{8\pi h}\int_{\partial V}{\rm d}^3x\,\epsilon^{\mu\nu\lambda}{\rm Tr}_{\rm ad}\,\Big(\mathcal{A}_\mu\partial_\nu\mathcal{A}_\lambda + \frac{2}{3}\mathcal{A}_\mu\mathcal{A}_\nu\mathcal{A}_\lambda\Big). \tag{2.53}$$

Concretely, when we write Ψ as in (2.45), the part of $i\Psi\,{\rm CS}(A)$ that is proportional to θ is already present in (2.29). The part proportional to $1/g_{\rm YM}^2$ appears upon writing the kinetic energy as $\{Q,\dots\}$ plus a multiple of the theta term, to arrive at (2.44). In the derivation of (2.44), one can assume that V has no boundary, since the integral $\int_V {\rm Tr}\,F\wedge F$ is in general non-zero even in that case. In Section 2.3, we will repeat the derivation of (2.44), for the case that V has a non-empty boundary. When we do this,

additional boundary terms will appear; this should come as no surprise, since one such term is already visible in (2.29) and Q-invariance implies that there must be more. In fact, the boundary couplings must be a function of $\mathcal{A}_w$ only (modulo Q-exact terms), since this is the only non-trivial Q-invariant combination of boundary fields.

One can determine the form of the full boundary couplings without any computation, using gauge invariance and dimensional analysis plus the fact that the boundary coupling is a function only of $\mathcal{A}_w$. These conditions imply that it must be simply a multiple of $\mathrm{CS}(\mathcal{A}_w)$; there is no other local, gauge-invariant functional of dimension three. For a reason that we will explain momentarily, the coefficient of $\mathrm{CS}(\mathcal{A}_w)$ is precisely $i\Psi$. So the generalization of (2.44) in the presence of a boundary is

$$I = \{Q, \dots\} + i\Psi\,\mathrm{CS}(\mathcal{A}_w). \tag{2.54}$$

When $\mathrm{CS}(\mathcal{A}_w)$ is written explicitly as a function of A and φ, the φ-dependent terms are given by local, gauge-invariant integrals, since

$$\begin{aligned}\mathrm{CS}(\mathcal{A}_w) = \mathrm{CS}(A) + \frac{1}{4\pi}\int_{\partial V} \mathrm{d}^3x\,\epsilon^{\mu\nu\lambda}\,\mathrm{Tr}\Big(&w\varphi_\mu F_{\nu\lambda} \\ &+ w^2\varphi_\mu D_\nu\varphi_\lambda + \frac{2w^3}{3}\varphi_\mu\varphi_\nu\varphi_\lambda\Big).\end{aligned} \tag{2.55}$$

Because those terms are local, gauge-invariant integrals over the boundary of V, they cannot be detected directly by a computation that assumes that this boundary is empty.

However, because $\mathrm{CS}(A)$ is not completely gauge-invariant, and must really be written as an integral over V, its coefficient is determined by the analysis of the case $\partial V = \varnothing$ in [67] and can be read off from (2.44), via the substitution (2.52). From this we learn that the coefficient of $\mathrm{CS}(A)$ in the boundary interaction is $i\Psi$, and in view of (2.55), the coefficient of $\mathrm{CS}(\mathcal{A}_w)$ must be the same. Still, one would naturally like to generalize (2.44) to the case $\partial V \neq \varnothing$, so as to see explicitly the origin of the φ-dependent boundary couplings. This is one of our next goals.

2.3. Localization and the boundary formula. Under favorable conditions, computations in topological field theory can be localized on configurations that obey $\{Q, \zeta\} = 0$, for all fermion fields ζ. Among the fermions of[8] $\mathrm{F} = -1$ in the present model are a selfdual two-form χ^+, an anti-selfdual two-form χ^-, and a scalar η (like all fields of $\mathcal{N} = 4$ super Yang–Mills theory, they are adjoint-valued). They have the property that $\mathcal{V}^+ = \{Q, \chi^+\}$, $\mathcal{V}^- = \{Q, \chi^-\}$, and $\mathcal{V}^0 = \{Q, \eta\}$ depend on A, φ only:

$$\begin{aligned}\mathcal{V}^+ &= (F - \varphi\wedge\varphi + t\,\mathrm{d}_A\varphi)^+, \\ \mathcal{V}^- &= (F - \varphi\wedge\varphi - t^{-1}\mathrm{d}_A\varphi)^-, \\ \mathcal{V}^0 &= D_\mu\varphi^\mu.\end{aligned} \tag{2.56}$$

[8]The fermion number F was defined in Section 2.2.1.

Here for any two-form α, we write α^+ and α^- for its selfdual and anti-selfdual projections. Localization on real fields A, φ can be achieved for real[9] t by adding a suitable term to the action I:

$$\begin{aligned} I \longrightarrow & I - \frac{1}{\epsilon}\left\{Q, \int_V \mathrm{Tr}\,(\chi^+\mathcal{V}^+ + \chi^-\mathcal{V}^- + \chi^0\mathcal{V}^0)\right\} \\ & = I - \frac{1}{\epsilon}\int_V \mathrm{Tr}\,((\mathcal{V}^+)^2 + (\mathcal{V}^-)^2 + (\mathcal{V}^0)^2 + \cdots), \end{aligned} \tag{2.57}$$

where ϵ is a small parameter and the omitted terms are fermion bilinears. For t real, $\mathcal{V}^+$, $\mathcal{V}^-$, and $\mathcal{V}^0$ are real, and the modified action diverges as $1/\epsilon$ unless the localization equations

$$(F - \varphi\wedge\varphi + t\,\mathrm{d}_A\varphi)^+ = \left(F - \varphi\wedge\varphi - t^{-1}\mathrm{d}_A\varphi\right)^- = D_\mu\varphi^\mu = 0 \tag{2.58}$$

are satisfied. So the path integral is supported, for $\epsilon \to 0$, on the space of solutions of those equations. On the other hand, the integral is independent of ϵ, since the term we have added to the action is of the form $\{Q, \dots\}$. The fact that this sort of argument is most straightforward for real t is not a major inconvenience, since for any Ψ (other than $\Psi = \infty$) there is always a convenient choice of real t.

There are also localization equations that depend on σ. For $t \neq \pm i$, they are

$$D_\mu\sigma = [\varphi_\mu, \sigma] = [\sigma, \bar\sigma] = 0. \tag{2.59}$$

They say that the gauge transformation generated by σ is a symmetry of the whole configuration. Under favorable conditions (for instance, if the gauge field is irreducible and has no continuous gauge symmetries, or if a boundary conditions sets σ to zero somewhere), they imply that σ is identically zero.

To understand explicitly the origin of the φ-dependent boundary terms in (2.55), we have to make more explicit the relation of the localization procedure of (2.57) to the physical action of $\mathcal{N} = 4$ Yang–Mills theory. The identity we need is actually the generalization of (3.33) of [67] to the case that $\partial V \neq \varnothing$:

$$\begin{aligned} & -\int_V \mathrm{d}^4x\,\mathrm{Tr}\left(\frac{t^{-1}}{t+t^{-1}}\mathcal{V}^+_{\mu\nu}\mathcal{V}^{+\mu\nu} + \frac{t}{t+t^{-1}}\mathcal{V}^-_{\mu\nu}\mathcal{V}^{-\mu\nu} + (\mathcal{V}^0)^2\right) \\ & \quad = \int_V \mathrm{d}^4x\,\sqrt{g}\,\mathcal{L}^{A,\varphi}_{\mathrm{kin}} + \frac{t-t^{-1}}{4(t+t^{-1})}\int_V \mathrm{d}^4x\,\epsilon^{\mu\nu\alpha\beta}\,\mathrm{Tr}\,F_{\mu\nu}F_{\alpha\beta} \\ & \qquad + \int_{\partial V}\mathrm{d}^3x\,\epsilon^{\mu\nu\lambda}\,\mathrm{Tr}\Big(-\frac{2}{t+t^{-1}}\varphi_\mu F_{\nu\lambda} - \frac{t-t^{-1}}{t+t^{-1}}\varphi_\mu D_\nu\varphi_\lambda \\ & \qquad\qquad + \frac{4}{3}\frac{1}{t+t^{-1}}\varphi_\nu\varphi_\nu\varphi_\lambda\Big). \end{aligned} \tag{2.60}$$

[9] According to (2.28), t is not real for physical values of the parameters; in fact, for weak coupling, it is close to $\pm i$. We are here using our freedom to change t as we wish while keeping Ψ fixed.

The left hand side of (2.60) is of the form $\{Q,\dots\}$ modulo fermion bilinears, by the same reasoning as in (2.57). One can write a more complete version of the formula that includes the fermions and also σ; this makes the formulas longer without contributing additional boundary terms. On the right hand side of (2.60), $\int \mathscr{L}_{\rm kin}^{A,\varphi}$ is (after including fermions and σ) the part of the bulk action of $\mathcal{N}=4$ super Yang–Mills theory that is proportional to $1/g_{\rm YM}^2$. The boundary terms that we want are the remaining terms on the right hand side of (2.60). Thus, after multiplying by $1/g_{\rm YM}^2$ and making the substitution (2.52) in one term, we can rewrite (2.60) as follows:

$$\begin{aligned}\frac{1}{g_{\rm YM}^2}\int_V {\rm d}^4x\,\sqrt{g}\,\mathscr{L}_{\rm kin}&\\ =\{Q,\dots\}+\frac{1}{g_{\rm YM}^2}\int_{\partial V}{\rm d}^3x\,\epsilon^{\mu\nu\lambda}\,{\rm Tr}\Big(&-\frac{t-t^{-1}}{t+t^{-1}}\Big(A_\mu\partial_\nu A_\lambda+\frac{2}{3}A_\mu A_\nu A_\lambda\Big)\\ &+\frac{2}{t+t^{-1}}\varphi_\mu F_{\nu\lambda}+\frac{t-t^{-1}}{t+t^{-1}}\varphi_\mu D_\nu\varphi_\lambda-\frac{4}{3}\frac{1}{t+t^{-1}}\varphi_\mu\varphi_\nu\varphi_\lambda\Big).\end{aligned}\tag{2.61}$$

When we add the boundary terms that have appeared in (2.61) to the boundary terms (2.29) that are already present in the physical theory, before twisting, we find that the action has the expected form

$$\{Q,\dots\}+i\,\Psi{\rm CS}(\mathscr{A}_w),\tag{2.62}$$

with the expected value $w=(t-t^{-1})/2$.

2.4. Relation to Chern–Simons theory. So far we have analyzed this problem starting with the D3–NS5 system. The coupling parameters $g_{\rm YM}$ and θ and the parameter a in the boundary condition were all real. This physical starting point has many advantages, such as the insight that it will give about the behavior under various nonperturbative dualities.

But let us see what we can say purely from the standpoint of topological field theory. Here we allow ourselves to continue all parameters to complex values. Keeping Ψ fixed, we may choose, roughly speaking, any value of t that we wish. The only restriction is that we may only vary t in such a way that the path integral continues to converge. What is convenient is to pick t to be real, for then, as we recalled in Section 2.3, there is a straightforward procedure to localize the path integral on solutions of the equations $\mathcal{V}^+=\mathcal{V}^-=\mathcal{V}^0=0$.

These are elliptic differential equations, as described in [108]. On rather general grounds, given a system of elliptic differential equations on a manifold V with boundary $\partial V=W$, the space of solutions of the equations gives a cycle Γ in the space of boundary data and this cycle is within a finite amount of being middle-dimensional. In the present problem, the boundary data are the fields $\mathscr{A}_w=A+w\varphi$ on W, and we want to interpret Γ as an integration cycle in the integral over $\mathscr{A}_w$.

We are actually now in a situation that has been analyzed in detail in Section 5 of [109]. Localization of the path integral on the space of solutions of the equations means that a path integral over bosons and fermions on the four-manifold V reduces to an integral over the purely bosonic fields $\mathcal{A}_w$ on the three-manifold $W = \partial V$. Localization further means that the integral over the boundary fields $\mathcal{A}_w$ reduces to an integral over the cycle Γ. In this reduction, the part of the action that is of the form $\{Q, \dots\}$ gets dropped, leaving only – in the present context – the boundary action $i\Psi\,\mathrm{CS}(\mathcal{A}_w)$.

Actually, at this stage we have a problem of index theory. The classical theory under discussion has the conserved fermion number F. This conservation law has an anomaly that is related in the usual way to the index theorem for the Dirac operator of the theory. This operator and its elliptic boundary condition are described in Appendix A of [109]. A nonzero index means that the four-dimensional path integral vanishes unless we insert a suitable operator violating F in the appropriate way. We say that Γ is a *middle-dimensional cycle* when the index vanishes, and in general that Γ *departs from being middle-dimensional* by an amount equal to the index. In the present problem, the index was analyzed[10] in Section 4.1.1 of [108]. It is independent of the choice of underlying G-bundle $E \to V$, simply because the fermions of given F transform in a real representation of G (namely the adjoint representation), and is proportional to the Euler characteristic of V.

We will be interested primarily in the case that the index vanishes. (A typical example of a similar problem in which the index is nonzero, so that an operator insertion is needed to get a nonzero path integral, is described in Section 2 of [109].) Then Γ is a middle-dimensional cycle. The four-dimensional path integral is generically nonzero and localization means that it reduces to an integral of the boundary fields over Γ:

$$\int_\Gamma D\mathcal{A}_w \exp(-i\Psi\,\mathrm{CS}(\mathcal{A}_w)). \tag{2.63}$$

This has been described in Section 5.2.2 of [109].

At this point, the precise value of w is not important. All that matters is that it has a nonzero imaginary part, so that $\mathcal{A}_w = A + w\varphi$ is a complex-valued connection. The integral (2.63) has no dependence on w except in the definition of $\mathcal{A}_w$, so we can eliminate w by simply writing $\mathcal{A}$ for $\mathcal{A}_w$. (In [109], w was set to i, but the analysis could have been made in the same way for any w with nonzero imaginary part.) Accordingly, we rewrite (2.63) with $w = i$:

$$\int_\Gamma D\mathcal{A} \exp(-i\Psi\,\mathrm{CS}(\mathcal{A})). \tag{2.64}$$

Now we should address the question of what are the possible values of Ψ. In

[10] The operator whose index we want is the operator $\mathrm{d}_A + \mathrm{d}_A^*$ mapping differential forms of odd degree to those of even degree. The requisite boundary conditions, which were described in Appendix A of [109]), are slightly unusual, but they are homotopic to standard boundary conditions in which the restriction of a differential form on V to ∂V vanishes. With these boundary conditions, the index is $-\chi(V)\dim G$.

our derivation starting with the D3–NS5 system, with physically sensible values of the parameters, Ψ has turned out to be an arbitrary nonzero real number, given according to (2.49) by $\Psi = |\tau|^2/\mathrm{Re}\,\tau$. From a topological field theory point of view, as in [109], one can make a more general choice of the twisting parameter t and then Ψ is an arbitrary nonzero complex number.[11] Both points of view are useful. The physical one based on the D3–NS5 system will enable us to understand the role of nonperturbative dualities. The topological field theory point of view leads among other things to holomorphy in Ψ, which we will make use of momentarily.

The relation of a "contour" integral such as (2.64) to ordinary Chern–Simons gauge theory with compact gauge group G has been discussed in [108]. Let $\mathfrak{g}$ and $\mathfrak{g}_{\mathbb{C}}$ be the Lie algebras of G and of its complexification $G_{\mathbb{C}}$, and let $\mathcal{U}$ be the space of all real gauge fields, that is all $\mathfrak{g}$-valued connections A on some principal G-bundle $E \to W$. And let $\mathcal{U}_{\mathbb{C}}$ be the complexification of $\mathcal{U}$, or in other words the space of all $\mathfrak{g}_{\mathbb{C}}$-valued connections on the complexification of E. We denote such a connection as $\mathcal{A}$. The path integral of ordinary Chern–Simons theory with the compact gauge group G is

$$\int_{\mathcal{U}} DA\, \exp(-ik\,\mathrm{CS}(\mathcal{A})), \tag{2.65}$$

and here the "level" k has to be an integer, in order to make the integrand of the path integral gauge-invariant. (There is no such restriction on Ψ in (2.64), as explained in [109], because of the choice of integration cycle Γ.) Usually one says that the path integral does not make sense if $k = 0$ (since one needs a nontrivial oscillatory factor $\exp(-ik\,\mathrm{CS}(\mathcal{A}))$ to define a sensible integral over the space of connections), and one chooses the orientation of W to restrict to the case $k > 0$. We will instead consider both signs of k.

It looks like the ordinary Chern–Simons path integral with gauge group G is the special case of (2.64) with $\Gamma = \mathcal{U}$, that is, with the integration cycle chosen to be the obvious cycle that parametrizes real gauge fields. To emphasize this, in (2.65) we have denoted the argument of the Chern–Simons function as a complex connection $\mathcal{A}$, although the integral is evaluated on the real cycle $\mathcal{U}$, where $\mathcal{A}$ reduces to a real connection A. However, before drawing conclusions about the relation of (2.64) to ordinary Chern–Simons theory, we have to be careful in comparing the holomorphic volume forms that appear in the two integrals.

The integration form that has been denoted as DA in (2.65) arises by analytic continuation to $\mathcal{U}_{\mathbb{C}}$ of the usual integration form (which we also call DA) of the Feynman integral of the $\mathfrak{g}$-valued theory. The corresponding form $D\mathcal{A}$ is induced from the four-dimensional path integral on V. Both DA and $D\mathcal{A}$ are Calabi–Yau volume forms on the same space, namely $\mathcal{U}_{\mathbb{C}}$. So, *a priori*, their ratio is an invertible

[11] Alternatively, one can reach generic Ψ by analytically continuing to complex values of the gauge theory theta-angle θ, and otherwise using the formulas of the present paper. Giving θ an imaginary part violates unitarity, and indeed it appears that reality of Ψ is related to unitarity.

holomorphic function on $\mathcal{U}_{\mathbb{C}}$. We propose that the relation is

$$DA = D\mathcal{A}\, \exp(-ih\, \mathrm{sign}(k)\, \mathrm{CS}(\mathcal{A}))\, \mathfrak{N}_0. \tag{2.66}$$

Here h is the dual Coxeter number of G, and $\mathrm{sign}(k)$ is the sign of the integer k. (Formulas somewhat analogous to (2.66) are described in Section 2.7.1 of [108].) In (2.66), we have included a possible multiplicative constant $\mathfrak{N}_0$, which is allowed by holomorphy. The constant $\mathfrak{N}_0$ might depend on the three-manifold W and the choice of a homomorphism $\rho\colon \pi_1(W) \to G_{\mathbb{C}}$ at $y = \infty$ to define the $\mathcal{N} = 4$ path integral, but holomorphy in Ψ, together with the fact that we have already incorporated the effects of the gauge theory theta-angle, does not allow contributions to $\mathfrak{N}_0$ beyond one-loop order.

The relation (2.66) should be demonstrated explicitly – and the constant $\mathfrak{N}_0$ calculated – by comparing the one-loop determinant for $\mathcal{N} = 4$ super Yang–Mills theory on V to the one-loop path integral of ordinary Chern–Simons theory on W. We will not make such an analysis in the present paper. Instead, we content ourselves with the following observation. Suppose that one expands the Chern–Simons path integral (2.65) around a critical point, that is, around a flat connection $\mathcal{A}_0$. The integrand of the path integral has a phase factor $\exp(-ik\mathrm{CS}(\mathcal{A}_0))$. As computed in [102], the phase of the one-loop determinant corrects this to $\exp(-ik'\,\mathrm{CS}(\mathcal{A}_0))$ where

$$k' = k + h\, \mathrm{sign}(k). \tag{2.67}$$

Usually, k is taken to be positive so this formula is written $k' = k + h$, but we want to allow both signs of k, which requires replacing h with $h\,\mathrm{sign}(k)$. (Chern–Simons theory on a three-manifold W is invariant under a reversal of orientation of W together with a change of sign of k; this means that k' must be an odd function of k. Concretely, the term in k' that is linear in h comes from an η-invariant that changes sign if the sign of k is changed.)

Now let us consider the analogous issue for $\mathcal{N} = 4$ super Yang–Mills on V, with a boundary condition that leads to a "contour" integral (2.64) in the space of $\mathfrak{g}_{\mathbb{C}}$-valued connections. The integral is holomorphic in Ψ, so a one-loop shift in the phase factor $\exp(-i\,\Psi\,\mathrm{CS}(\mathcal{A}_0))$ would have to be holomorphic in Ψ. Since there is no holomorphic function that restricts to $\mathrm{sign}(\Psi)$ when Ψ is real, such a term cannot arise.

Our proposal is that no such shift arises from the one-loop determinant of $\mathcal{N} = 4$ super Yang–Mills theory. Instead, the shift is contained in the comparison (2.66) between the path integral measures of the two theories. There is no problem in holomorphy here, since the left hand side is only defined when k is a nonzero integer. According to our proposal, in comparing $\mathcal{N} = 4$ super Yang–Mills theory on V to Chern–Simons theory on W, we should use not the naive $\Psi = k$ but

$$\Psi = k + h\, \mathrm{sign}(k). \tag{2.68}$$

To be more exact, from $\mathcal{N} = 4$ super Yang–Mills theory on V, we can generate a theory that works for general (nonzero) complex Ψ. It can be compared to Chern–Simons theory when Ψ is an integer; in making this comparison we should use (2.68).

As is clear both from Section 2.2.4 of the present paper and from the analysis in [109], we can add knots and Wilson loop operators to this analysis. $\mathcal{N} = 4$ super Yang–Mills theory with supersymmetric Wilson lines inserted on $W = \partial V$ gives an unusual integration cycle in Chern–Simons theory on W with the same Wilson line insertions. A more complete microscopic explanation of the origin of the knots will be presented in Section 5.1.3.

2.5. Choice of *V*. Now we will explain the choice of V that will be most useful in the rest of this paper.

Given an oriented three-manifold W, we want to pick in a natural and general way an oriented four-manifold V with $\partial V = W$. There is no way to do this if V is supposed to be compact. Instead we will pick $V = W \times \mathbb{R}_+$, where $\mathbb{R}_+$ is a half-line $y \geq 0$. Thus y corresponds to the normal coordinate to the boundary, which earlier has been called x^3.

Since V is not compact, we need a boundary condition at $y = \infty$. The boundary condition will be given by a y-independent solution of the localization equations (2.58). As explained in [108], such solutions correspond to conjugacy classes of homomorphism[12] from $\pi_1(W)$, the fundamental group of W, to $G_{\mathbb{C}}$, the complexification of G. We let $\rho\colon \pi_1(W) \to G_{\mathbb{C}}$ be such a homomorphism.

Since $V = W \times \mathbb{R}_+$ has two ends – the boundary at $y = 0$ and the end at $y = \infty$ – we have to be more careful with the formula (2.62) for the action. The complete version of the formula has contributions from both ends:

$$I = \{Q, \dots\} + i\,\Psi \mathrm{CS}(\mathcal{A}) - i\,\Psi \mathrm{CS}(\mathcal{A}_\infty). \tag{2.69}$$

Here we write simply $\mathcal{A}$ for the complex connection at $y = 0$, and $\mathcal{A}_\infty$ for its counterpart at $y = \infty$. $\mathcal{A}_\infty$ is completely determined by the boundary condition at $y = \infty$ and in particular by the choice of ρ, so the term we have added is simply a constant. It is more precise to include the resulting constant in (2.64), so the $\mathcal{N} = 4$ path integral on $W \times \mathbb{R}_+$ is really

$$\mathfrak{N} \int_\Gamma D\mathcal{A} \exp(-i\,\Psi\,\mathrm{CS}(\mathcal{A})), \tag{2.70}$$

where $\mathfrak{N}$ is a normalization factor

$$\mathfrak{N} = \exp(i\,\Psi\,\mathrm{CS}(\mathcal{A}_\infty)). \tag{2.71}$$

[12] To be more precise [22], the solutions correspond to homomorphisms that obey a mild condition of semi-stability: their monodromies should not be strictly triangular.

2.6. Some key details. We now run into an important point, which has also been discussed in Section 5.2.2 of [109]. If W is compact, then $W \times \mathbb{R}_+$ is macroscopically one-dimensional, and we must worry about infrared divergences.

If ρ is irreducible (which we take to mean that the homomorphism $\rho\colon \pi_1(W) \to G_{\mathbb{C}}$ commutes with at most a finite subgroup of $G_{\mathbb{C}}$), then our boundary condition at $y = \infty$ makes the theory "massive" – in the effective one-dimensional physics obtained by compactification on W, all bosons and fermions are massive. Under these conditions, the choice of ρ satisfactorily specifies the boundary conditions.

If instead ρ is reducible – it leaves unbroken a subgroup of G of positive rank – then our boundary condition at $y = \infty$ leads to a reduced one-dimensional theory in which the potential energy as a function of scalar fields has flat directions: there are some scalar fields (such as some components of σ) that can acquire expectation values, at no cost in energy. In one dimension, quantum fluctuations of massless scalars are inevitable and important. The boundary condition at $y = \infty$ is in this case not adequately specified by the choice of ρ; one also needs a quantum wavefunction describing the initial conditions for the massless scalar fields at $y = \infty$. Here we view y as a Euclidean time coordinate.

The dependence on ρ presents a number of problems for the constructions that we will make in the rest of this paper. Our next step, in Section 3.1, will be electric-magnetic duality. At a minimum, to proceed in a situation in which ρ is important, we would need to know how ρ transforms under electric-magnetic duality. Not much is known about this, though a little can be gleaned (for some special choices of W) from [99] and [59]. The reducible ρ's are certainly important for understanding the standard Chern–Simons path integral, since when expressed in terms of cycles associated to flat bundles, it certainly receives contributions from reducible flat bundles.

What happens to the choice of ρ under electric-magnetic duality is a question that presumably can be answered, in principle. The infrared divergences that arise in the reducible case pose another problem that may be more serious. After making electric-magnetic duality, we will in Section 4 make a T-duality to introduce a new time coordinate, and then we will want to consider quantum states that propagate in the time direction. Describing quantum states that propagate in the time direction is, at least at first sight, incompatible with specifying a boundary condition by fixing a quantum state that propagates in the y direction. One would at least need a better language to describe what happens here.

Presumably, none of these problems are insuperable, but there clearly is some work to be done to overcome them.

There is actually a straightforward way to circumvent these problems. This is the approach we will take in most of this paper; it also is the approach that leads to Khovanov homology. Instead of taking W to be compact, we will take $W = \mathbb{R}^3$. (It then is essential to include knots or Wilson loop operators, since Chern–Simons theory on $\mathbb{R}^3$ is trivial without them.) For $W = \mathbb{R}^3$, fluctuations of massless scalar fields on $V = W \times \mathbb{R}_+$ do not present a problem, because V has four non-compact directions. Also, as $\mathbb{R}^3$ is simply-connected, when we take $W = \mathbb{R}^3$, there is a unique

choice of ρ (corresponding to the trivial flat connection), and this choice must map to itself under electric-magnetic duality. So as long as we restrict ourselves to knots in $\mathbb{R}^3$, we avoid all technical problems related to infrared divergences and the behavior of ρ under electric-magnetic duality.

There are additional technical advantages in taking $W = \mathbb{R}^3$. Our approach in this paper naturally leads to an integral (2.64) over a cycle Γ defined by solving flow equations on $V = W \times \mathbb{R}_+$. Γ depends on the choice of ρ, so we might denote it in more detail as Γ_ρ. Khovanov homology is related instead to ordinary real Chern–Simons theory, the integration cycle being the real cycle $\mathcal{U}$. In general, as described in [108], one can expand $\mathcal{U}$ as an integer linear combination of the Γ_ρ's, but it may be hard to determine the coefficients explicitly. For $W = \mathbb{R}^3$, as ρ is unique, all integration cycles are integer multiples of a fundamental one, and the relation is simply[13] $\Gamma = \mathcal{U}$. So the integration cycle that emerges naturally from $\mathcal{N} = 4$ super Yang–Mills theory in four dimensions is equivalent to the usual one of ordinary Chern–Simons theory on the boundary.

Furthermore, the normalization factors $\mathfrak{N}$ and $\mathfrak{N}_0$ of (2.71) and (2.66) equal 1 for $W = \mathbb{R}^3$. We have $\mathfrak{N} = 1$ because $\mathcal{A}_\infty$ is trivial. And $\mathfrak{N}_0 = 1$ on $\mathbb{R}^3$ because we are studying a topological field theory. A "constant" arising from a one-loop determinant on $\mathbb{R}^3$ would be a shift in the ground state energy per unit volume, but such a shift is not possible in a topological field theory.

So there are many advantages to taking $W = \mathbb{R}^3$. Some but not all of these advantages persist in the following more general case. Let W_0 be a rational homology sphere and let $W = W_0 \backslash p$ be W_0 with a point p omitted. W is not compact and we pick on W a metric that near its noncompact end looks like the flat metric on $\mathbb{R}^3$. In this type of example, there are no infrared divergences, but there are in general non-trivial choices of ρ, and to proceed one would need to understand how ρ transforms under electric-magnetic duality, and how to expand $\mathcal{U}$ as a linear combination of the Γ_ρ's.

Khovanov homology has been defined in the literature for knots in $\mathbb{R}^3$ (or S^3). It has proved difficult so far to generalize Khovanov homology to other three-manifolds. The difficulties may be related to some of the points made above. We note, however, that results in [38] appear to be part of an analog of Khovanov homology for the case $W = \mathbb{R} \times C$, with C a Riemann surface.

3. S-duality

To learn something new about Chern–Simons gauge theory, we will apply dualities to the framework analyzed in Section 2. The relevant dualities are standard. Here we consider S-duality and in Section 4, we follow with T-duality.

[13] $\mathcal{U}$ is precisely Γ, rather than a more general integer multiple of Γ, because in general when the real integration cycle is expressed in terms of cycles associated to critical points, the cycles associated to real critical points always enter with coefficient 1, as explained in [108], equation (3.39).

3.1. Electric-magnetic duality. We begin by applying electric-magnetic duality to $\mathcal{N} = 4$ super Yang–Mills theory on $V = W \times \mathbb{R}_+$.

The gauge group G is transformed to the Goddard–Nuyts–Olive or Langlands dual group, which we will denote as $G^\vee$. The $G^\vee$ gauge theory has a theta-angle and gauge coupling, which we call $\theta^\vee$ and $g^\vee_{\mathsf{YM}}$. As usual, we define

$$\tau^\vee = \frac{\theta^\vee}{2\pi} + \frac{4\pi i}{(g^\vee_{\mathsf{YM}})^2}. \tag{3.1}$$

The standard relation between $\tau^\vee$ and τ, generalized [3] to the case that G is not simply-laced, is

$$\tau^\vee = -\frac{1}{\mathfrak{n}_{\mathfrak{g}}\tau}, \tag{3.2}$$

where $\mathfrak{n}_{\mathfrak{g}}$ is the ratio of length squared of long and short roots of G, or equivalently of $G^\vee$. (Thus, $\mathfrak{n}_{\mathfrak{g}} = 1$ if G is simply-laced.) The formula (3.2) can be written as $\tau^\vee = (a\tau + b)/(c\tau + d)$ where

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \pm \begin{pmatrix} 0 & -\sqrt{\mathfrak{n}_{\mathfrak{g}}} \\ \sqrt{\mathfrak{n}_{\mathfrak{g}}} & 0 \end{pmatrix}. \tag{3.3}$$

The two choices of sign differ by the possibility of combining electric-magnetic duality with a discrete chiral symmetry. (This symmetry is an element of the center of the R-symmetry group $\mathrm{SU}(4)_R$; it reverses the sign of the twisting parameter t and maps $(A, \varphi) \to (A, -\varphi)$.) The two choices correspond to duality of the D3–NS5 system with a D3–D5 or D3–$\overline{\text{D5}}$ system, respectively. There is no natural choice of which is which. Either way, the boundary condition of the D3–NS5 system maps to a dual boundary condition, which we will discuss in Section 3.3. Wilson operators supported at $y = 0$ map to 't Hooft operators supported at $y = 0$; these are described in Section 3.6 and modify the boundary conditions.

The family of twisted topological field theories that is relevant in the present paper is mapped to itself by electric-magnetic duality. The twisting parameter $t^\vee$ of the dual description with gauge group $G^\vee$ is related to the twisting parameter t in the original description by

$$t^\vee = \pm \frac{\tau}{|\tau|} t. \tag{3.4}$$

This formula is a special case of (2.47); the sign is the same as the one in (3.3). For the D3–NS5 system, we have $t = \pm\sqrt{\bar\tau/\tau}$ according to (2.28), and this leads to the amazingly simple

$$t^\vee = \pm 1. \tag{3.5}$$

The sign does not matter, as the two choices are exchanged by the discrete chiral symmetry mentioned in the last paragraph. In this paper, we will take $t^\vee = 1$. The localization equations in the $G^\vee$ gauge theory then take a particularly simple form:

$$F - \varphi \wedge \varphi + \star \mathrm{d}_A \varphi = 0 = \mathrm{d}_A \star \varphi. \tag{3.6}$$

The transformation law (2.48) for the canonical parameter Ψ tells us that the parameter $\Psi^\vee$ of the dual theory is related to Ψ by

$$\Psi^\vee = -\frac{1}{\mathfrak{n}_{\mathfrak{g}}\Psi}. \tag{3.7}$$

On the other hand, since $t^\vee = 1$, the formula (2.45) for $\Psi^\vee$ reduces to

$$\Psi^\vee = \frac{\theta^\vee}{2\pi}. \tag{3.8}$$

Combining these formulas,

$$\theta^\vee = 2\pi\Psi^\vee = -\frac{2\pi}{\mathfrak{n}_{\mathfrak{g}}\Psi}. \tag{3.9}$$

For $G^\vee = \mathrm{SU}(N)$, we define the instanton number of the $G^\vee$ gauge theory by

$$\mathrm{P} = \frac{1}{32\pi^2}\int_V \epsilon^{\mu\nu\alpha\beta}\,\mathrm{Tr}\,F_{\mu\nu}F_{\alpha\beta}, \tag{3.10}$$

where Tr is the trace in the N-dimensional representation. For any $G^\vee$, we can take

$$\mathrm{P} = \frac{1}{2h^\vee}\frac{1}{32\pi^2}\int_V \epsilon^{\mu\nu\alpha\beta}\,\mathrm{Tr}_{\mathrm{adj}}\,F_{\mu\nu}F_{\alpha\beta}, \tag{3.11}$$

where[14] $h^\vee$ is the dual Coxeter number of $G^\vee$, and $\mathrm{Tr}_{\mathrm{adj}}$ is the trace in the adjoint representation of $G^\vee$. The symbol Tr will be used as an abbreviation for $\mathrm{Tr}_{\mathrm{adj}}/2h^\vee$ even if $G^\vee$ is not $\mathrm{SU}(N)$. We will eventually modify the definition (3.11) by subtracting a c-number term, that is a term that does not depend on the gauge field A; see (3.30) below.

The role of $\theta^\vee$ in the path integral is simply to weight a field of instanton number P by a factor $\exp(-i\theta^\vee\mathrm{P})$. We set

$$q = \exp(-i\theta^\vee) = \exp(2\pi i/\mathfrak{n}_{\mathfrak{g}}\Psi), \tag{3.12}$$

so that the $\theta^\vee$-dependent factor by which we weight a field of instanton number P is q^{P}. Recalling (2.68), we see that when we compare the $G^\vee$ gauge theory to Chern–Simons theory on $W = \partial V$ with gauge group G, we must take

$$q = \exp\Big(\frac{2\pi i}{\mathfrak{n}_{\mathfrak{g}}(k + h\,\mathrm{sign}(k))}\Big). \tag{3.13}$$

At least for simply-laced G, this is essentially the standard definition of q in Chern–Simons gauge theory (the formula is usually written for positive k, and what we call q is sometimes called q^2 or q^{-1}). Hence, for example, the Jones polynomial of a knot in $\mathbb{R}^3$ (and its generalizations for other groups and representations) is essentially a Laurent polynomial in this variable; for a precise statement, see (1.7).

[14] Thus, in our notation, h is the dual Coxeter number of G and $h^\vee$ is the dual Coxeter number of $G^\vee$. (Note that some authors use $h^\vee$ for the dual Coxeter number of G.)

3.2. Computing the partition function. Now let us discuss how to compute the partition function of the $G^\vee$ gauge theory on V. Because $t^\vee$ is real, the model is analogous to a two-dimensional A-model (or four-dimensional Donaldson theory) and computations can be carried out by an appropriate procedure of counting of classical solutions of the localization equations (3.6). The value $t^\vee = 1$ makes the procedure particularly simple. As $\Psi^\vee$ is independent of $g_{\mathrm{YM}}^\vee$, to calculate the partition function for given $\Psi^\vee$, we can take $g_{\mathrm{YM}}^\vee$ to be arbitrarily small. The partition function then reduces to a sum over classical solutions of the localization equations. The expected dimension of the moduli space of solutions of those equations is given by the index of a certain Dirac-like operator. As is typical of A-type topological field theories, the operator in question is the fermion kinetic operator of the theory, whose index equals the anomaly in the fermion number F. So the expected dimension of the moduli space must vanish in order for the twisted $\mathcal{N} = 4$ path integral on V without any operator insertions to be non-vanishing.[15]

Let us suppose that this is the case and consider the contribution to the path integral from a given solution of the localization equations. For simplicity, assume that in expanding around such a solution, there are no bosonic or fermionic zero modes and no unbroken gauge symmetries. This is the generic state of affairs when the index vanishes. In expanding around such a solution, since we can take $g_{\mathrm{YM}}^\vee$ to be arbitrarily small, we can make a one-loop approximation to the path integral, which – apart from a factor coming from the classical action – reduces to the ratio of fermion and boson determinants. The determinants are equal up to sign, because of supersymmetry, and the boson determinant is always positive. So the ratio of determinants is ± 1, depending on the sign of the fermion determinant. The factor in the path integral from the classical action is q^{P}, coming from the part of the classical action proportional to $\theta^\vee$.

The sum of the contributions of all solutions with $\mathrm{P} = n$ is then $a_n q^n$ for some integer a_n; here a_n is simply the sum of contributions $+1$ and -1 from classical solutions with $\mathrm{P} = n$ and positive or negative fermion determinant. The partition function is the sum of $a_n q^n$ over all values of n:

$$Z(q) = \sum_n a_n q^n. \tag{3.14}$$

As explained in Section 2.5, the $\mathcal{N} = 4$ partition function $Z(q)$ will be most simply related to Chern–Simons theory if $V = \mathbb{R}^3 \times \mathbb{R}_+$, in which case $Z(q)$ and the Chern–Simons path integral on $\mathbb{R}^3$ should simply coincide. To make this case interesting, we include knots in $\mathbb{R}^3$ on the Chern–Simons side and the corresponding loop operators

[15] When the index is non-zero, we make a suitable operator insertion to replace the twisted $\mathcal{N} = 4$ partition function with a non-vanishing path integral. (This can actually only be done when the index is positive, because the cohomology of Q in the space of local operators vanishes for $\mathrm{F} < 0$.) As in other theories of A-model type, the operator insertions have the effect of constraining the solutions of the localization equations and reducing to a situation much like what prevails when the index vanishes. We omit the details, as we do not need them and they are standard in topological field theories of this type.

in the boundary of V in the $\mathcal{N} = 4$ description. The formula $Z(q)$ has been obtained in a dual description by $G^\vee$ gauge theory, so the loop operators are 't Hooft operators (rather than the Wilson operators that were introduced in Section 2.2.4). The presence of 't Hooft operators affects the coefficients a_n in the partition function because it affects the boundary conditions along ∂V, as we will describe in Section 3.6.

The claim that the sum (3.14) reproduces the knot invariants of Chern–Simons theory is one of the main claims of the present paper. For a direct verification of this for the special case corresponding to the Jones polynomial (that is, $G = \mathrm{SU}(2)$ with loop operators associated to the two-dimensional representation of G) see [41].

For future reference, we can rewrite (3.14) as follows. Let S be the set of classical solutions of the localization equations. For $s \in S$, let n_s be the value of P for the corresponding solution, and denote the sign of the fermion determinant obtained when one expands around that solution as $(-1)^{g_s}$. Then

$$Z(q) = \sum_{s\in S} q^{n_s}(-1)^{g_s}. \tag{3.15}$$

What values of the instanton number n occur in (3.14)? Suppose first that $G^\vee$ is simply-connected. Then n is an integer if V is compact and without boundary, but if V has a boundary or an end at infinity, then $n \in \mathbb{Z} + \delta$, where the constant δ depends on the boundary conditions and the behavior at infinity. (We will analyze this dependence in Section 3.5.) If $G^\vee$ is not simply-connected but V is compact and without boundary, then $n \in \mathbb{Z}/w$, where the integer w depends only on $G^\vee$ (for example, $w = 4$ if $G^\vee = \mathrm{SO}(3)$, since the instanton number of an SO(3) bundle $W \to V$ is congruent to $\int_V w_2(E)^2/4 \bmod \mathbb{Z}$). If $G^\vee$ is not simply-connected and V has a boundary or a non-compact end, then $n \in \mathbb{Z}/w + \delta$ for some constant δ. Despite these details, we will loosely refer to a sum of the form (3.14) as a Laurent polynomial if a_n vanishes except for finitely many values of n.

Given that the Chern–Simons path integral for a knot in $\mathbb{R}^3$ can be expressed as in (3.14), can we get a new understanding of the fact that these functions are actually Laurent polynomials? This is true if the localization equations have solutions only for finitely many values of P, since a_n certainly vanishes if there are no solutions at all with $\mathrm{P} = n$. It is shown in [67], Section 3.3, that if V is a compact four-manifold without boundary, then the localization equations (for any value of t other than 0 or ∞) have no solutions except for $\mathrm{P} = 0$. Hopefully, for $\partial V \neq \varnothing$, with the boundary conditions of Sections 3.3 and 3.6, and possibly with a noncompact end, there is a more general result giving a bound on $|\mathrm{P}|$ for any solution. This will ensure that the path integral is a Laurent polynomial.

3.2.1. Some further details. In our simplified explanation of (3.14), we have omitted a few details that will be important in some generalizations.

First of all, under electric-magnetic duality, the action may obtain a c-number term of the form $\alpha\chi(V) + \beta\sigma(V)$ where $\chi(V)$ and $\sigma(V)$ are the Euler characteristic

and signature of V and α, β are universal constants. Such an effect has been described in [99] in the context of a different twist of $\mathcal{N} = 4$ super Yang–Mills theory. If it occurs in the present context, this would multiply the right hand side of (3.14) by $\exp(\alpha\chi(V) + \beta\sigma(V))$. This may be important for some applications, though not for the case $V = W \times \mathbb{R}_+$ that we focus on in the present paper.

Second, we should discuss the role of unbroken gauge symmetries. Given a solution of the localization equations, we write H for the subgroup of $G^\vee$ consisting of gauge transformations that leave fixed the given solution. We call a solution *reducible* if H is a Lie group of positive dimension and *irreducible* if H is a finite group, in which case we denote the number of its elements as $\#H$. Reducible solutions (such as the trivial solution with $A = \varphi = 0$) are inevitably present if $\partial V = \varnothing$. In expanding around a reducible solution, there are flat directions in the classical potential (for example, the potential vanishes for some components of σ), and one has to learn how to integrate over this space of flat directions in order to determine the contribution of a reducible solution to the path integral. This is a rather delicate question, and we will not investigate it here.

There is also some subtlety concerning irreducible solutions when H is non-trivial. For compact V, the contribution of an irreducible solution with non-trivial H is actually not $\pm q^n$ but $\pm q^n/\#H$, where the factor $1/\#H$ results from the process of dividing by the volume of the gauge group. Suppose that V has a nonempty boundary and we use the boundary condition described in Section 3.3. This boundary condition explicitly breaks $G^\vee$ down to its center, which we denote as $\mathcal{Z}(G^\vee)$. The center is always a symmetry of any classical solution, so in this situation we always have $H = \mathcal{Z}(G^\vee)$. If in addition V is compact, (3.14) should be multiplied by $1/\#\mathcal{Z}(G^\vee)$, reflecting the fact that $\mathcal{Z}(G^\vee)$ acts trivially on the space of fields. However, if V also has a noncompact end (as in our basic example $V = W \times \mathbb{R}_+$), one divides only by gauge transformations that are trivial at infinity, and hence the factor of $1/\#\mathcal{Z}$ does not arise.

For $V = W \times \mathbb{R}_+$, we have to define a boundary condition at infinity. We do this just as we did for the original D3–NS5 system: we pick a y-independent solution of the localization equations at infinity. In the present case, this corresponds to a homomorphism $\rho^\vee\colon \pi_1(W) \to G^\vee_{\mathbb{C}}$. The partition function (3.14) can be defined for each $\rho^\vee$, so we really get a family of partition functions $Z_{\rho^\vee}(q)$, labeled by $\rho^\vee$. Similarly, the integral (2.64) is really a family of path integrals I_ρ, labeled by homomorphisms $\rho\colon \pi_1(W) \to G_{\mathbb{C}}$. One expects that electric-magnetic duality will lead to formulas of the general nature

$$Z_{\rho^\vee}(q) = \sum_\rho m_{\rho^\vee,\rho} I_\rho(q), \tag{3.16}$$

with some matrix $m_{\rho^\vee,\rho}$. But little is clear about the nature of this matrix. This problem was pointed out in Section 2.5. Luckily, for the important case $W = \mathbb{R}^3$, we avoid this question.

3.3. The dual boundary condition. We are mainly interested in the case that the four-manifold V has a boundary, so we need to describe the appropriate boundary condition in the $G^\vee$ gauge theory. (We describe here the boundary condition away from possible 't Hooft operators. The more elaborate boundary condition that must be used near an 't Hooft operator is described in Section 3.6.)

For $G^\vee = G = \mathrm{U}(N)$, the boundary conditions that we want are those of the D3–D5 system, or equivalently, the Dp-D$(p+2)$ system for any p. This boundary condition, which is of a rather surprising nature, was first formulated in [25] by comparing to known results about the Nahm transform of BPS monopoles. More intuitive explanations have been given in [16], [81], and [21] in terms of the D$(p+2)$-brane theory and a "fuzzy funnel." A formulation of the boundary condition purely in field theory terms, along with a generalization to any $G^\vee$, has been given in [39].

The boundary condition of the D3–D5 system is defined not by imposing a condition on the fields or their normal derivatives, as in the case of familiar boundary conditions such as Dirichlet and Neumann, but by specifying the singular behavior that the fields should have near the boundary. (This is somewhat like the procedure used to define an 't Hooft operator, or a disorder operator in statistical mechanics; these are also defined by specifying a desired singularity.) The desired behavior is described by giving a model solution of the equations (3.6) that has the desired singularity. In the context of topological field theory, the model solution has to obey the equations in order to preserve the desired topological supersymmetry at $t^\vee = 1$.

In fact, the boundary condition of the D3–D5 system has much more symmetry than that; it is half-BPS, and is invariant under translations and rotations and in fact even conformal transformations that leave fixed the boundary. It is convenient to define the model solution on the half-space $x^3 \geq 0$, and to write y for x^3. In the model solution, the gauge field A vanishes, as does the normal part of the one-form φ. We write $\vec{\varphi} = \sum_{i=0}^{2} \varphi_i \, \mathrm{d}x^i$ for the tangential part of φ. Rotation and translation invariance tell us to look for a model singular solution such that φ is a function of y only. Given all this, the equations (3.6) reduce to Nahm's equations

$$\frac{\mathrm{d}\vec{\varphi}}{\mathrm{d}y} + \vec{\varphi} \times \vec{\varphi} = 0. \tag{3.17}$$

Here $\vec{\varphi} \times \vec{\varphi}$ is the triple of elements of $\mathfrak{g}$ defined by $(\vec{\varphi} \times \vec{\varphi})_0 = [\varphi_1, \varphi_2]$ plus cyclic permutations of indices, or equivalently by $(\vec{\varphi} \times \vec{\varphi})_i = [\varphi_{i+1}, \varphi_{i-1}]$, where we consider the integer-valued label i to be defined modulo 3.

Conformal invariance of the D3–D5 boundary condition means the boundary condition is defined by a solution in which

$$\vec{\varphi} = \vec{t}/y \tag{3.18}$$

for some constant elements $\vec{t}$ of the Lie algebra $\mathfrak{g}^\vee$. Nahm's equations then reduce to

$$[t_i, t_j] = \epsilon_{ijk} t_k, \quad i, j, k = 0, 1, 2, \tag{3.19}$$

where ϵ_{ijk} is the antisymmetric tensor with $\epsilon_{012} = 1$. Equation (3.19) is equivalent to saying that the elements $\vec{t}$ are the images of a standard set of SU(2) generators under some Lie algebra homomorphism $\xi\colon \mathfrak{su}(2) \to \mathfrak{g}^\vee$.

Having picked ξ, the boundary condition on $\vec{\varphi}$ is

$$\vec{\varphi} = \frac{\vec{t}}{y} + \cdots, \tag{3.20}$$

where the ellipses refer to terms less singular than $1/y$. The other three scalar fields (the normal part of φ and the real and imaginary parts of σ) vanish at $y = 0$, regardless of ξ. This is deduced in [39] as a consequence of supersymmetry; in a D3–D5 brane construction, it asserts that scalar fields that describe motion of the D3-branes normal to the D5-brane must vanish on the boundary. The gauge field A obeys a shifted version of Dirichlet boundary conditions, as described in Section 3.4 below.

The procedure just sketched, with any choice of ξ, leads to a half-BPS boundary condition that preserves the desired supersymmetry. However, as explained in [39], the boundary condition we want (S-dual to the generalized Neumann boundary conditions that were our starting point in Section 2) corresponds to the case ξ is a "principal embedding" [71] of $\mathfrak{su}(2)$ in $\mathfrak{g}^\vee$. A principal embedding is unique up to conjugacy, for any $G^\vee$.

For $G^\vee = \mathrm{SU}(N)$ or $\mathrm{U}(N)$, a principal embedding is defined by picking an SU(2) subgroup of $G^\vee$ such that the fundamental N-dimensional representation of $G^\vee$ restricts to an irreducible representation of SU(2). For $G^\vee = \mathrm{U}(N)$, the principal embedding arises for N D3-branes ending on a single D5-brane; other choices of ξ can be realized with N D3-branes ending on multiple D5-branes.

For all other groups, a principal embedding is, roughly speaking, as close to irreducible as possible. For example, for $G^\vee = \mathrm{SO}(2k+1)$, the fundamental $(2k+1)$-dimensional representation is irreducible under a principal SU(2) subgroup. This is possible because an irreducible $(2k+1)$-dimensional representation of SU(2) is real, and hence the SU(2) matrices acting in this representation can be embedded in $\mathrm{SO}(2k+1)$. For $G^\vee = \mathrm{SO}(2k)$, the best we can do is to pick an SU(2) subgroup under which the fundamental representation decomposes as $2k = (2k-1)+1$, and this is a principal SU(2) subgroup. For $G^\vee = \mathrm{Sp}(2k)$, a principal SU(2) subgroup is one under which the fundamental $2k$-dimensional representation of G transforms irreducibly; this is possible because an irreducible $2k$-dimensional representation of SU(2) is pseudoreal, so the representation matrices can be embedded in $\mathrm{Sp}(2k)$. For all these classical groups, the principal embedding arises for N D3-branes ending on a single D5-brane in the presence of an orientifold plane. To give one example involving an exceptional Lie group, for $G^\vee = \mathsf{G}_2$, the principal SU(2) embedding is characterized by the fact that the 7-dimensional representation of G_2 transforms irreducibly under a principal SU(2) subgroup of G_2.

We will later need to know a few more basic facts about a principal $\mathfrak{su}(2)$ subalgebra of $\mathfrak{g}$. If G is a simple Lie group of rank r, then its Lie algebra $\mathfrak{g}$ decomposes under

a principal $\mathfrak{su}(2)$ subalgebra as a direct sum of precisely r irreducible representations of dimensions $2j_i + 1$, $i = 1, \dots, r$. (The j_i are always integers.) For $G = \mathrm{SU}(N)$, the j_i are $1, 2, 3, \dots, N-1$ and of course in general

$$\sum_{i=1}^{r}(2j_i + 1) = \dim G. \tag{3.21}$$

The ring of invariant polynomials on the Lie algebra $\mathfrak{g}$ is freely generated by r fundamental Casimir invariants, which are homogeneous of degrees $d_i = j_i + 1$, $i = 1, \dots, r$. For $\mathrm{SU}(N)$, these invariants are the functions $\mathrm{Tr}\, a^d$, $d = 2, \dots, N$.

As a point of terminology, we will refer to the singularity that $\vec{\varphi}$ has at the boundary for the case of a principal $\mathfrak{su}(2)$ embedding as a regular Nahm pole. Referring to this singularity as a Nahm pole requires no explanation. The term "regular" refers to the fact that the raising operator of a principal $\mathfrak{su}(2)$ subalgebra is a regular element of the complex Lie algebra $\mathfrak{g}_{\mathbb{C}}$. (An element of this Lie algebra is called *regular* if the subalgebra that commutes with it has the minimum possible dimension – the rank of G.) For a fuller explanation, see the discussion of (3.53).

3.4. Embedding the tangent bundle. So far we have described the behavior near the boundary for the case that $V = \mathbb{R}^3 \times \mathbb{R}_+$, $\partial V = \mathbb{R}^3$. Now we want to generalize to the case that the boundary of V is an arbitrary three-manifold W with Riemannian metric g_{ij}. We assume that, near its boundary, V looks like a product $W \times \mathbb{R}_+$.

Let us first consider the case that $G^\vee$ is $\mathrm{SU}(2)$ or $\mathrm{SO}(3)$. The gauge field A, restricted to W, is a connection on a $G^\vee$ bundle $E \to W$.

In Section 3.3, for $W = \mathbb{R}^3$, we described the singular part of $\vec{\varphi}$ as $\vec{t}/y$. In the context of the twisted topological field theory, since $\vec{\varphi}$ is interpreted as a one-form, an identification of the Lie algebra $\mathfrak{su}(2)$ with the tangent space to $\mathbb{R}^3$ is implicit here. To make it explicit, we introduce the Kronecker delta δ_i^a and write, in more detail,

$$\vec{\varphi} \cdot \mathrm{d}\vec{x} = \frac{\sum_{i,a} \delta_i^a\, t_a\, \mathrm{d}x^i}{y} + \cdots, \tag{3.22}$$

where t_a are a standard set of $\mathfrak{su}(2)$ generators, obeying $[t_a, t_b] = \epsilon_{abc} t_c$ and (therefore) $\mathrm{Tr}\, t_a t_b = -\delta_{ab}/2$. It is convenient to define a quadratic form on the $\mathfrak{su}(2)$ Lie algebra by $(x, y) = -2\,\mathrm{Tr}\, xy$, so $(t_a, t_b) = \delta_{ab}$.

In the case of a general W, the generalization of (3.22) can only be

$$\vec{\varphi} = \frac{\sum_{i,a} e_i^a\, t_a\, \mathrm{d}x^i}{y} + \cdots, \tag{3.23}$$

where now e_i^a is some tensor that, at any point $p \in W$, reduces to δ_i^a, up to a gauge transformation, in any locally Euclidean coordinate system at p. Such a coordinate system is one in which the metric at p is $g_{ij} = \delta_{ij}$. A covariant way to state the

condition on e_i^a without any restriction on the coordinate system or any choice of gauge is to say that

$$(e_i^a t_a, e_j^b t_b) = g_{ij}, \tag{3.24}$$

which implies that in a locally Euclidean coordinate system, $e_i^a = \delta_i^a$ up to a gauge transformation. An equivalent statement is

$$e_i^a e_j^b \delta_{ab} = g_{ij}. \tag{3.25}$$

But this is a familiar condition in Riemannian geometry. The object e is usually called the *vierbein;* it establishes an isomorphism between the bundle $\mathrm{ad}(E)$ with its natural $\mathfrak{su}(2)$-invariant quadratic form and the tangent bundle TW of W with the quadratic form determined by the metric tensor of W.

Now we have to look more closely at the equations (3.6). As $\vec{\varphi} \sim 1/y$, the equations have terms of order $1/y^2$. By taking the t_i to obey the $\mathfrak{su}(2)$ commutation relations, we ensure vanishing of the $1/y^2$ terms in the equations. We still must consider the terms of order $1/y$ in the equations. Here we find that we need

$$D_i e_j - D_j e_i = 0, \tag{3.26}$$

where $D_i = \partial_i + [A_i, \ \cdot\ ]$ is the usual gauge theory connection. This is another basic equation in Riemannian geometry. It uniquely determines the restriction of A to W to be the Riemannian connection on TW. In fact, this equation is usually taken as the definition of the Riemannian connection on the tangent bundle. We will denote the Riemannian connection on TW as ω.

This is all there is to say if $G^\vee = \mathrm{SO}(3)$: the $G^\vee$ bundle $E \to V$, restricted to the boundary $W = \partial V$, is the tangent bundle to W, and the connection restricted to W is the Riemannian connection. For $G^\vee = \mathrm{SU}(2)$, the $G^\vee$-bundle $E \to W$ is not completely determined by the above description of $\mathrm{ad}(E)$; the additional data required is a choice of spin structure.

The extension of this discussion to any $G^\vee$ is straightforward. The polar part of $\vec{\varphi}$ establishes an isomorphism between TW and a subbundle of $\mathrm{ad}(E)$, and this subbundle corresponds to an $\mathfrak{su}(2)$ subalgebra of $\mathfrak{g}$. The case we want is that the subalgebra is principal. Equation (3.26) says that the gauge field A, restricted to the boundary, is valued in this $\mathfrak{su}(2)$ subalgebra and that its restriction to $\mathfrak{su}(2)$ is the Riemannian connection. Differently put, the bundle $\mathrm{ad}(E)$ is associated to TW by a principal embedding $\mathfrak{su}(2) \subset \mathfrak{g}$. If the center $\mathcal{Z}(G^\vee)$ of $G^\vee$ is trivial, then the $G^\vee$ bundle $E \to W$ is completely characterized by this description of $\mathrm{ad}(E)$. Otherwise, if W is not simply-connected, the global description of E may involve some additional discrete data analogous to a choice of spin structure: the holonomies of E around noncontractible loops in W are not uniquely determined by the Riemannian structure of W, but can be modified by tensoring with a homomorphism $\pi_1(W) \to \mathcal{Z}(G^\vee)$.

3.5. The framing anomaly

3.5.1. A gravitational coupling. This last result presents us with a quandary. According to Section 3.2, the contribution of a given classical solution to the partition function is $\pm q^n$, where n is the instanton number of that solution. But the boundary conditions of Section 3.4 do not lead to a natural definition of the instanton number.

The instanton number of a $G^\vee$-bundle $E \to V$ is a topological invariant if V is a four-manifold without boundary. It remains a topological invariant if V has a non-empty boundary and we are given a trivialization of E on $W = \partial V$.

We have just discovered that instead of being trivialized on W, E is identified on W with the tangent bundle TW to W; the gauge field A restricted to W is similarly identified with the Riemannian connection ω on TW, or more precisely with its $G^\vee$-valued image $\xi(\omega)$, where $\xi\colon \mathfrak{su}(2) \to \mathfrak{g}^\vee$ is a principal embedding. This means that the instanton number P is not invariant under a change of metric of V. In general, under any change in the gauge field A, the change in P is given by the change in the Chern–Simons invariant of the restriction of A to the boundary W:

$$\delta \mathrm{P} = \frac{1}{2\pi}\delta \mathrm{CS}(A). \tag{3.27}$$

(This is the content of (2.52), for example.) Since when restricted to W we have $A = \xi(\omega)$, we can equivalently write

$$\delta \mathrm{P} = \frac{1}{2\pi}\delta \mathrm{CS}(\xi(\omega)). \tag{3.28}$$

In turn, $\mathrm{CS}(\xi(\omega))$ is (modulo the standard 2π ambiguity) the same as $\mathfrak{b}\,\mathrm{CS}(\omega)$ where $\mathrm{CS}(\omega)$ is the Chern–Simons invariant of ω as an SU(2) connection (before embedding it in $G^\vee$), and $\mathfrak{b}$ is an integer, analyzed in Section 3.5.3, that results from the embedding. So we can slightly simplify (3.28) to

$$\delta \mathrm{P} = \frac{\mathfrak{b}}{2\pi}\delta \mathrm{CS}(\omega). \tag{3.29}$$

If V is a compact manifold with boundary, there is a simple cure for this. We simply modify the definition (3.11) of P by subtracting the integral over V of a suitable curvature integral. The curvature integral is a multiple of $\int_V \mathrm{Tr}\, R \wedge R$, with R the Riemann tensor of V. This integral is a topological invariant if $\partial V = \varnothing$, and in general its variation is a multiple of $\delta\mathrm{CS}(\omega)$. We pick the coefficient to cancel the boundary term in the variation of P. Thus, we replace the definition (3.11) with

$$\widehat{\mathrm{P}} = \frac{1}{2h^\vee}\frac{1}{32\pi^2}\int_V \epsilon^{\mu\nu\alpha\beta}\,\mathrm{Tr}_{\mathrm{adj}}\, F_{\mu\nu}F_{\alpha\beta} - \frac{\mathfrak{b}}{4}\frac{1}{32\pi^2}\int_V \epsilon^{\mu\nu\alpha\beta}\,\mathrm{Tr}_{TV}\, R_{\mu\nu}R_{\alpha\beta}, \tag{3.30}$$

where we view the Riemann tensor as a two-form with values in endomorphisms of

the tangent bundle TV of V and take the trace accordingly.[16] With the boundary condition of Sections 3.3 and 3.4, $\widehat{\mathrm{P}}$ is an integer-valued topological invariant. The modification of P amounts to adding to the underlying Lagrangian a coupling of the gauge-theory theta-angle to $\mathrm{Tr}_{TV} R \wedge R$, in addition to its usual coupling to the gauge theory instanton density. If V has no boundary, this modification does not affect the topological invariance of the theory, while if V has a boundary, it eliminates the dependence on the Riemannian metric of the boundary.

3.5.2. The product case and the framing anomaly. What has just been described does not quite work if V is the noncompact four-manifold $W \times \mathbb{R}_+$ that will be essential in our applications. Let us discuss this case closely. We always assume a product metric on $W \times \mathbb{R}_+$; considering more general metrics does not add anything.

On $V = W \times \mathbb{R}_+$, we should first worry about a possible problem in defining P at infinity, as well as the problem at the boundary of V. At infinity on $\mathbb{R}_+$, we take a boundary condition that is given by a homomorphism $\rho^\vee\colon \pi_1(W) \to G^\vee_{\mathbb{C}}$ (as in the last paragraph of Section 3.2.1). Such a homomorphism is given by a complex-valued connection $\mathcal{A} = A + i\varphi$ that is independent of y. The complex-valued Chern–Simons invariant $\mathrm{CS}(\mathcal{A})$ is, of course, independent of the metric of W, and, given that $\mathcal{A}$ is flat, the real part of $\mathrm{CS}(\mathcal{A})$ coincides with $\mathrm{CS}(A)$. So $\mathrm{CS}(A)$ is independent of the metric of W. Hence varying the metric of W does not produce a contribution at infinity to the change in P; the only such contribution comes at $y = 0$, that is, at the boundary of V. Still, if $\rho^\vee$ is non-trivial, the constant value of $\mathrm{CS}(A)$ does represent a contribution to P. Because of this contribution as well as the contribution at $y = 0$, the values of P are not integers. However, differences in values of P continue to be integers.

We pause to explain this last important statement. The statement is clear if $G^\vee$ is simply-connected, for then any two bundles that obey the boundary conditions differ by a twist by an element of $\pi_3(G^\vee)$; as usual this twist shifts the instanton number by an integer. But even if $G^\vee$ is not simply-connected, differences in the values of P are still integers in the special case of $V = W \times \mathbb{R}_+$. Let us explain the reason for this for the case $G^\vee = \mathrm{SO}(3)$. In this case, a $G^\vee$ bundle $E \to V$ has an invariant $w_2(E) \in H^2(V, \mathbb{Z}_2)$, and if V is a compact four-manifold without boundary, the instanton number of the bundle E is congruent to[17] $\int_V w_2(E)^2/4 \bmod \mathbb{Z}$. This is

[16] If V is spin and we pick one of the spin bundles of V, say the bundle S_+ of spinors of positive chirality, then we can use in (3.30) a trace in S_+, rather than $1/4$ of a trace in TV. Even if V has a boundary, but assuming the metric is a product near the boundary, the two formulas differ by a topological invariant, a multiple of the Euler characteristic of V.

[17] This is a standard topological result. First, let us explain why $\int_V w_2(E)^2$ can be evaluated mod 4 even though $w_2(E)$ is defined only mod 2. For simplicity, we make a very mild assumption that $W_3(M) = 0$, which implies that $w_2(E)$ can be lifted to a class $x \in H^2(M, \mathbb{Z})$. Though x is only uniquely determined mod 2, $\int_M x^2$ is well defined mod 4. This is so simply because $(x + 2y)^2 = x^2 + 4(xy + y^2)$ so $\int_M x^2$ is invariant mod 4 under $x \to x + 2y$. So $\frac{1}{4}\int_M w_2(E)^2$ is well defined mod $\mathbb{Z}$. Now we wish to show that this number coincides with the instanton number of E mod $\mathbb{Z}$. By obstruction theory, this is true for all $\mathrm{SO}(3)$ bundles E with a given value of $w_2(E)$ if it is true for one such bundle. (The basic

why, potentially, values of P might not differ by integers. However, for $V = W \times \mathbb{R}_+$, our boundary condition at $y = 0$ says that $E|_W = TW$, and hence (as any oriented three-manifold is spin), the restriction of $w_2(E)$ to W vanishes. Since $V = W \times \mathbb{R}_+$ is contractible onto W, this ensures that $w_2(E)$ vanishes altogether, so the $G^\vee$ bundle E is liftable to a $\widehat{G}^\vee$ bundle, where $\widehat{G}^\vee = \mathrm{SU}(2)$ is the universal cover of $G^\vee$. This being so, we can replace $G^\vee$ by $\widehat{G}^\vee$ in analyzing the possible values of P, and these differ by integers just as if $G^\vee$ is simply connected. For any $G^\vee$, the argument proceeds in the same way, using the boundary condition at $y = 0$ to show that E can be lifted to a bundle with structure group $\widehat{G}^\vee$.

We still have to face the metric dependence of P that comes from the behavior at $y = 0$. On $V = W \times \mathbb{R}_+$, we cannot eliminate the metric-dependence of P by subtracting a curvature integral, as above. For a product metric on V, the integral $\int_V \mathrm{Tr}\, R \wedge R$ vanishes. If we use a more general metric, adding such a term would merely move the problem from $y = 0$ to $y = \infty$. Instead, we will have to proceed as in [102], where a precisely analogous problem arose in analyzing Chern–Simons theory on a three-manifold W.

If $\mathrm{CS}(\omega)$, the Chern–Simons function of the spin connection, were a well-defined real-valued function, we could eliminate the problem by subtracting from P a multiple of this function to define

$$\widehat{\mathrm{P}} = \mathrm{P} - \frac{\mathfrak{b}}{2\pi}\mathrm{CS}(\omega). \tag{3.31}$$

$\widehat{\mathrm{P}}$ would then be an integer-valued topological invariant that we would use instead of P in the formula for the partition function.

Actually, $\mathrm{CS}(\omega)$ has the usual 2π ambiguity, and is not well defined as a real-valued function unless we are given more information. The additional information we need is known as a "framing," a trivialization (up to homotopy) of the bundle in question. We have defined $\mathrm{CS}(\omega)$ as the Chern–Simons invariant of the Riemannian connection regarded as an $\mathrm{SU}(2)$ connection on the spin bundle, so the information we need to define $\mathrm{CS}(\omega)$ as a real-valued function is a framing of the spin bundle. Actually, we will proceed in a slightly different way. $\mathrm{CS}(\omega)$ has a dependence on the choice of spin structure of W, and this is unnatural in our problem (unless $G^\vee$ is such that the boundary condition of Section 3.3 entails a choice of spin structure). Although $\mathrm{CS}(\omega)$ depends on the spin structure, its variation in a change in metric does not (the dependence of $\mathrm{CS}(\omega)$ on the spin structure is a topological invariant); this is why (3.29) for the metric dependence of P does not depend on a spin structure. In redefining P to eliminate its metric-dependence, we want to avoid introducing an unnatural dependence on spin structure; we can accomplish this by simply rewriting (3.31) in terms of the Chern–Simons invariant of the Riemannian connection ω

idea here is that any two such bundles differ by a twist by $\pi_3(\mathrm{SO}(3)) = \mathbb{Z}$, and such a twist shifts the instanton number by an integer.) So it suffices to consider a convenient choice of E. For such a choice, let $\mathscr{L}$ be a complex line bundle with $c_1(\mathscr{L}) = w_2(E)$ mod 2, and let $E = \mathbb{R} \oplus \mathscr{L}$ where $\mathbb{R}$ is a trivial real line bundle and $\mathscr{L}$ is viewed as a real bundle of rank 2. Then $w_2(E) = c_1(\mathscr{L})$ mod 2 and the instanton number of E is $\frac{1}{4}\int_M c_1(\mathscr{L})^2$.

regarded as an SO(3) connection on TW, the tangent bundle of W. In [102], the Chern–Simons invariant of ω as an SO(3) connection was called $\mathrm{CS}_{\mathrm{grav}}$. The relation between the $\mathrm{CS}(\omega)$ and $\mathrm{CS}_{\mathrm{grav}}$ is simply

$$\mathrm{CS}_{\mathrm{grav}} = 4\,\mathrm{CS}(\omega). \tag{3.32}$$

The factor of 4 reflects the fact that the trace of a product of Lie algebra elements (such as $F \wedge F$) in the three-dimensional representation of SO(3) is four times the trace of the same product in the two-dimensional representation of SU(2). To define $\mathrm{CS}_{\mathrm{grav}}$ as a real-valued function, the topological data that we need is a framing of the tangent bundle TW. This is usually called simply a *framing of* W.

Given a framing, $\mathrm{CS}_{\mathrm{grav}}$ becomes a well-defined real-valued function, and we eliminate the metric-dependence of P by defining, as in (3.31):

$$\widehat{\mathrm{P}} = \mathrm{P} - \frac{\mathfrak{b}}{2\pi}\mathrm{CS}(\omega) = \mathrm{P} - \frac{\mathfrak{b}}{8\pi}\,\mathrm{CS}_{\mathrm{grav}}. \tag{3.33}$$

The quantity $\widehat{\mathrm{P}}$ is an invariant, valued in a coset of $\mathbb{Z}$ in $\mathbb{R}$ that depends on the choice of $\rho^\vee$ at infinity and on the framing but not on the metric of W.

Replacing P by $\widehat{\mathrm{P}}$ introduces in the partition function Z an extra factor

$$q^{-\mathfrak{b}\,\mathrm{CS}(\omega)/2\pi} = q^{-\mathfrak{b}\,\mathrm{CS}_{\mathrm{grav}}/8\pi}. \tag{3.34}$$

Under a unit change of framing, with $\mathrm{CS}_{\mathrm{grav}} \to \mathrm{CS}_{\mathrm{grav}} + 2\pi$, $\widehat{\mathrm{P}}$ as defined in (3.33) maps to $\widehat{\mathrm{P}} - v/4$. So under a unit change of framing, the partition function transforms by

$$Z \to Z q^{-\mathfrak{b}/4}. \tag{3.35}$$

Precisely such a dependence on a choice of framing appears in Chern–Simons theory. In Section 3.5.3, we will compare the framing anomaly as we have computed it in (3.35) in $\mathcal{N} = 4$ super Yang–Mills theory to the standard framing anomaly as found in Chern–Simons theory.

The relation of what has just been said to the treatment in Section 3.5.1 is that if one is given a compact V with boundary W, then the curvature integral on V gives a natural lift of $\mathrm{CS}_{\mathrm{grav}}$ (or $\mathrm{CS}(\omega)$) to a real-valued function. On $V = W \times \mathbb{R}_+$, there is no natural lift and we simply have to pick one.

Actually, something slightly less than a framing of TW is enough. In comparing two framings of TW, one runs into an integer winding number, associated with $\pi_3(\mathrm{SO}(3)) = \mathbb{Z}$, and, depending on the topology of W, one also encounters some two-torsion information derived from $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$. The two-torsion information is not relevant for the framing anomaly of Chern–Simons theory. There is a convenient way to eliminate it [6]. Two framings of TW that induce the same framing of $TW \oplus TW$ lead to the same definition of $\mathrm{CS}_{\mathrm{grav}}$. One can therefore consider the basic concept needed to define $\mathrm{CS}_{\mathrm{grav}}$ to be a framing of $TW \oplus TW$. A framing of $TW \oplus TW$ is called a *two-framing*. Globally, by making use of the signature theorem

on a four-manifold with boundary, one can define a canonical two-framing for any three-manifold W. This canonical two-framing is often used, explicitly or otherwise, in writing formulas for the Chern–Simons partition function. Because there is no local recipe for constructing it, it is natural to allow any framing (or two-framing) and determine how the partition function changes in a change of framing.

3.5.3. Comparison with Chern–Simons theory. According to [102], the framing dependence of Chern–Simons theory on a three-manifold W arises from the fact that to cancel an anomalous dependence of the partition function Z on the metric of W, we must pick a framing of W and include in the definition of Z a factor

$$\exp\Big(\frac{ic(k)\mathrm{sign}(k)\,\mathrm{CS}_{\mathrm{grav}}}{24}\Big). \tag{3.36}$$

Here $c(k)$ is the central charge of G current algebra at level $|k|$:

$$c(k) = \frac{k\ \dim(G)}{k + h\,\mathrm{sign}(k)}, \tag{3.37}$$

where $\dim(G)$ is the dimension of the gauge group G and h is its dual Coxeter number. Both equations (3.36) and (3.37) are usually written for $k > 0$; we have included factors of $\mathrm{sign}(k)$ so that they are valid for any nonzero integer k. (The required factors are determined by the fact that the partition function is invariant under $k \to -k$ together with a reversal of the orientation of W, which changes the sign of $\mathrm{CS}_{\mathrm{grav}}$.)

It is convenient to expand

$$c(k) = \dim(G) - \frac{h\ \dim(G)\,\mathrm{sign}(k)}{k + h\,\mathrm{sign}(k)}. \tag{3.38}$$

Here the first term, $\dim(G)$, arises in the one-loop approximation to Chern–Simons theory. In fact, it comes from the metric-dependence of an Atiyah–Patodi–Singer η-invariant, as explained in [102]. When inserted in (3.36), this term gives a factor

$$\exp(i\ \dim(G)\ \mathrm{sign}(k)\,\mathrm{CS}_{\mathrm{grav}}/24). \tag{3.39}$$

This factor is not analytic in k or q and hence will not match any computation in $\mathcal{N} = 4$ super Yang–Mills theory.

Instead, we interpret this factor as part of the constant $\mathfrak{N}_0$ in the relation (2.66) between two different holomorphic volume forms on the space of complex-valued connections. One of these, which we denote as DA, arises by analytic continuation of the path integral measure of Chern–Simons theory (with a compact gauge group G), while the second, which we denote as $D\mathcal{A}$, is induced from $\mathcal{N} = 4$ super Yang–Mills theory (together with a boundary condition defined by a flat connection $\mathcal{A}_\infty$ at

$y = \infty$, associated with some homomorphism $\rho\colon \pi_1(W) \to G_{\mathbb{C}}$). If what we have just found were a complete formula for $\mathfrak{N}_0$, we would have

$$DA \cong D\mathcal{A} \exp(-ih \operatorname{sign}(k) \operatorname{CS}(\mathcal{A}) + i \dim(G) \operatorname{sign}(k) \operatorname{CS}_{\text{grav}}/24). \tag{3.40}$$

Unfortunately, this cannot quite be a complete formula. Because of the factor of $1/24$ multiplying $\operatorname{CS}_{\text{grav}}$, the formula actually leaves unspecified a 24^{th} root of unity in the relation between DA and $D\mathcal{A}$. There is actually yet another root of unity that should be included; this is a fourth root of unity that arises on the Chern–Simons side from a spectral flow invariant that is described in [32]. It seems that $\mathfrak{N}_0$ depends on ρ, at least by these roots of unity, as well as on the metric of W. The factor involving $\operatorname{CS}_{\text{grav}}$ and the roots of unity all come from the η invariant which arises in the one-loop approximation to Chern–Simons theory evaluated at the flat connection $\mathcal{A}_\infty$. Perhaps $\mathfrak{N}_0$ should simply be written in terms of this η-invariant. Luckily, in this paper we mostly take $W = \mathbb{R}^3$ and $\mathcal{A}_\infty = 0$, enabling us to avoid these issues.

The higher order terms turn out to have a more clear-cut interpretation. We write $c(k) = \dim(G) + \Delta c$, where $\Delta c = -h \operatorname{sign}(k) \dim(G)/(k + h \operatorname{sign}(k))$ is the part of $c(k)$ that in Chern–Simons theory comes from diagrams of two or more loops. The natural perturbative expansion in Chern–Simons theory is in powers of $1/k$; Δc has contributions of all orders in this expansion. On the other hand, in $\mathcal{N} = 4$ super Yang–Mills theory, the natural expansion parameter is $1/\Psi$, where $\Psi = k + h \operatorname{sign}(k)$, so in this expansion, Δc is purely a two-loop effect. This fact remains to be explained.

In any case, the framing anomaly associated to Δc has a straightforward interpretation in the S-dual description by $G^\vee$ gauge theory. The part of (3.36) involving Δc is $\exp(-ih \dim(G) \operatorname{CS}_{\text{grav}}/24(k + h \operatorname{sign}(k)))$. Under an elementary change of framing $\operatorname{CS}_{\text{grav}} \to \operatorname{CS}_{\text{grav}} + 2\pi$, this factor changes by

$$\exp\Big(-\frac{2\pi i h \dim(G)}{24(k + h \operatorname{sign}(k))}\Big) = q^{-h \dim(G) \mathfrak{n}_{\mathfrak{g}}/24}, \tag{3.41}$$

where q was defined in (3.12). For the S-dual description, the equivalent formula (3.35) says that in an elementary change of framing, the partition function changes by a factor of $q^{-\mathfrak{b}/4}$. So obviously to reconcile the two formulas, we need $\mathfrak{b} = \mathfrak{n}_{\mathfrak{g}}\, h \dim(G)/6$.

So let us evaluate $\mathfrak{b}$. We start with an SU(2) gauge field A of instanton number 1. Such a gauge field has the property that if $\operatorname{Tr}_{\mathfrak{su}(2)}$ is the trace in the adjoint representation of SU(2), then

$$1 = \frac{1}{2 \cdot 2} \cdot \frac{1}{32\pi^2} \int_V \epsilon^{\mu\nu\alpha\beta} \operatorname{Tr}_{\mathfrak{su}(2)} F_{\mu\nu} F_{\alpha\beta}. \tag{3.42}$$

In the denominator, we have replaced $2h^\vee$ in the definition of the instanton number by $2 \cdot 2$, since $h^\vee = 2$ for SU(2). Now $\mathfrak{b}$ is defined as the instanton number of the $G^\vee$ gauge field $\xi(A)$, where ξ is a principal embedding $\mathfrak{su}(2) \to \mathfrak{g}$. Hence

$$\mathfrak{b} = \frac{1}{2 \cdot h^\vee} \frac{1}{32\pi^2} \int_V \epsilon^{\mu\nu\alpha\beta} \operatorname{Tr}_{\mathfrak{g}} \xi(F_{\mu\nu}) \xi(F_{\alpha\beta}). \tag{3.43}$$

The trace is now taken in the adjoint representation of $G^\vee$, and to be pedantic, we have written $\xi(F)$ for the $\mathfrak{g}$-valued image of F. The ratio of traces in (3.43) and (3.42) is the same as the ratio of the traces of the quadratic Casimir operator of $\mathfrak{su}(2)$ in the two representations (namely $\mathfrak{g}$ and $\mathfrak{su}(2)$). The value of the Casimir operator in an irreducible representation of $\mathfrak{su}(2)$ of dimension $2j+1$ is $j(j+1)$, and its trace is $j(j+1)(2j+1)$. So the ratio of the two traces is $\sum_{i=1}^r j_i(j_i+1)(2j_i+1)/6$, where (as discussed at the end of Section 3.3) $\mathfrak{g}$ is a direct sum of $\mathfrak{su}(2)$ modules of dimensions $2j_i+1$. So finally

$$\mathfrak{b} = \sum_{i=1}^{r} \frac{j_i(j_i+1)(2j_i+1)}{3h^\vee}\,. \tag{3.44}$$

The desired relation $\mathfrak{b} = \mathfrak{n}_\mathfrak{g}\,\dim(G)\,h/6$ hence becomes

$$\sum_{i=1}^{r} j_i(j_i+1)(2j_i+1) = \frac{1}{2}\mathfrak{n}_\mathfrak{g}\,\dim(G)\,hh^\vee. \tag{3.45}$$

As a check, this relation holds for G if and only if it holds for $G^\vee$. Indeed, the j_i, $\mathfrak{n}_\mathfrak{g}$, and $\dim G$ are invariant under the exchange $G \leftrightarrow G^\vee$, while h and $h^\vee$ are exchanged.

For a proof of this relation, see [83], Proposition 3.1. It is actually not difficult to verify the relation by hand for all simple Lie groups, whether of type A, B, C, D, E, F, or G. As an example, if G and therefore also $G^\vee$ are of type G_2, then the j_i are 1 and 5, while $\mathfrak{n}_g = 3$, $\dim(G) = 14$, and $h = h^\vee = 4$. The left and right of (3.45) both equal 336.

3.6. 't Hooft operators in the boundary

3.6.1. Preliminaries. In Section 2.2.4, we have shown that, when the gauge theory theta-angle is nonzero, the D3–NS5 system admits supersymmetric Wilson line operators at, and only at, the boundary of a four-manifold V. Dually, the same must be true for the D3–D5 system, but now with supersymmetric 't Hooft operators rather than Wilson operators. Our goal in the present section will be to concretely explain how to define these 't Hooft operators.

In general, 't Hooft operators are analogous to disorder operators in statistical mechanics – and also analogous to the D3–D5 boundary condition that we have described in Section 3.3. Just as our boundary condition was described by specifying the singularity that fields must have along the boundary of V, so an 't Hooft operator is defined, as explained in [65], by describing the singular behavior that four-dimensional fields should have along a chosen one-manifold S, which usually is taken to lie in the interior of V. To explain what singular behavior one wants, one selects a local model solution of the supersymmetric Yang–Mills equations on $\mathbb{R}^4\backslash\mathbb{R}$ (i.e., $\mathbb{R}^4$ with $\mathbb{R}$ removed) with a singularity of some desired type along $\mathbb{R}$. Normally, one picks a solution that is invariant under rotations and translations (and possibly

conformal motions) of $\mathbb{R}^4$ that map $\mathbb{R}$ to itself, and possibly under some supersymmetries. Concretely, for the usual half-BPS 't Hooft operators, the requisite singular solutions are very simple: they are obtained by embedding an abelian Dirac monopole in the nonabelian Yang–Mills gauge group. Once a singularity type is chosen, one calculates in the presence of a 't Hooft operator supported on a one-manifold $S \subset V$ by doing gauge theory on $V \backslash S$ with fields that have a singularity along S of the chosen type.

In our problem, we want to follow the same general ideas, with one important difference: V is a four-manifold with boundary W, and S is contained in W. (We expect from duality that S must be contained in W, but we can also see this directly by following the analysis of Wilson–'t Hooft operators in Section 6.2 of [67].[18]) But the basic idea of defining an 't Hooft operator by specifying a model solution still applies.

For the model solution, we now take V to be a half-space, say the space $x^3 \geq 0$ in a Euclidean space with coordinates $x^0, \dots, x^3$. And we take S to be a straight line in the boundary of V, say the line $x^1 = x^2 = x^3 = 0$. We look for a solution of the Yang–Mills equations on V that is invariant under symmetries that map S to itself, that is, under translations of x^0 and rotations of the $x^1 - x^2$ plane. In addition, as we want an 't Hooft operator that preserves the supersymmetry Q of our topological field theory, the singular solution should obey the supersymmetric equations (3.6). (Actually our 't Hooft operator will preserve more supersymmetry than just the one supercharge Q, which it will accomplish by obeying a stronger system of equations, as described later.) The solution should become trivial for $x^3 \to \infty$, far from the position of the 't Hooft operator. At a generic boundary point, it must have the boundary behavior of the regular Nahm pole as described in Section 3.3. This in particular means that the desired singular solution cannot be a simple abelian one, like the singular solution used to describe an 't Hooft operator away from the boundary. At a boundary point that is located on the line S, the singular behavior is more complicated. That more complicated behavior is exactly what we wish to determine.

We will carry out this program in full for $G = \mathrm{SU}(2)$. For G of higher rank, we carry out some of the steps but the precise singular solution of relevance is not yet known.

3.6.2. First reduction of the equations. As just explained, we want to find on the half-space V given by $x^3 \geq 0$ a special type of solution of the supersymmetric equations

$$F - \varphi \wedge \varphi + \star \mathrm{d}_A \varphi = 0 = \mathrm{d}_A \star \varphi. \tag{3.46}$$

[18] It is shown there that 't Hooft operators away from the boundary preserve the topological symmetry only if $\Psi = 0$. It is also shown, however, that for any rational value of Ψ, there are combined Wilson–'t Hooft operators in bulk (as one would expect from S-duality). These are undoubtedly important for understanding special properties of the theory at rational values of Ψ.

The solution should be invariant under translations in x^0, should become trivial for $x^3 \to \infty$, and away from the line S given by $x^1 = x^2 = x^3 = 0$, its boundary behavior should coincide with the regular Nahm pole described in Section 3.3.

A drastic simplification comes from the fact that in solving the equations, we can set $A_0 = \varphi_3 = 0$. The reader may choose to view this as a lucky ansatz that can be used to simplify the equations. However, there are also several ways to predict *a priori* that the solution we want has $A_0 = \varphi_3 = 0$. For one thing, one can use a vanishing argument similar to that discussed in (4.13) of [108] to prove that a solution on V with the desired asymptotic behavior has $A_0 = \varphi_3 = 0$. (The proof is standard: one squares the equations (3.46), integrates over V, and then integrates by parts, showing that in any solution, A_0 and φ_3 are annihilated by strictly positive linear differential operators.) Alternatively, one can use supersymmetry. Obeying (3.46) ensures invariance under one supersymmetry, but duality with the boundary Wilson lines studied in Section 2.2.4 indicates that the 't Hooft operators of interest should preserve four global supercharges (half of the supercharges preserved by the half-BPS boundary condition). The extra supersymmetry puts additional constraints on the solution, leading to the structure that we describe momentarily.

The equations obtained from (3.46) after setting $A_0 = \varphi_3 = 0$ can be described as follows. Define the three operators

$$\begin{aligned} \mathcal{D}_1 &= \frac{D}{Dx^1} + i\frac{D}{Dx^2} = \frac{\partial}{\partial x^1} + i\frac{\partial}{\partial x^2} + [A_1 + iA_2, \cdot], \\ \mathcal{D}_2 &= D_3 - i[\varphi_0, \cdot]] = \frac{\partial}{\partial x^3} + [A_3 - i\varphi_0, \cdot], \\ \mathcal{D}_3 &= [\varphi_1 - i\varphi_2, \cdot]. \end{aligned} \tag{3.47}$$

Thus, $\mathcal{D}_1$ and $\mathcal{D}_2$ are first order differential operators, while $\mathcal{D}_3$ is of order zero. In (3.47), for an adjoint-valued field Λ, the symbol $[\Lambda, \cdot]]$ represents the commutator with Λ.

With this understood, the equations (3.46) take the form

$$[\mathcal{D}_i, \mathcal{D}_j] = 0, \quad i, j = 1, \dots, 3 \tag{3.48}$$

together with

$$\sum_{i=1}^{3} [\mathcal{D}_i, \mathcal{D}_i^\dagger] = 0. \tag{3.49}$$

Here $\mathcal{D}_i^\dagger$ is the adjoint of the differential operator $\mathcal{D}_i$. Concretely, (3.49) takes the form

$$F_{12} - [\varphi_1, \varphi_2] - D_3\varphi_0 = 0. \tag{3.50}$$

To similarly make (3.48) explicit is immediate from the definitions of the $\mathcal{D}_i$.

Before trying to understand these equations, let us describe some special cases. If we set $A_1 = A_2 = 0$ and take the fields to be independent of x^1 and x^2, we get

Nahm's equations. If we set $A_3 = \varphi_0 = 0$ and take the fields to be independent of $y = x^3$, we get Hitchin's equations. Finally, if we set $\varphi_1 = \varphi_2 = 0$, we get the Bogomolny equations. So our system is a hybrid of all those equations. This hybrid was encountered in [67] and called the extended Bogomolny equations (see (10.36) of that paper, where the equations are written in the gauge $A_y = 0$). The main interest there was the role in these equations of 't Hooft operators in the bulk (and their interpretation in terms of Hecke modifications of Higgs bundles). Our concern here will instead be the more subtle case of 't Hooft operators in the boundary.

It is also helpful to consider some analogous equations. For an interesting analogy, consider gauge theory of a connection A on $\mathbb{R}^6 \cong \mathbb{C}^3$. We endow $\mathbb{C}^3$ with complex coordinates z^i, $i = 1, \dots, 3$, and define

$$\mathcal{D}_i = \frac{\partial}{\partial \overline{z^i}} + A_{\bar{i}}. \tag{3.51}$$

In other words, the $(0,1)$ part of the connection is $\sum_i \mathrm{d}\overline{z^i}\,\mathcal{D}_i$. The equations $[\mathcal{D}_i, \mathcal{D}_j] = 0$ assert that the $(0,2)$ part of the curvature vanishes, so that the connection defines a holomorphic bundle, while the remaining equation $\sum_i [\mathcal{D}_i, \mathcal{D}_i^\dagger] = 0$ can be solved only if the holomorphic bundle is semi-stable, and, according to a theorem of Donaldson and of Uhlenbeck and Yau, it has a unique solution in that case. The combined equations are known as the *hermitian Yang–Mills equations,* and can be formulated on a general complex manifold, not necessarily $\mathbb{C}^3$. Physically, the hermitian Yang–Mills equations are familiar in the context of the heterotic string on a Calabi–Yau threefold. In that context, solutions of those equations preserve four supercharges, and the same is true for the equations (3.48) and (3.49), though we will not demonstrate this here.

As in the other cases that we have just mentioned, the key to understanding (3.48) and (3.49) is to first observe that the equations (3.48) have a larger gauge symmetry than the full system. The full system of equations is invariant under an ordinary gauge transformation

$$\mathcal{D}_i \longrightarrow g\mathcal{D}_i g^{-1}, \quad i = 1, \dots, 3, \tag{3.52}$$

where g is $G^\vee$-valued. But (3.48), since they involve only the operators $\mathcal{D}_i$ and not their adjoints, are invariant under complex-valued gauge transformations, that is gauge transformations in which we allow g to be valued in $G^\vee_{\mathbb{C}}$, the complexification of $G^\vee$. The space of solutions of (3.48), modulo complex-valued gauge transformations, is naturally a complex manifold. In all the problems that we have mentioned – including Nahm's equations, Hitchin's equations, the Bogomolny equations, the hermitian Yang–Mills equations, and also our present problem – the remaining equation (3.49) can be interpreted as an equation for vanishing of the moment map. In other words, in each case, one can define a symplectic structure on the space of fields such that the moment map for the action of the compact gauge group ($G^\vee$ in our problem) is the left hand side of (3.49). One then aims to compare (i) the space of solutions of the full system of equations, modulo $G^\vee$-valued gauge transformations, to (ii) the

solutions of the holomorphic equations modulo $G_{\mathbb{C}}^{\vee}$-valued gauge transformations. Typically, one aims to show (as in the result of Donaldson and Uhlenbeck–Yau concerning the hermitian Yang–Mills equations) that (i) and (ii) coincide after correcting (ii) to incorporate a certain condition of stability. In our present problem, the desired boundary condition at $y = 0$ ensures that the gauge group acts freely on the space of solutions, and one may hope that in a proper formulation – which will have to take into account the boundary behavior in an essential way – (i) and (ii) – will simply coincide.

3.6.3. The holomorphic data. The holomorphic data in this problem are easily described. Since a holomorphic $G_{\mathbb{C}}^{\vee}$-bundle over the complex z-plane is trivial, we can make a complex gauge transformation to go to a gauge in which $A_1 + iA_2 = 0$, so that $\mathcal{D}_1$ reduces to $\partial_1 + i\partial_2 = 2\partial_{\bar{z}}$. But actually, since $[\mathcal{D}_1, \mathcal{D}_2] = 0$, we can do better: we can make a complex gauge transformation setting $A_1 + iA_2 = A_3 - i\varphi_0 = 0$. In this gauge, $\mathcal{D}_1 = 2\,\partial/\partial\bar{z}$ and $\mathcal{D}_2 = \partial/\partial x^3$. The equations $[\mathcal{D}_1, \mathcal{D}_3] = [\mathcal{D}_2, \mathcal{D}_3] = 0$ then say that $\varphi = \varphi_1 - i\varphi_2$ is holomorphic in z and independent of $y = x^3$. We are still free to make a gauge transformation by a holomorphic map $g(z)\colon \mathbb{C} \to G_{\mathbb{C}}^{\vee}$.

In short, the holomorphic data consist of a $\mathfrak{g}_{\mathbb{C}}^{\vee}$-valued holomorphic function $\varphi(z)$, modulo conjugation by a $G_{\mathbb{C}}^{\vee}$-valued holomorphic function $g(z)$. What sort of function $\varphi(z)$ we should consider depends on what behavior we want at infinity. Let us remember that vacuum states of $\mathcal{N} = 4$ super Yang–Mills theory are specified by the asymptotic values of the scalar fields (which moreover must commute with each other to ensure the vanishing of the classical potential energy). In particular, a choice of vacuum state at infinity determines the conjugacy class of $\varphi = \varphi_1 - i\varphi_2$ at $y = \infty$. For the present paper, the most convenient vacuum to consider is the one in which the scalar fields simply vanish at infinity. So we will look for solutions of the extended Bogomolny equations in which $\varphi \to 0$ at infinity. In any event, the real interest in the present section is in the singular behavior of the solution near special boundary points where 't Hooft operators are inserted, and we do not care too much about what happens far away. For our immediate purposes, asking for φ to vanish at infinity is just a convenient auxiliary condition that will make it easier to find a solution with the singularity we want.

The equation $[\mathcal{D}_2, \mathcal{D}_3] = 0$ is equivalent to $\partial_3\varphi = -[A_3 - i\varphi_0, \varphi]$. It says that the x^3 derivative of φ is a commutator of φ with some matrix, so that the conjugacy class of φ is independent of $y = x^3$. It is not correct to conclude from this and the fact that φ vanishes at $y = \infty$ that φ is identically zero. The correct conclusion is only that φ is nilpotent. To prove nilpotency, let $\mathcal{P}$ be a homogeneous invariant polynomial of positive degree on the complex Lie algebra $\mathfrak{g}_{\mathbb{C}}$. Since the conjugacy class of φ is independent of y, we have $\partial_y \mathcal{P}(\varphi) = 0$. So if φ vanishes at infinity, then $\mathcal{P}(\varphi)$ vanishes for all y. An element $\varphi \in \mathfrak{g}_{\mathbb{C}}$ such that $\mathcal{P}(\varphi) = 0$ for all $\mathcal{P}$ of the assumed kind is nilpotent. So φ is nilpotent for all y (and z).

A simple example of a solution in which φ is everywhere nilpotent but not zero and

approaches zero at infinity is the basic Nahm pole solution (3.18) with $\vec{\varphi} = \vec{t}/y$, where $\vec{t}$ are images of a standard set of $\mathfrak{su}(2)$ generators under an embedding $\xi : \mathfrak{su}(2) \to \mathfrak{g}$. In this solution, $\varphi = (t_1 - it_2)/y$ is indeed nilpotent (it is a lowering operator with respect to t_0). Its conjugacy class is independent of y (this is proved by conjugating by t_0) and it vanishes for $y \to \infty$.

We are actually interested in the case that ξ is a principal embedding, which is equivalent to the condition that φ is a regular nilpotent element of $\mathfrak{g}_{\mathbb{C}}$. We pause to explain this concept. Every complex simple Lie algebra has a finite set of nilpotent conjugacy classes. For example, a nilpotent element $\varphi \in \mathfrak{sl}(n, \mathbb{C})$ can be conjugated to a Jordan canonical form in which all matrix elements vanish except just above the main diagonal:

$$\varphi = \begin{pmatrix} 0 & * & 0 & \dots & 0 \\ 0 & 0 & * & \dots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \dots & * \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}, \tag{3.53}$$

and moreover the matrix elements just above the main diagonal are all 1 or 0. The conjugacy classes of nilpotent elements of $\mathfrak{sl}(n, \mathbb{C})$ are classified by the pattern of 1's and 0's, up to obvious permutations of blocks. An element of a complex Lie algebra $\mathfrak{g}_{\mathbb{C}}$ is called *regular* if the subalgebra of $\mathfrak{g}_{\mathbb{C}}$ that commutes with it is as small as possible, that is if its dimension equals r, the rank of the algebra. There is always a unique nilpotent conjugacy class of maximal dimension, known as *the regular nilpotent conjugacy class.* This is the class containing the raising and lowering operators for a principal $\mathfrak{su}(2)$ subalgebra. For $\mathfrak{sl}(n, \mathbb{C})$, the regular nilpotent conjugacy class is the one with a single Jordan block (all elements labeled $*$ in (3.53) actually equal 1). A generic nilpotent element is contained in this regular nilpotent conjugacy class. In particular, in the solution associated to the principal $\mathfrak{su}(2)$ embedding, φ is a regular nilpotent element.

Finally, we can describe the solutions that are relevant for boundary 't Hooft operators. We look for a solution in which $\varphi(z)$ is holomorphic in z and everywhere nilpotent. Moreover, for a generic value of z, the behavior for $y \to 0$ must coincide with the model solution (3.18), so φ is a regular nilpotent. At isolated points $z = z_j$, $j = 1, \dots, s$, φ is in a more special nilpotent conjugacy class. These are the points at which 't Hooft operators are inserted.

For example, for the case that $G^\vee = \mathrm{SU}(2)$, any everywhere nilpotent $\varphi(z)$ is conjugate to

$$\varphi(z) = \begin{pmatrix} 0 & f(z) \\ 0 & 0 \end{pmatrix}, \tag{3.54}$$

for some holomorphic function $f(z)$. Only the zeroes of f and the degrees of their zeroes have an invariant meaning, since where $f(z)$ is not zero, we can set

$\varphi = g\varphi_1 g^{-1}$, with

$$\varphi_1 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \tag{3.55}$$

and

$$g(z) = \begin{pmatrix} f(z)^{1/2} & 0 \\ 0 & f(z)^{-1/2} \end{pmatrix}. \tag{3.56}$$

The case of a single 't Hooft operator is the case that the function $f(z)$ has only one zero, say of order $\mathfrak{r}$:

$$\varphi = \begin{pmatrix} 0 & z^{\mathfrak{r}} \\ 0 & 0 \end{pmatrix} \tag{3.57}$$

In Section 3.6.4, we will find for each positive integer $\mathfrak{r}$ a unique solution of the extended Bogomolny equations with this φ and the appropriate asymptotic behavior at the boundary $y = 0$ and at infinity.

For a more systematic explanation of the above formula, let us recall that GNO or Langlands duality associates to a representation of G a dual magnetic weight of $G^\vee$. This magnetic weight is a conjugacy class of homomorphisms from $\mathbb{C}^*$ to $G^\vee_{\mathbb{C}}$. For $G = \mathrm{SO}(3)$, the homomorphism to $G^\vee_{\mathbb{C}} = SL(2,\mathbb{C})$ associated to the spin j representation of G is

$$z \longrightarrow g(z) = \begin{pmatrix} z^j & 0 \\ 0 & z^{-j} \end{pmatrix}. \tag{3.58}$$

For $G = \mathrm{SU}(2)$, j may be half-integral and then the formula should be written in the spin 1 representation; $g(z)$ is well defined as a homomorphism from $\mathbb{C}^*$ to $G^\vee_{\mathbb{C}} = \mathrm{SO}(3)_{\mathbb{C}}$. In all cases, the relation between φ and g is $\varphi = g\varphi_1 g^{-1}$, so that in the notation of (3.57), $\mathfrak{r} = 2j$.

The analog of this for $G = \mathrm{SU}(n)$ is hopefully clear. Instead of (3.57), we look for a solution with

$$\varphi = \begin{pmatrix} 0 & z^{\mathfrak{r}_1} & 0 & \dots & 0 \\ 0 & 0 & z^{\mathfrak{r}_2} & \dots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \dots & z^{\mathfrak{r}_{n-1}} \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}, \tag{3.59}$$

where the $\mathfrak{r}_i$ are non-negative integers, not all zero, representing the highest weight of a representation of G. More generally, for any $G^\vee$, the corresponding formula is obtained as follows. Pick a principal $\mathfrak{su}(2)$ embedding and within it a Cartan subalgebra. Relative to this choice, let φ_1 be a raising operator of the chosen $\mathfrak{su}(2)$ subalgebra, and let $T^\vee_{\mathbb{C}}$ be the maximal torus of $G^\vee_{\mathbb{C}}$ that commutes with the chosen Cartan subalgebra of $\mathfrak{su}(2)$. Pick a homomorphism $g(z)\colon \mathbb{C}^* \to T^\vee_{\mathbb{C}}$ such that $\varphi = g\varphi_1 g^{-1}$ has no pole at $z = 0$. The choices for $g(z)$ are in natural correspondence with the highest weights of G representations, and therefore with Wilson operators of G gauge

theory. By solving the extended Bogomolny equations with the corresponding φ and identifying the singular behavior at $y = z = 0$, we get our candidate for the definition of the boundary 't Hooft operator in $G^\vee$ gauge theory that is dual to a given Wilson operator of G.

In Section 3.6.4, we will explicitly find the relevant solutions of the extended Bogomolny equations for $G = \mathrm{SU}(2)$. For G of higher rank, this remains open.

3.6.4. Solving the equations for SU(2). Starting with the holomorphic data (3.57), with all other fields vanishing, we want to make a complex gauge transformation $\mathcal{D}_i \to g\mathcal{D}_i g^{-1}$ so as to obey the extended Bogomolny equations. Since the $\mathcal{D}_i$ will obey $[\mathcal{D}_i, \mathcal{D}_j] = 0$ for any choice of g, we really need only chose g to obey the remaining condition $\sum_i [\mathcal{D}_i, \mathcal{D}_i^\dagger] = 0$.

The extended Bogomolny equations are invariant under $\varphi \to e^{i\alpha}\varphi$ with α a real constant. The holomorphic data (3.57) are invariant under this symmetry, up to a diagonal gauge transformation. So it is natural to choose g so as to preserve the symmetry. This means that g must be diagonal:

$$g = \begin{pmatrix} e^{v/2} & 0 \\ 0 & e^{-v/2} \end{pmatrix}. \tag{3.60}$$

Moreover, using the invariance of the extended Bogomolny equations under unitary gauge transformations (those valued in $G^\vee$ rather than its complexification), we can take v to be real. After transforming $\mathcal{D}_i \to g\mathcal{D}_i g^{-1}$, we find

$$\begin{aligned} A_1 + iA_2 &= -\frac{(\partial_1 + i\partial_2)v}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \\ F_{12} &= \frac{i(\partial_1^2 + \partial_2^2)v}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \\ \varphi_0 &= -\frac{i\,\partial_3 v}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \\ \varphi &= z^{\mathfrak{r}} e^{v}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}. \end{aligned} \tag{3.61}$$

And finally, the "moment map" equation $\sum_i [\mathcal{D}_i, \mathcal{D}_i^\dagger] = 0$ becomes

$$-\Big(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial y^2}\Big)v + |z|^{2\mathfrak{r}} \exp(2v) = 0, \tag{3.62}$$

where we write y for x_3 and z for $x_1 + ix_2$.

This equation has the simple exact solution

$$v = -\mathfrak{r}\log|z| - \log y, \tag{3.63}$$

corresponding to

$$\varphi = \frac{(z/\bar{z})^{\mathfrak{r}/2}}{y}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}. \tag{3.64}$$

This solution is singular at $z = 0$, but the singularity can actually be removed by a unitary gauge transformation $\varphi \to h\varphi h^{-1}$ with

$$h = \begin{pmatrix} (z/\bar{z})^{-\mathfrak{r}/4} & 0 \\ 0 & (z/\bar{z})^{\mathfrak{r}/4} \end{pmatrix}. \tag{3.65}$$

After this gauge transformation, we arrive at the basic solution (3.18) in which the gauge field A vanishes while φ is $1/y$ times a raising operator. This is the solution that defines the boundary condition we want at boundary points with $z \neq 0$, that is, anywhere away from the insertion of the 't Hooft operator.

To describe an 't Hooft operator at the boundary, we want a solution with the same behavior as (3.63) for $y \to 0$ with $z \neq 0$, but regular along the open ray $z = 0$, $y \neq 0$. Exactly what will happen near $z = y = 0$ will be determined by the equations. That will be the answer to our question: the 't Hooft operator of charge $\mathfrak{r}$ will be defined by the singularity that the equation forces upon us at $z = y = 0$.

It is useful to make a small change of variables:

$$v = -(\mathfrak{r}+1)\log|z| + u. \tag{3.66}$$

The desired behavior of u is hence

$$\begin{cases} u \sim \log|z| - \log y & \text{for } y \to 0 \text{ with } z \neq 0, \\ u \sim (\mathfrak{r}+1)\log|z| & \text{for } z \to 0 \text{ with } y \neq 0. \end{cases} \tag{3.67}$$

(The second condition ensures that v is regular at $z = 0$, $y > 0$.) In terms of u, the equation becomes

$$-\Big(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial y^2}\Big)u + |z|^{-2}\exp(2u) = 0. \tag{3.68}$$

Writing the equation this way makes visible a scaling symmetry $z \to \lambda z$, $y \to \lambda y$. There is also an obvious symmetry of rotation of the z-plane.

It is natural to expect the fields produced by an 't Hooft operator at $y = z = 0$ to be scale-invariant and rotation-symmetric. For a rotation-symmetric solution, writing $r = |z|$, the equation becomes

$$-((r\partial_r)^2 + (r\partial_y)^2))u + \exp(2u) = 0. \tag{3.69}$$

Scale-invariance means that u is a function only of $s = r/y$. Acting on a function with this property, we can substitute $r\partial_r \to s\,\partial_s$, $r\partial_y \to -s^2\,\partial_s$, so the equation becomes

$$-\Big(\Big(s\frac{\mathrm{d}}{\mathrm{d}s}\Big)^2 + \Big(s^2\frac{\mathrm{d}}{\mathrm{d}s}\Big)^2\Big)u + e^{2u} = 0. \tag{3.70}$$

This equation can be neatly solved by transforming from s to another coordinate $\tau(s)$ with the property that

$$\left(s\frac{\mathrm{d}}{\mathrm{d}s}\right)^2 + \left(s^2\frac{\mathrm{d}}{\mathrm{d}s}\right)^2 = \frac{\mathrm{d}^2}{\mathrm{d}\tau^2}. \tag{3.71}$$

This equation is conveniently equivalent to

$$\left(\sqrt{s^2+s^4}\,\frac{\mathrm{d}}{\mathrm{d}s}\right)^2 = \frac{\mathrm{d}^2}{\mathrm{d}\tau^2}, \tag{3.72}$$

leading to

$$\frac{\mathrm{d}s}{\sqrt{s^2+s^4}} = \mathrm{d}\tau. \tag{3.73}$$

This equation can be integrated, but for the moment let us refrain from doing so. In terms of τ, our equation (3.70) becomes

$$\frac{\mathrm{d}^2u}{\mathrm{d}\tau^2} = \exp(2u). \tag{3.74}$$

This implies that

$$\frac{\mathrm{d}u}{\sqrt{e^{2u}+b^2}} = \pm\,\mathrm{d}\tau, \tag{3.75}$$

with an integration constant b^2. Setting

$$e^{u(\tau)} = b\,p(\tau), \tag{3.76}$$

we get

$$\frac{1}{b}\frac{\mathrm{d}p}{\sqrt{p^4+p^2}} = \pm\mathrm{d}\tau, \tag{3.77}$$

and comparing to (3.73), we see that we can eliminate τ:

$$\frac{1}{b}\frac{\mathrm{d}p}{\sqrt{p^4+p^2}} = \pm\frac{\mathrm{d}s}{\sqrt{s^4+s^2}}. \tag{3.78}$$

Using now the indefinite integral

$$\int \frac{\mathrm{d}t}{\sqrt{t^4+t^2}} = -\log\left(\frac{t}{\sqrt{1+t^2}-1}\right) + C, \tag{3.79}$$

we find that

$$\frac{p}{\sqrt{1+p^2}-1} = N\left(\frac{s}{\sqrt{1+s^2}-1}\right)^{\pm b}, \tag{3.80}$$

for a constant N. For $y \to 0$ with fixed $z \neq 0$, we have $s \to \infty$, and according to (3.67), we want $u \to \infty$ in this limit, and hence also $p \to \infty$. It then follows from (3.80) that we must set $N = 1$. Compatibility with (3.67) for $s \to 0$ (that is,

for $z \to 0$ with fixed $y \neq 0$) gives $b = \mathfrak{r} + 1$ (and also tells us to use the plus sign in the exponent in (3.80)). Taking these values and solving for p, we get

$$p(s) = \frac{2s^{\mathfrak{r}+1}}{(\sqrt{1+s^2}+1)^{\mathfrak{r}+1} - (\sqrt{1+s^2}-1)^{\mathfrak{r}+1}}. \tag{3.81}$$

The original variable $v(s)$ is

$$e^{v(s)} = \frac{(\mathfrak{r}+1)p(s)}{|z|^{\mathfrak{r}+1}}. \tag{3.82}$$

This is the solution in the presence of a single 't Hooft operator that is dual to a Wilson operator with $j = \mathfrak{r}/2$. More generally, the singularity of this solution at $y = z = 0$ defines what we mean by a boundary 't Hooft operator of this magnetic charge.

To understand the solution a little better, let us evaluate the gauge field on the boundary plane $y = 0$. From (3.82), we have $v = -\log y - \mathfrak{r} \log z + \text{constant} + \mathcal{O}(y)$, so from (3.61) we get

$$A_i = \frac{\epsilon_{ij} x_j}{x_1^2 + x_2^2} \frac{\mathfrak{r}}{2} \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} + \mathcal{O}(y). \tag{3.83}$$

This is a familiar type of two-dimensional U(1) gauge field, except that here it is embedded in SU(2). It describes a point vortex with $\mathfrak{r}/2$ magnetic flux quanta, located at $z = 0$. The gauge field is flat in the boundary, away from $z = 0$. The monodromy around the point $z = 0$ is

$$\begin{pmatrix} e^{i\pi\mathfrak{r}} & 0 \\ 0 & e^{-i\pi\mathfrak{r}} \end{pmatrix}. \tag{3.84}$$

As long as $\mathfrak{r}$ is an integer, the monodromy is ± 1, and in fact it is always 1 when regarded as an element of $G^\vee$. (We recall that odd $\mathfrak{r}$ corresponds to half-integral $j = \mathfrak{r}/2$, and hence to $G = \mathrm{SU}(2)$, $G^\vee = \mathrm{SO}(3)$.)

3.6.5. Solutions with a line singularity. In Section 6, we will actually want some additional solutions of the same equations that have a singularity not just at $z = y = 0$, but along the whole ray $z = 0$, $y \geq 0$. We denote this ray as ℓ.

Some new solutions correspond to the case $\mathfrak{r} = -1$ of the ansatz (3.61). Thus, the holomorphic data are given by $\varphi = g\varphi_1 g^{-1}$, with g as in (3.60) and

$$\varphi_1 = \begin{pmatrix} 0 & z^{-1} \\ 0 & 0 \end{pmatrix}. \tag{3.85}$$

For $\mathfrak{r} = -1$, v and u coincide. As for the asymptotic behavior of the solution, for $y \to 0$ or $s \to \infty$, we want the usual behavior

$$v \sim \log|z| - \log y = \log s, \quad s \longrightarrow \infty, \tag{3.86}$$

so as to agree at a generic point on the boundary with the usual solution with a regular Nahm pole. Along the line ℓ, we look first for a solution that is singular but less singular than $1/|z|$. For φ to be less singular than $1/|z|$ means that we need $v \to -\infty$ for $|z| \to 0$, but for A to be less singular than $1/|z|$ means that $|v|$ should diverge more slowly than $\log|z|$. These conditions force us to take $b = 0$, which is not a surprise since in general we had $b = \mathfrak{r} + 1$. For $b = 0$, the substitution (3.76) is not useful, but we can directly combine (3.75) and (3.73) to get (with $v = u$)

$$\frac{\mathrm{d}v}{e^v} = \frac{\mathrm{d}s}{\sqrt{s^2 + s^4}}. \tag{3.87}$$

Using (3.79) and adjusting the integration constant to match what we want for $s \to \infty$, we find the unique solution

$$e^v = \frac{1}{\log(s/(\sqrt{1+s^2} - 1))}. \tag{3.88}$$

A slightly more general solution in which we do not take $b = 0$ is also of interest. To find this solution, we simply combine (3.76) and (3.80). We set $v = u$ as we still assume $\mathfrak{r} = -1$, and we keep $N = 1$ to leave the behavior unchanged for $y \to 0$ or $s \to \infty$. The solution is

$$e^v = \frac{2bs^b}{(\sqrt{1+s^2}+1)^b - (\sqrt{1+s^2}-1)^b}. \tag{3.89}$$

The asymptotic behavior is

$$\begin{cases} v \sim \log s & \text{for } s \to \infty, \\ v \sim b \log s & \text{for } s \to 0. \end{cases} \tag{3.90}$$

The Nahm pole for $y \to 0$ or $s \to \infty$ is unchanged, and in particular, if we restrict to the boundary plane at $y = 0$, then the monodromy around the point $z = 0$ remains trivial (as an element of[19] $G^\vee = \mathrm{SO}(3)$), just as in (3.84). However, the singularity along ℓ at a point with $y > 0$ is controlled by the behavior for $z \to 0$ with fixed y, or in other words for $s \to 0$. This monodromy can be determined by the same computation that led to (3.84), simply replacing the behavior $v \sim -\mathfrak{r}\log|z|$ assumed there by $v \sim b\log|z|$. So the monodromy is

$$\begin{pmatrix} e^{-i\pi b} & 0 \\ 0 & e^{i\pi b} \end{pmatrix}. \tag{3.91}$$

[19] For $G^\vee = \mathrm{SU}(2)$, to make the monodromy in the boundary plane trivial, we modify the solution by twisting by a flat line bundle on the complement of ℓ whose monodromy around ℓ is -1. Differently put, we modify the solution by the gauge transformation (3.65), with $\mathfrak{r} = -1$.

For a further generalization, we continue to require that the singularity in the holomorphic data corresponds to a simple pole at $z = 0$, but we drop the assumption that φ is nilpotent. So we take $\varphi = g\varphi_1 g^{-1}$, with

$$\varphi_1 = \frac{\lambda}{z}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{3.92}$$

where λ is an arbitrary nonzero complex number. (Equivalently, we could take $\varphi_1 = M/z$, where M is any 2×2 matrix of determinant $-\lambda^2$, but then we would have to slightly alter the rest of the ansatz.) So

$$\varphi = g\varphi_1 g^{-1} = \frac{\lambda}{z}\begin{pmatrix} 0 & e^v \\ e^{-v} & 0 \end{pmatrix}. \tag{3.93}$$

Keeping the rest of the ansatz (3.61) unchanged, the equation (3.62) is replaced by

$$-\Big(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial y^2}\Big)v + \frac{|\lambda|^2}{|z|^2}\left(e^{2v} - e^{-2v}\right) = 0. \tag{3.94}$$

We assume that v is a function only of $s = |z|/y$ with

$$\begin{cases} v \sim \log s & \text{for } s \to \infty, \\ v \text{ bounded} & \text{for } s \to 0. \end{cases} \tag{3.95}$$

Equation (3.70) is replaced by

$$-\Big(\Big(s\frac{\mathrm{d}}{\mathrm{d}s}\Big)^2 + \Big(s^2\frac{\mathrm{d}}{\mathrm{d}s}\Big)^2\Big)v + |\lambda|^2(e^{2v} - e^{-2v}) = 0. \tag{3.96}$$

Introducing τ as in (3.73), we get now

$$\frac{\mathrm{d}v}{\sqrt{e^{2v} + e^{-2v} + 2E}} = |\lambda|\,\mathrm{d}\tau = |\lambda|\frac{\mathrm{d}s}{\sqrt{s^2 + s^4}}, \tag{3.97}$$

where E is an integration constant. For v to be regular for all $s \geq 0$, we have to take $E = -1$, whereupon we get

$$\frac{\mathrm{d}v}{e^v - e^{-v}} = |\lambda|\frac{\mathrm{d}s}{\sqrt{s^2 + s^4}}, \tag{3.98}$$

leading to

$$\frac{e^v - 1}{e^v + 1} = \Big(\frac{\sqrt{s^2 + 1} - 1}{s}\Big)^{2|\lambda|}, \tag{3.99}$$

so that

$$\begin{cases} v \sim \log s - \log|\lambda| + \cdots & \text{for } s \to \infty, \\ v \sim\ 2(s/2)^{2|\lambda|} & \text{for } s \to 0. \end{cases} \tag{3.100}$$

Equation (3.99) is equivalent to

$$e^v = \frac{1 + ((\sqrt{s^2+1}-1)/s)^{2|\lambda|}}{1 - ((\sqrt{s^2+1}-1)/s)^{2|\lambda|}}\,. \tag{3.101}$$

Taking $\lambda \to 0$, we get

$$e^v \sim \frac{1}{|\lambda| \log(s/(\sqrt{s^2+1}-1))}\,. \tag{3.102}$$

Thus, even though the form of the differential equation (3.94) suggests that the solution might become regular in the limit $\lambda \to 0$, this is not the case. However, if we shift v by $-\log|\lambda|$, then (3.102) coincides with the solution (3.88) in which φ is nilpotent. Modulo the shift in v (and an ordinary gauge transformation that depends on the argument of λ), the ansatz (3.93) converges for $\lambda \to 0$ to the ansatz (3.85) with a nilpotent pole. Thus, starting with the solution (3.99) in which φ has a pole at $z = 0$ with distinct eigenvalues $\pm\lambda$, and taking the limit $\lambda \to 0$, we get the solution (3.88) in which φ has a pole with nilpotent residue. An analogous phenomenon is known for solutions of Hitchin's equations with a regular singularity [91].

In the language of Section 6.3, the solution (3.89) has $\alpha^\vee \neq 0$ with $\beta^\vee = \gamma^\vee = 0$, while the solution (3.101) has $\beta^\vee, \gamma^\vee \neq 0$ with $\alpha^\vee = 0$. The solution (3.88) is the limit for $\alpha^\vee, \beta^\vee, \gamma^\vee \to 0$. It would be desirable to find a solution with generic values of $\alpha^\vee, \beta^\vee, \gamma^\vee$ (that is, a solution in which φ has a pole at $z = 0$ whose residue has distinct eigenvalues and the monodromy around the ray ℓ is generic). This appears to require a more complicated ansatz than the one we have used.

3.6.6. Two-sided solutions. The solutions that we have studied so far have been motivated by the problem of D3-branes on $\mathbb{R}^3 \times \mathbb{R}_+$, with D3–D5 boundary conditions and 't Hooft operators in the boundary. It is also of interest to consider a two-sided problem[20] of D3-branes on $\mathbb{R}^3 \times I$, where I is a compact interval, for instance the unit interval $0 \le y \le 1$, and we assume that the D3-branes end on D5-branes both at $y = 0$ and at $y = 1$. A time-independent configuration of 't Hooft operators is still described by the three-dimensional equations (3.48) and (3.49). Now we want a solution that describes 't Hooft operators on both components of the boundary.

A simple modification of the above ansatz gives examples of solutions of that type. (It does not give the most general such solutions.) We set

$$\varphi_1 = \begin{pmatrix} 0 & f(z) \\ h(z) & 0 \end{pmatrix} \tag{3.103}$$

where $f(z)$ and $h(z)$ are two polynomials. Zeroes of f and of h will be, respectively, the positions of 't Hooft operators at $y = 0$ and at $y = 1$. We take $\varphi = g\varphi_1 g^{-1}$ with g

[20] This problem is related to Chern–Simons theory on the boundary with a complex gauge group, as will be described elsewhere.

as in (3.60), and we leave the rest of the ansatz (3.61) unchanged. Equation (3.62) for v becomes

$$-\Big(\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial x_2^2}+\frac{\partial^2}{\partial y^2}\Big)v+|f|^2e^{2v}-|h|^2e^{-2v}=0. \tag{3.104}$$

To understand what sort of solution to look for, first consider the case that f and h are constants, so that no 't Hooft operators are present. Then one can look for a solution[21] that depends only on y. An elementary integration gives an implicit form of the solution

$$y=C-\int_0^v\frac{\mathrm{d}w}{\sqrt{|f|^2e^{2w}+|h|^2e^{-2w}+E}}\,, \tag{3.105}$$

with constants C, E. These constants can be adjusted in a unique way to ensure that $v\to+\infty$ for $y\to 0$ and $v\to-\infty$ for $y\to 1$. Then one has $v\sim-\log y-\log|f|$ for $y\to 0$, and $v\sim\log(1-y)+\log|h|$ for $y\to 1$. At both $y=0$ and $y=1$, the solution has a regular Nahm pole. Looking at the way v was introduced in (3.60), we see that a sign change of v can be compensated by a Weyl transformation that exchanges the two eigenvalues of a diagonal matrix; the structures at $y=1$ and $y=0$ are related in this way.

In general, for any polynomials f, h, we look for a solution such that $v\to+\infty$ for $y\to 0$ and $v\to-\infty$ for $y\to 1$. Then near $y=0$, the term $-|h|^2e^{-2v}$ is unimportant in (3.104). The analysis of the boundary behavior is the same as in the one-sided case; near a boundary point at which f is not zero, we have $v\sim-\log y-\log|f|$, while near a point at which f is zero, the boundary behavior is given by the appropriate model solution with an 't Hooft operator. Similarly, near $y=1$, the term $|f|^2e^{2v}$ is unimportant. The behavior near $y=1$ is the same as the behavior near $y=0$ with the substitutions $v\to-v$, $f\to h$, $y\to 1-y$.

3.7. The framing anomaly for knots. We have described the singularity associated to an 't Hooft operator supported on a knot K for the idealized case that K is a copy of $\mathbb{R}$ linearly embedded in $W=\mathbb{R}^3$. For the general case, we simply require that there should be a singularity along K that in the directions normal to K looks like this ideal solution. Away from K, the structure must be what we have already described in Sections 3.3 and 3.4.

An important consequence of this is the framing anomaly for knots. We will describe this for $G^\vee=\mathrm{SO}(3)$, which in any event is the case that we understand the 't Hooft operator in most detail. We consider an 't Hooft operator of spin j supported on K. In the absence of the 't Hooft operator, the restriction $E|_W$ of E to W coincides with TW, the tangent bundle to W, as we have seen in Section 3.4. In what follows, we are only concerned with the behavior along W, so we write simply E for $E|_W$. In the presence of the 't Hooft operator, E is modified along K and we denote this modification as $E_{(j)}$. The Riemannian connection ω on E is modified to

[21] This solution is related to one of the original solutions of Nahm's equation.

a connection on $E_{(j)}$ that we will call $\omega_{(j)}$. In the absence of the 't Hooft operator, a step in defining the partition function was to define a real-valued Chern–Simons function $\mathrm{CS}(\omega)$ (or $\mathrm{CS}_{\mathrm{grav}}$, but this refinement is not relevant in discussing the framing anomaly for knots). Similarly, to define the partition function in the absence of the 't Hooft operator, we need to be able to define a real-valued Chern–Simons function $\mathrm{CS}(\omega_{(j)})$. A framing of W makes it possible to define a lift of $\mathrm{CS}(\omega)$ to a real-valued function, but does not suffice for defining a natural real-valued $\mathrm{CS}(\omega_{(j)})$.

The additional information we need turns out to be a framing of K. For $K \subset W$ a knot, let NK be the normal bundle to K in W. The fibration $NK \to K$ has structure group SO(2) (we have taken W orientable from the beginning, since this is required in the definition of Chern–Simons theory, and K is certainly orientable, so NK is orientable). Since K is a one-manifold and SO(2) is connected, it follows that the fibration $NK \to K$ is trivial. But it has different homotopy classes of trivializations; given any one trivialization, any other can be found by twisting the first by a map from $K \cong S^1$ to SO(2). In other words, two trivializations differ by an element of $\pi_1(\mathrm{SO}(2)) \cong \mathbb{Z}$. A framing of K is a trivialization of NK up to homotopy. As we will see below, a real-valued function $\mathrm{CS}(\omega_j)$ can be defined if we are given framings of both W and K. Thus, the knot invariants that we obtain in the $G^\vee$ description can be naturally understood as invariants of framed knots in a framed three-manifold.[22]

Similarly, the knot invariants of Chern–Simons theory are most naturally defined for framed knots. Let us recall some details of this that will help in understanding what to look for on the $G^\vee$ side. The tangent bundle TW, when restricted to a knot K, is a direct sum $TK \oplus NK$, where TK is the tangent bundle to K. Unless K is a geodesic, this decomposition is not invariant under parallel transport along K. However, the Riemannian connection ω on TW induces a natural SO(2) connection ϖ on NK. Parallel transport of a vector in NK with respect to ϖ is defined as transport with respect to ω with a projection back to NK. Concretely, with respect to the decomposition $TW|_K = TK \oplus NK$, ϖ is the lower right block of ω:

$$\omega = \begin{pmatrix} 0 & * \\ * & \varpi \end{pmatrix}. \tag{3.106}$$

The holonomy of the connection ϖ is an element of SO(2) that we can write $\exp(\tau I)$ with

$$I = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{3.107}$$

For a "bare" knot, τ takes values in $\mathbb{R}/2\pi\mathbb{Z}$, but for a framed knot, τ is $\mathbb{R}$-valued. Indeed, once a framing is picked, the connection ϖ becomes $\varpi = \lambda I$, where now λ

[22] Here we can make a remark that parallels what was said about framings of three-manifolds at the end of Section 3.5.2. A knot $K \subset \mathbb{R}^3$ has a canonical framing (relative to which its self-linking number vanishes). Formulas for the Jones polynomial and related invariants are usually written relative to this canonical framing. Because the canonical framing cannot be found locally, it is natural to define the invariants for an arbitrary framing. In any event, in a general three-manifold W, a knot does not have a canonical framing.

is an ordinary one-form, and τ is simply $\oint_K \lambda$. If the framing of K is shifted by one unit (by making an SO(2)-valued gauge transformation of $NK \to K$ with winding number 1 around K), τ transforms by $\tau \to \tau + 2\pi$.

As essentially found for abelian Chern–Simons theory in [84] and more generally in [102], in computing the expectation value of a Wilson loop operator $\mathcal{W}_R(K)$ in Chern–Simons theory on W with gauge group G, one runs into an analog of what was described for three-manifolds in Section 3.5.3. The expectation value of $\mathcal{W}_R(K)$ is not independent of the metric of W unless one modifies its classical definition by including a factor that depends on τ:

$$\mathcal{W}_R(K) \longrightarrow \mathcal{W}_R(K) \exp(i d_R \tau). \tag{3.108}$$

Here d_R is a constant that can be usefully characterized using the relation of three-dimensional Chern–Simons theory to conformal field theory in two dimensions. For $k > 0$, d_R is the dimension of the primary field associated to the representation R in two-dimensional current algebra with symmetry group G at level k. Thus

$$d_R = \frac{c_2(R)}{k + h\,\mathrm{sign}(k)}, \tag{3.109}$$

where $c_2(R)$ is the value in the representation R of the quadratic Casimir operator of G (normalized to equal h in the adjoint representation). This formula is usually written only for $k > 0$; we have extended it to all nonzero integers k so that d_R is an odd function of k (this reflects the fact that for $k < 0$, Chern–Simons theory is related to an antiholomorphic rather than holomorphic current algebra in two dimensions). It follows from (3.108), (3.109), and the definition of q in (3.13) that under a unit change in framing of K, the Wilson loop operator transforms by

$$\mathcal{W}_R(K) \longrightarrow \mathcal{W}_R(K) q^{\mathrm{n_g} c_2(R)}. \tag{3.110}$$

For example, if $G = \mathrm{SU}(2)$ and R is the spin j representation, then

$$\mathcal{W}_R(K) \longrightarrow \mathcal{W}_R(K) q^{j(j+1)}. \tag{3.111}$$

The difference between E and $E_{(j)}$ is local along K, so to understand what happens in the dual $G^\vee$ description, it suffices to consider a local model of the neighborhood of $K \subset W$. We take such a neighborhood to be $W_0 = S^1 \times D$ where D is a disc of radius R. We assume that W is the union of two pieces W_0 and W_1, glued along their common boundary $\Xi = S^1 \times \widetilde{S}^1$, where $\widetilde{S}^1$ is the boundary of D. W_1 may be arbitrarily complicated, but W_0 will be very simple. To describe W_0, we introduce an angular coordinate α on S^1 and polar coordinates r, β ($0 \leq r \leq R$) on D, and we take the obvious flat metric:

$$\mathrm{d}s^2 = \mathrm{d}\alpha^2 + \mathrm{d}r^2 + r^2 \mathrm{d}\beta^2, \tag{3.112}$$

but with a twist of the following sort. We take β to be an ordinary angular variable,

$$\beta \cong \beta + 2\pi, \tag{3.113}$$

while under a 2π shift of α, we rotate $\mathbb{R}^2$ by an angle τ:

$$\alpha \to \alpha + 2\pi, \quad \beta \to \beta - \tau. \tag{3.114}$$

The definition of W_0 only depends on τ mod 2π, since $\beta \to \beta + 2\pi$ is an equivalence anyway. We take the knot K to be located at $r = 0$. Relative to the obvious orthonormal frame field

$$e_1 = \mathrm{d}\alpha, \quad e_2 = \mathrm{d}(r\cos\beta), \quad e_3 = \mathrm{d}(r\sin\beta), \tag{3.115}$$

the Riemannian connection ω simply vanishes. However, it has a nontrivial monodromy around S^1 because the orthonormal frame used in (3.115) has a monodromy under (3.114):

$$\begin{pmatrix} e_2 \\ e_3 \end{pmatrix} \longrightarrow \exp(\tau I) \begin{pmatrix} e_2 \\ e_3 \end{pmatrix}. \tag{3.116}$$

It is convenient to work with a single-valued orthonormal frame consisting of e_1 and

$$\begin{pmatrix} \tilde{e}_2 \\ \tilde{e}_3 \end{pmatrix} = \exp\Big(-\frac{\tau\alpha}{2\pi} I\Big) \begin{pmatrix} e_2 \\ e_3 \end{pmatrix}. \tag{3.117}$$

Unlike all the previous formulas, this one depends on τ as a real number, not just an angle. In fact, when restricted to K, $\tilde{e}_2$ and $\tilde{e}_3$ define a framing of K. This framing is shifted by n units if we modify (3.117) by $\tau \to \tau + 2\pi n$. The orthonormal frame $e_1, \tilde{e}_2, \tilde{e}_3$ also defines a framing of W_0, but this framing contains no relevant topological information.[23] We assume that the framing of W_0 given by $e_1, \tilde{e}_2, \tilde{e}_3$ (or at least the corresponding two-framing) is somehow matched to a framing of W_1, giving a framing of W. We want to see what happens to $\mathrm{CS}(\omega_{(j)})$ when we vary the framing of K while keeping fixed the framing or two-framing of W.

Relative to the orthonormal frame $e_1, \tilde{e}_2, \tilde{e}_3$, the Riemannian connection is

$$\omega = \frac{\tau\,\mathrm{d}\alpha}{2\pi} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}. \tag{3.118}$$

It is clumsy to write such a formula with a first row and column of zeroes. Everything of interest will happen in the lower right 2×2 block, and the 2×2 matrices will all be easily constructed from the SO(2) generator I of (3.107). So we will abbreviate a formula such as this one as

$$\omega = \frac{\tau\,\mathrm{d}\alpha}{2\pi} I. \tag{3.119}$$

Now we want to include the 't Hooft operator. As in (3.83) (which however was written in the two-dimensional representation while now we are in the adjoint

[23] Because $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$, the topological class of the framing of W_0 depends on n precisely mod 2. But the two-torsion information contained in a framing is not relevant in Chern–Simons theory. A convenient way to eliminate it [6] is to pass from a framing of TW to the corresponding framing of $TW \oplus TW$.

representation), this means that the Riemannian connection ω is replaced by a connection ω^* that is obtained from the Riemannian connection by adding a singular vortex of flux $2j$ acting on the normal bundle. In the same abbreviated notation as in (3.119), we take

$$\omega^* = 2j\left(\mathrm{d}\beta + \frac{\tau}{2\pi}\mathrm{d}\alpha\right)I + \frac{\tau\,\mathrm{d}\alpha}{2\pi}I. \tag{3.120}$$

This formula was chosen so that for fixed α it agrees with the singular vortex connection (3.83), and also so that ω^* is gauge-equivalent to ω for $r \neq 0$. The gauge transformation between them is

$$\mathrm{d} + \omega = \exp(-s)(\mathrm{d} + \omega^*)\exp(s), \tag{3.121}$$

with

$$s = -2j\left(\beta + \frac{\tau\alpha}{2\pi}\right)I. \tag{3.122}$$

s has been defined so that $\exp(s)$ is single-valued on the complement of the knot K.

We want to modify ω^* slightly near $r = 0$ to remove its singularity. We introduce a cutoff function $g(r)$ such that $g(r) = 1$ for $r > \epsilon$ (with some very small $\epsilon << R$) but $g(r) \sim r^2$ for $r \to 0$. We modify ω^* to

$$\hat{\omega} = 2j\left(g(r)\mathrm{d}\beta + \frac{\tau\,\mathrm{d}\alpha}{2\pi}\right)I + \frac{\tau\,\mathrm{d}\alpha}{2\pi}I. \tag{3.123}$$

(One can think of this modification as meaning that instead of restricting the bundle E literally to the boundary W of $V = W \times \mathbb{R}_+$, we restrict it to a three-cycle that coincides with the boundary away from knots, but near a knot K bends slightly into the interior of V to avoid the singularity along K.)

Now we can describe the desired bundle $E_{(j)} \to W$ and the connection $\omega_{(j)}$ on this bundle whose Chern–Simons function we want. On W_1, $E_{(j)}$ coincides with TW_1, and the connection is the Riemannian connection ω. On W_0, $E_{(j)}$ is a trivial bundle with connection $\hat{\omega}$ defined in (3.123). On the common boundary Ξ of W_0 and W_1, the bundles and connections are glued together with the gauge transformation (3.121). The framing (or more exactly the two-framing) of TW_0 that is given by e_1, $\tilde{e}_2$, $\tilde{e}_3$ has an extension over W that will be kept fixed while varying the framing of K. Everything is in place to compute a real-valued Chern–Simons function $\mathrm{CS}(\omega_{(j)})$ and determine its dependence on the framing of K. We use (2.53), in which $\mathrm{CS}(A)$ is defined for any connection A using a trace in the adjoint representation (and we set $h = 2$). In the present context, it is convenient to evaluate the right hand side of (2.53) as the sum of an integral over W_1 with the connection ω, an integral over W_0 with the connection $\hat{\omega}$, and a correction term on the common boundary Ξ of W_0 and W_1 that involves the gauge transformation between ω and $\hat{\omega}$:

$$\begin{aligned}\mathrm{CS}(\omega_{(j)}) =& \frac{1}{16\pi}\int_{W_1}\mathrm{Tr}_{\mathrm{ad}}\left(\omega\wedge\mathrm{d}\omega + \frac{2}{3}\omega\wedge\omega\wedge\omega\right)\\ &+ \frac{1}{16\pi}\int_{W_0}\mathrm{Tr}_{\mathrm{ad}}\,\hat{\omega}\wedge\mathrm{d}\hat{\omega} - \frac{1}{16\pi}\int_{\Xi}\mathrm{Tr}_{\mathrm{ad}}\,\mathrm{d}s\wedge\hat{\omega}.\end{aligned} \tag{3.124}$$

($\mathrm{Tr}_{\mathrm{ad}}$ is the trace in the adjoint representation of SO(3); some minor simplifications in (3.124) reflect the fact that $\hat{\omega}$ and the gauge transformation relating it to ω are actually abelian, taking values in an SO(2) subgroup. Evaluation of (3.124) uses $\mathrm{Tr}_{\mathrm{ad}}\, I^2 = -2$ and the orientation of W_0 given by $e_1 \wedge e_2 \wedge e_3$.) The terms in (3.124) that depend on the framing of K are the integrals over W_0 and Ξ. A straightforward evaluation gives

$$\mathrm{CS}(\omega_{(j)}) = -\tau j(j+1) + \cdots, \tag{3.125}$$

where the ellipses come from the integral over W_1 and do not depend on the framing of K. Using (3.34) (with $v = 1$ for $G^\vee = \mathrm{SO}(3)$), the dependence of the partition function on $\mathrm{CS}(\omega_{(j)})$ is a factor of $q^{-\mathrm{CS}(\omega_{(j)})/2\pi}$. So finally, under a unit change in framing, $\tau \to \tau + 2\pi$, the partition function is multiplied by $q^{j(j+1)}$, just as in Chern–Simons theory.

There is another issue that could be treated here using these ideas. This is to show that, for $W = \mathbb{R}^3$, with a knot K labeled by the spin j representation of SU(2), and using our boundary conditions, the instanton number P takes values in $\mathbb{Z} + j$. Setting $j = 1/2$, this accounts for the fact that the Jones polynomial is actually $q^{1/2}$ times a Laurent polynomial in q. More generally, for $W = \mathbb{R}^3$ with a link L with ν components labeled by $j_1, \dots, j_\nu$, P takes values in $\mathbb{Z} + \sum_{s=1}^{\nu} j_s$. We will postpone these issues and consider them in Section 5.4 from a higher-dimensional perspective. Similarly, in Section 5.4, we will give a new and possibly more transparent computation of the framing anomaly for knots.

4. *T*-duality and Khovanov homology

4.1. Lift to five dimensions

4.1.1. Five-dimensional super Yang–Mills and *T*-duality. So far we have found a new way to calculate the partition function of three-dimensional Chern–Simons gauge theory with gauge group G, using $G^\vee$ gauge theory in four dimensions. To get to Khovanov homology takes an additional step: we need a fifth dimension.

From a field theory point of view, we can try to proceed by claiming that four-dimensional maximally supersymmetric Yang–Mills theory is the theory obtained at low energies by compactifying five-dimensional maximally supersymmetric Yang–Mills theory on a circle. Thus, instead of considering four-dimensional $\mathcal{N} = 4$ super Yang–Mills theory on a four-manifold V, we consider the corresponding five-dimensional theory on $V \times S^1$ (with supersymmetry-preserving boundary conditions in going around S^1). The twisting along V and the boundary conditions at the boundary of V preserve the same supersymmetry that they did in the purely four-dimensional formulation of the theory. (The boundary condition of Section 3.3 can be lifted to five dimensions in an obvious way; three of the scalar fields have the singular behavior at the boundary described there.) In particular, the topological

supercharge Q that is familiar in four dimensions is still a symmetry when the model is lifted to five dimensions.

Once the model is lifted to $V \times S^1$, we can pick a point $p \in S^1$ and construct a physical Hilbert space $\mathcal{H}(V)$ associated to quantization on the codimension one submanifold $V \times p$. The path integral on $V \times S^1$ can then be written as a trace in $\mathcal{H}(V)$. In the present approach, $\mathcal{H}(V)$ plays the role of the space that was called by that name in our introductory sketch of Khovanov homology in Section 1.2. Q automatically acts on $\mathcal{H}(V)$, as it generates a symmetry of the theory. We write $\mathcal{K}(V)$ for the cohomology of Q, acting on $\mathcal{H}(V)$. Then $\mathcal{K}(V)$ is our candidate for the generalization to this situation of Khovanov homology. (Since we do not have a proof that the cohomology of Q is equivalent to Khovanov homology as defined in the literature, even if one specializes to the situation of knots in $\mathbb{R}^3$ where Khovanov homology has been defined, we denote the cohomology of Q as $\mathcal{K}$ and write k for Khovanov homology.)

From a D-brane point of view, the lift from four to five dimensions amounts to T-duality. Thus, for the case that the gauge group is $G^\vee = \mathrm{U}(N)$, consider a system of N D3-branes wrapped on V, with some twisting of the normal bundle to V to preserve supersymmetry. This picture was described in Section 2.2.3. Without changing anything essential in that discussion, we can take one of the spacetime directions transverse to V to be compactified on a circle $\widetilde{S}^1$. Explicitly, we replace what in Section 2.2.3 was $T^*V_0 \times \mathbb{R}^2$ by $T^*V_0 \times \mathbb{R} \times \widetilde{S}^1$. Then we perform T-duality on $\widetilde{S}^1$, converting the spacetime to $T^*V_0 \times \mathbb{R} \times S^1$. The D3-branes wrapped on $V \subset V_0$ are converted to D4-branes wrapped on $V \times S^1$. If as in Section 2.2.3, the D3-branes end on a D5-brane (wrapped on T^*W with $W = \partial V$), then T-duality converts the D3-branes to D4-branes that end on a D6-brane (wrapped on $T^*W \times S^1$). So, when the appropriate geometry exists, the lift to five dimensions simply amounts to T-duality from the D3–D5 system that we have studied so far to a D4–D6 system.

None of the approaches just mentioned is entirely satisfactory. The disadvantage of the description by five-dimensional super Yang–Mills theory is that this theory is not ultraviolet complete. The brane construction also has a few drawbacks, which were described in Section 2.2.3. The appropriate Calabi–Yau geometry may not exist for generic V, and even if it exists, it may entail unnatural choices. The brane construction does not help very much with exceptional gauge groups. Also, the brane construction and the full string theory have many degrees of freedom that are not relevant to the problem of defining an analog of Khovanov homology and relating it to Chern–Simons theory.

There is a completely satisfactory alternative to the approaches that we have summarized so far. Five-dimensional maximally super Yang–Mills theory has a canonical ultraviolet completion in the six-dimensional (0,2) superconformal field theory. This gives a general and economical framework for the topic considered in the present paper, and for many purposes it is probably the most powerful framework. In Section 5, we will develop a top-down approach to the subject with this starting

point. As an illustration of the power of this viewpoint, we will show that in the six-dimensional picture, the existence of supersymmetric Wilson and 't Hooft operators precisely at the boundary of V follows from standard facts, while in the four and five-dimensional pictures, this seems to require the detailed computations in Sections 2.2.4 and 3.6.

But some important points, especially the representation (4.10) of the Chern–Simons partition function as a trace in Khovanov homology, do not require the six-dimensional machinery. So it seems reasonable to begin with an explanation in five dimensions.

4.1.2. The bigrading. To agree with Khovanov homology, $\mathcal{K}(V)$ should admit a $\mathrm{U}(1)\times\mathrm{U}(1)$ action, so that it will be $\mathbb{Z}\times\mathbb{Z}$ graded.[24] One generator of $\mathrm{U}(1)\times\mathrm{U}(1)$ is the instanton number, evaluated on the four-cycle V. The definition is the same as it was in Section 3.1:

$$\mathrm{P}=\frac{1}{32\pi^2}\int_V \epsilon^{\mu\nu\alpha\beta}\,\mathrm{Tr}\,F_{\mu\nu}F_{\alpha\beta}\,. \tag{4.1}$$

However, the physical interpretation is different: in the five-dimensional interpretation, P is an operator acting on quantum states that are obtained by quantizing fields on V, while in the four-dimensional interpretation, P was a term in the classical action.

The other generator of $\mathrm{U}(1)\times\mathrm{U}(1)$ is an R-symmetry generator F that is left unbroken by the twisting procedure that is used to define a topological field theory. In the four-dimensional analysis of Section 2.2, we began with the R-symmetry group SO(6) of $\mathcal{N}=4$ super Yang–Mills theory in four dimensions, and twisted by identifying an SO(4) subgroup of SO(6) with the Riemannian holonomy of V. This left an unbroken subgroup $\mathrm{SO}(2)\subset\mathrm{SO}(6)$, and we defined the generator of this $\mathrm{SO}(2)\cong\mathrm{U}(1)$ to be F. When we lift to five dimensions, the R-symmetry group is reduced to SO(5), so embedding an SO(4) holonomy group in the R-symmetry group would not leave an unbroken SO(2). To compensate for this, we specialize to $V=W\times\mathbb{R}_+$ (or $V=W\times S$ for any one-manifold S), with W a three-manifold. This ensures that the holonomy group of V reduces to SO(3), so that its embedding in the R-symmetry group, which is now SO(5), again leaves an unbroken SO(2). We again denote the generator of this symmetry as F. For general V, we do not get a $\mathbb{Z}$-grading by F, but there is always a $\mathbb{Z}_2$-grading that distinguishes bosonic states from fermionic ones. When F can be defined, the $\mathbb{Z}_2$-grading by statistics is the mod 2 reduction of the $\mathbb{Z}$-grading by F. It turns out, however, that the lift to five dimensions is useful primarily when the conserved charge F can be defined, so we will be mainly interested in that case.

[24] This is a slight simplification as in general the eigenvalues of the symmetry generators F and P may lie in a coset of $\mathbb{Z}\times\mathbb{Z}\subset\mathbb{R}\times\mathbb{R}$. The most important consequence of this was described in Section 3.5. More generally, if $G^\vee$ is not simply-connected, the eigenvalues of P may lie in a coset of $\mathbb{Z}/w\subset\mathbb{R}$ for some integer w, rather than in a coset of $\mathbb{Z}$. This last effect, which was discussed in relation to (3.14), is not directly relevant to Khovanov homology, because it does not arise for $V=W\times\mathbb{R}_+$ with the sort of boundary conditions that we impose on ∂V.

Of course, when V has a boundary, to define F, the boundary condition must be F-invariant. But there is no problem with this. We use the boundary condition of Section 3.3, lifted to five dimensions. Three of the five scalar fields of five-dimensional maximally supersymmetric Yang–Mills theory have expectation values that diverge at the boundary, leaving an unbroken SO(2)-symmetry that rotates the other two. The two scalars that are rotated by F play the role of the complex field σ of Section 2.2.1. In any supersymmetric classical solution, σ vanishes and the value of F also vanishes. Quantum mechanically, for a quantum state associated to a given classical solution, the eigenvalue of F is computed by summing over the F quantum numbers of all fermions in the filled Dirac sea. In that sense, it makes sense to refer to F as a fermion number.

A more detailed and complete explanation of many of these matters is given in Section 5 in the context of an ultraviolet completion of five-dimensional super Yang–Mills theory in six dimensions. For now, it is enough to know that, not for all V, but for V of the form $W \times \mathbb{R}_+$, $\mathcal{K}(V)$ is bigraded, like Khovanov homology.

Since Khovanov homology has been defined in the literature only for links in $\mathbb{R}^3$, to make a precise conjecture about the relation of $\mathcal{K}(V)$ to Khovanov homology, we must restrict to $V = \mathbb{R}^3 \times \mathbb{R}_+$. For Khovanov homology, we consider a link $L \subset \mathbb{R}^3$ consisting of a disjoint union of embedded circles $K_i \subset \mathbb{R}^3$. We label each K_i by an irreducible representation R_i of a compact Lie group G. In the four-dimensional description of Section 2 via G gauge theory, we include supersymmetric Wilson operators of the representations R_i, supported on $K_i \times \{0\}$, where $\{0\}$ is the endpoint of $\mathbb{R}_+$. In the S-dual description in Section 3, the gauge group is $G^\vee$, the Goddard–Nuyts–Olive or Langlands dual of G, and the Wilson operators in the boundary of V are converted to the dual 't Hooft operators of $G^\vee$ gauge theory. The description of 't Hooft operators in the boundary of V is somewhat subtle and was described in Section 3.6. In this situation, $\mathcal{K}(V)$ is a candidate for Khovanov homology.

4.1.3. Notation. As we move to five dimensions, the cast of characters will get longer. To make the arguments easier to follow, in the rest of the paper we write V_4 and W_3 for the four-manifold and three-manifold that earlier we have called simply V and W. Thus W_3 is always the boundary of V_4.

4.2. Procedure for computing $\mathcal{K}$. Now we would like to sketch the concrete procedure for computing $\mathcal{K}(V_4)$, for a four-manifold V_4, via five-dimensional supersymmetric Yang–Mills theory. This procedure is in no way novel; it is a standard procedure in topological applications of supersymmetric theories; typical examples involve Morse theory [100] or Floer cohomology [31]. We sketch the procedure here for completeness.

We want to describe a procedure to determine the space of quantum ground states of twisted super Yang–Mills theory on the five-manifold $M_5 = \mathbb{R} \times V_4$. For comparison to Chern–Simons theory (or Khovanov homology), we take $V_4 = W_3 \times \mathbb{R}_+$ for

some W_3, but the general procedure to describe the space of ground states holds for any V_4.

First of all, the condition for a five-dimensional field configuration to preserve the Q-symmetry gives a system of elliptic differential equations in five dimensions. It is straightforward to derive these equations, and we will do so in Section 5.2; see (5.36) for the final result. But for now, we do not need the details. All we need to know is that these are elliptic differential equations that, in the time-independent case, specialize to the familiar four-dimensional equations

$$F - \varphi \wedge \varphi + \star \mathrm{d}_A \varphi = 0 = \mathrm{d}_A \star \varphi. \tag{4.2}$$

The first approximation to finding the space of quantum ground states is to find the space of classical ground states. A classical ground state is a time-independent classical solution of the five-dimensional equations for unbroken supersymmetry. So in other words, a classical ground state is a solution of (4.2) on the four-manifold V_4. For simplicity we are going to assume that this equation has a finite set of solutions, up to gauge transformation, and further that these solutions are all nondegenerate (there are no bosonic zero modes in expanding around a given solution). Let S be the set of these solutions. If V_4 has a non-empty boundary, then on ∂V_4 we impose the boundary conditions of Section 3.3; with these boundary conditions, the solutions are automatically all irreducible (they leave unbroken only a finite group of gauge symmetries, in fact the center of $G^\vee$). If V_4 has no boundary, we assume for simplicity that the solutions are all irreducible.

Nondegeneracy means that the expansion around a given classical solution gives, at least perturbatively, a single quantum state of zero energy. We will let $\mathcal{K}_0$ be the space of quantum ground states in the classical approximation; it has a basis consisting of a single state ψ_s for each $s \in S$. We let n_s be the instanton number P for the s^{th} classical solution, as defined in (4.1). Assuming that $V_4 = W_3 \times \mathbb{R}_+$ for some W_3, we let f_s be the fermion number F of the s^{th} classical solution. (It equals the value of F for the filled Dirac sea that one obtains in expanding around the s^{th} solution.) For any V_4, $\mathcal{K}_0$ is $\mathbb{Z} \times \mathbb{Z}_2$-graded, where the $\mathbb{Z}$-grading is by the eigenvalue of P, and the $\mathbb{Z}_2$ distinguishes fermionic states from bosonic ones. For $V_4 = W_3 \times \mathbb{R}_+$, $\mathcal{K}_0$ is $\mathbb{Z} \times \mathbb{Z}$ graded by the eigenvalues of P and F.

Now we want to consider quantum corrections to this spectrum. Once one has an asymptotic approximation to the space of supersymmetric states – in this case $\mathcal{K}_0$ – states can only disappear from the supersymmetric spectrum in Bose–Fermi pairs. The reason for this is familiar: eigenstates of the supersymmetric Hamiltonian with a nonzero energy occur in pairs, corresponding to a bosonic state and a fermionic state of the same energy. In the $\mathbb{Z} \times \mathbb{Z}_2$-graded case, a pair of states that are going to disappear must have the same P eigenvalue (since P commutes with Q) and opposite statistics. In the $\mathbb{Z} \times \mathbb{Z}$-graded case, a pair of states that are going to disappear from the supersymmetric spectrum must have the same eigenvalue of P and eigenvalues of F that differ by 1. (The last statement is a consequence of the commutation relation

$[\mathrm{F}, Q] = Q$, which implies that a supermultiplet of energy eigenstates with nonzero energy consists of a pair of states with values of F differing by ± 1.)

In perturbation theory, nothing happens to the supersymmetric spectrum. Indeed, perturbation theory around a given classical solution only "knows" about a single approximate supersymmetric state, namely the one obtained by quantizing that classical solution. In perturbation theory, there is no way for that approximate supersymmetric ground state to pair up with another one and disappear. However, just as in supersymmetric quantum mechanics or Floer cohomology, instanton effects involving tunneling from one classical solution to another can lift a pair of supersymmetric states away from zero energy. In the present context, instantons are solutions of the five-dimensional supersymmetric equations, the ones that are presented in (5.36) and whose reduction to the time-independent case agrees with (4.2). An instanton that interpolates between one solution of (4.2) in the past and another in the future can lift away from zero energy the supersymmetric quantum states that correspond to the two solutions.

Let $\mathcal{K}$ be the exact supersymmetric spectrum that we get after allowing for the effects of instantons. A precise and general recipe for computing $\mathcal{K}$ is that it is the cohomology of a certain operator acting on $\mathcal{K}_0$. This operator is simply Q evaluated in the space $\mathcal{K}_0$ generated by the approximate supersymmetric states ψ_s. A precise formula for Q, up to conjugation, is

$$Q\psi_s = \sum_{\{t\in S\,|\,f_t - f_s = 1\}} n_{st}\psi_t\,, \tag{4.3}$$

where n_{st} is computed by summing over instantons that begin at the s^{th} solution in the past and end on the t^{th} solution in the future. Such solutions come in one-parameter families generated by time translation invariance; each such family contributes 1 or -1 to n_{st}, depending on the sign of the fermion determinant that arises in linearizing around the given solution, after removing the zero mode that comes from time-translation invariance. The details are standard in Floer cohomology and related theories, and will not be described here.

4.2.1. Relation to Chern–Simons theory. Now we want to explain how $\mathcal{K}(V_4)$, as just described, is related to the S-dual four-dimensional construction of Section 3. For brevity, we focus on the $\mathbb{Z}\times\mathbb{Z}$-graded case $V_4 = W_3\times\mathbb{R}_+$, so that we also will get a link to Chern–Simons theory on W_3. The general case is similar, except that the function $L(q, y)$ that is introduced shortly is only defined for $y = -1$ since the grading is only by $\mathbb{Z}\times\mathbb{Z}_2$.

First of all, if we know $\mathcal{K}(V_4)$, then we can compute the function

$$L(q, y) = \mathrm{Tr}_{\mathcal{K}(V_4)}\, q^{\mathrm{P}} y^{\mathrm{F}}. \tag{4.4}$$

For $V_4 = W_3\times\mathbb{R}_+$, this function is an invariant of W_3, or of W_3 together with the knot or link it may contain, if any. However, there is no convenient way to represent this function by a path integral.

To get a trace associated to V_4, we should consider a path integral on the five-manifold $M_5 = V_4 \times S^1$. If $\mathcal{H}$ is the Hilbert space of all physical states of five-dimensional super Yang–Mills theory (not necessarily annihilated by Q), H is the Hamiltonian acting on $\mathcal{H}$, and β is the circumference of S^1, then a path integral on M_5 with an insertion of the operator $q^{\mathrm{P}} y^{\mathrm{F}}$ can compute

$$G(q, y) = \mathrm{Tr}_{\mathcal{H}}\, q^{\mathrm{P}} y^{\mathrm{F}} \exp(-\beta H). \tag{4.5}$$

However, this trace receives contributions from states of nonzero energy. A pair of states with $H = E$, $\mathrm{P} = n$, and $\mathrm{F} = f$, $f + 1$ contribute

$$q^n \exp(-\beta E)(y^f + y^{f+1}) \tag{4.6}$$

to $G(q, y)$. To make this contribution vanish, we must choose y so that $y^f + y^{f+1} = 0$; in other words, we need to take $y = -1$. Otherwise, $G(q, y)$ is not a topological invariant. If we set $y = -1$, $G(q, y)$ reduces to $L(q, y)$.

The study of Khovanov homology has shown that the function $L(q, y)$ contains quite a lot of information that we lose if we set $y = -1$. However, the case $y = -1$ is the case that can be represented by a path integral on M_5. For this value of y, the trace in (4.4) or (4.5) computes what is usually called the *index* of the operator Q, or more precisely the equivariant generalization of this index to take account of the symmetry generated by P. (We get the ordinary index of Q if we set $q = 1$.) As is usual, the index of an operator is more readily computed by a path integral than are other topological invariants.

Not only can $L(q, -1)$ be represented by a five-dimensional path integral on M_5; it can more simply be represented by a path integral on V_4. The reason for this is as follows. Approximate supersymmetric states that are lifted from the spectrum by instanton effects do not contribute to $L(q, -1)$ (since they have the same value of P and have F differing by 1). So we can calculate $L(q, -1)$ in the space $\mathcal{K}_0(V_4)$ of approximate supersymmetric ground states, instead of the space $\mathcal{K}(V_4)$ of states of exactly zero energy:

$$L(q, -1) = \mathrm{Tr}_{\mathcal{K}_0(V_4)}\, q^{\mathrm{P}} (-1)^{\mathrm{F}}. \tag{4.7}$$

Before looking at this formula more closely, let us note as an aside that we could also, of course, define a more general trace in $\mathcal{K}_0(V_4)$:

$$\widetilde{L}(q, y) = \mathrm{Tr}_{\mathcal{K}_0(V_4)}\, q^{\mathrm{P}} y^{\mathrm{F}}. \tag{4.8}$$

But in general, one should not expect $\widetilde{L}(q, y)$ to be a topological invariant. The reason is that, unlike $\mathcal{K}(V_4)$, $\mathcal{K}_0(V_4)$ is not, in general, a topological invariant. In general, one should expect supersymmetric classical solutions to appear and disappear in pairs as the metric on V_4 is varied; when this occurs, $\widetilde{L}(q, y)$ will jump with no change in $L(q, y)$. Concretely, when one varies the metric of V_4 so that a pair of time-independent classical solutions appears, there also appears a time-dependent

instanton solution that interpolates between them and ensures that the extra two states that have appeared in $\mathcal{K}_0(V_4)$ do not contribute to $\mathcal{K}(V_4)$.

Since we want to study topological invariants, we set $y = -1$. Now let us go back to the formula (4.7) for $L(q,-1)$. This trace is a sum over classical solutions of the time-independent equations (4.2); as before, we assume that the solutions are nondegenerate and parametrized by a finite set S. For each $s \in S$, we write n_s and f_s for the P and F eigenvalues of the approximate ground state ψ_s. The explicit formula for $L(q,-1)$ is then

$$L(q,-1) = \sum_{s\in S} q^{n_s}(-1)^{f_s}. \tag{4.9}$$

But this coincides with the formula (3.15) for the purely four-dimensional path integral on V_4 provided the sign $(-1)^{g_s}$ of the four-dimensional fermion determinant coincides with $(-1)^{f_s}$. The justification for that last statement is that as one varies the metric of V_4 or the background fields A, φ in the Dirac operator, the sign of the four-dimensional fermion determinant is reversed whenever it has a zero mode; but these are precisely the points at which, from a five-dimensional point of view, the value of f_s jumps by ± 1. (This argument does not fix an additive constant in g_s; this constant depends on a choice of trivialization of the determinant line bundle in four dimensions. We fix the constant to reconcile the four- and five-dimensional formulas.)

In turn, we know that for $V_4 = W_3 \times \mathbb{R}_+$, the four-dimensional path integral (3.15) equals the Chern–Simons path integral $Z^{\rm CS}_{W_3}(q)$ on W_3. Putting everything together, we have obtained the relation

$$Z^{\rm CS}_{W_3}(q) = {\rm Tr}_{\mathcal{K}(W_3\times\mathbb{R}_+)}\, q^{\rm P}(-1)^{\rm F} \tag{4.10}$$

between Chern–Simons theory on W_3 and our candidate $\mathcal{K}(W_3 \times \mathbb{R}_+)$ for the generalized Khovanov homology. But in general, something is hidden in the way we have written this formula.

On the left hand side of this formula, the possible integration cycles of the Chern–Simons theory on W_3 that must be used for computing $Z^{\rm CS}_{W_3}$ are associated to critical points of the $G_{\mathbb{C}}$-valued Chern–Simons function on W_3 – in other words, to homomorphisms $\rho\colon \pi_1(W_3) \to G_{\mathbb{C}}$. On the right hand side, $\mathcal{K}(W_3 \times \mathbb{R}_+)$ is defined using a homomorphism $\rho^\vee\colon \pi_1(W_3) \to G^\vee_{\mathbb{C}}$ to set the boundary condition at infinity. To use the formula in general, we would have to understand the relation between ρ and $\rho^\vee$ determined by S-duality. A more precise version of the formula would involve a sum as in (3.16) with an unknown matrix $m_{\rho^\vee,\rho}$. We can avoid this problem if we specialize to $W_3 = \mathbb{R}^3$ with a link whose components are labeled by Wilson operators on the left hand side of (4.10) or by the dual 't Hooft operators on the right hand side. Then ρ and $\rho^\vee$ are both trivial, so we do not need to analyze an S-duality transformation between them. The relation (4.10) becomes – conjecturally – the classical relation between Khovanov homology (and its generalization to arbitrary representations of compact Lie groups) and the Jones polynomial (and more general knot invariants derived from Chern–Simons theory), as described in (1.9) of the introduction.

4.3. Lie groups that are not simply-laced. We are now going to explain a possibly surprising fact: when the gauge group G of Chern–Simons theory is not simply-laced, there is a perfectly good alternative to what has just been explained.

Although this is a general fact, we will, to be concrete, explain it first for the case that $G = \mathrm{Sp}(2n)$ for some n. The GNO or Langlands dual group is then $G^\vee = \mathrm{SO}(2n+1)$. And this is a subgroup of the simply-laced Lie group $G^* = \mathrm{SO}(2n+2)$. G^* admits an outer automorphism that we will call ζ that leaves fixed $G^\vee$. In the $(2n+2)$-dimensional representation of G^*, ζ acts by the matrix $\mathrm{diag}(1, 1, \dots, 1, -1)$.

As is clear from the explicit description in Section 3.3, a principal $\mathfrak{su}(2)$ subalgebra of $\mathrm{SO}(2n+2)$ can actually be conjugated into the Lie algebra of $\mathrm{SO}(2n+1)$. With this choice, it commutes with ζ. This means that the boundary condition of the D3–D5 system, as described in Section 3.3, or its T-dual, the boundary condition of the D4–D6 system, as studied in this section, is ζ-invariant.

Hence, taking the gauge group to be G^*, ζ acts on the set S^* of solutions of the four-dimensional equations (4.2). We denote this space as S^*, rather than S (as before), to emphasize that we are taking the gauge group to be G^* rather than $G^\vee$. The set S of solutions of the equations (4.2) with gauge group $G^\vee$ is simply the set of fixed points of ζ acting on S^*. We will likewise write $\mathcal{K}_0^*(V_4)$ and $\mathcal{K}^*(V_4)$ for the spaces of approximate and exact quantum ground states in the G^* theory, while $\mathcal{K}_0(V_4)$ and $\mathcal{K}(V_4)$ will be the corresponding spaces for gauge group $G^\vee$.

Since ζ acts on the set S^*, it also acts on the vector space $\mathcal{K}_0^*(V_4)$, which is simply constructed to have one basis vector ψ_s for every $s \in S^*$. ζ is also a symmetry of the five-dimensional "instanton" equations that lift some states in $\mathcal{K}_0^*(V_4)$ (this is hopefully natural even though we will not actually construct those equations until Section 5), so it acts on $\mathcal{K}^*(V_4)$ as well.

Using the ζ action on $\mathcal{K}^*(V_4)$, we can now define a new trace that generalizes (4.4):

$$L_\zeta^*(q, y) = \mathrm{Tr}_{\mathcal{K}^*(V_4)}\, q^{\mathrm{P}} y^{\mathrm{F}} \zeta. \tag{4.11}$$

Here for brevity, but also because it is the most interesting case, we assume that $V_4 = W_3 \times \mathbb{R}_+$ so that we can define the F-symmetry. Note that ζ commutes with P and with F, as well as with Q.

Just as in the discussion of (4.4), to represent $L_\zeta^*(q, y)$ by a path integral in a simple way is only possible if $y = -1$. So let us consider the relation of $L_\zeta^*(q, -1)$ to Chern–Simons theory. Just as in (4.7), in computing $L_\zeta^*(q, -1)$, we can replace the trace in $\mathcal{K}^*(V_4)$ by a trace in $\mathcal{K}_0^*(V_4)$:

$$L_\zeta^*(q, -1) = \mathrm{Tr}_{\mathcal{K}_0^*(V_4)}\, q^{\mathrm{P}} (-1)^{\mathrm{F}} \zeta. \tag{4.12}$$

We can evaluate the trace in (4.12) by summing over the basis of $\mathcal{K}_0^*$ given by the vectors ψ_s, $s \in S^*$. In this basis, we evaluate the trace by summing over the diagonal matrix elements of $q^{\mathrm{P}}(-1)^{\mathrm{F}}\zeta$. Since P and F are diagonal in the chosen basis, the trace receives contributions only from diagonal matrix elements of ζ. The action of ζ in this basis is easily described. ζ is a permutation matrix determined by the

action of ζ on the set S^*. ζ either leaves fixed a given $s \in S^*$ or exchanges a pair of elements. Nonzero diagonal matrix elements of ζ are all 1 and correspond to ζ-invariant elements of S^*. But the ζ-invariant elements of S^* make up precisely the set S of $G^\vee$-valued solutions of the four-dimensional localization equations. Hence

$$L^*_\zeta(q,-1) = \sum_{s\in S} q^{\mathrm{P}}(-1)^{\mathrm{F}} = \mathrm{Tr}_{\mathcal{K}(V_4)}\, q^{\mathrm{P}}(-1)^{\mathrm{F}}. \tag{4.13}$$

Since we got the same result for $L(q,-1)$ in (4.9), we learn that $L^*_\zeta(q,-1) = L(q,-1)$. Since we have already identified $L(q,-1)$ with the Chern–Simons partition function of $G = \mathrm{Sp}(2n)$, we actually now have two alternative formulas for this function:

$$Z^{\mathrm{CS}}_{W_3}(q) = L^*_\zeta(q,-1) = L(q,-1). \tag{4.14}$$

Both of these formulas amount to ways of writing the Chern–Simons partition function as a trace:

$$Z^{\mathrm{CS}}_{W_3}(q) = \mathrm{Tr}_{\mathcal{K}(W_3\times\mathbb{R}_+)}\, q^{\mathrm{P}}(-1)^{\mathrm{F}} = \mathrm{Tr}_{\mathcal{K}^*(W_3\times\mathbb{R}_+)}\, q^{\mathrm{P}}(-1)^{\mathrm{F}}\zeta. \tag{4.15}$$

Actually, the attentive reader may notice a small gap in this derivation: we have assumed that for a given $G^\vee$-valued classical solution, the values of P and $(-1)^{\mathrm{F}}$ are the same whether calculated in $G^\vee$ or after embedding of the solution in G^*. For P, this is a classical fact about the instanton number, but a proof of what we want for $(-1)^{\mathrm{F}}$ is not clear at the moment[25] and this is a gap in our explanation. A proof may follow from a vanishing theorem for the five-dimensional Dirac operator.

We have treated the case of $G = \mathrm{Sp}(2n)$, but a similar derivation works for any gauge group that is not simply-laced. For $G = \mathrm{SO}(2n+1)$, we have $G^\vee = \mathrm{Sp}(2n)$. We can take G^* to be the simply-laced Lie group $\mathrm{SU}(2n)$, which admits an outer automorphism ζ that leaves fixed $G^\vee$. Once again, a principal $\mathfrak{su}(2)$ subalgebra of $G^\vee$ embeds as a principal $\mathfrak{su}(2)$ subalgebra of G^*. This is clear from the description of the principal subgroups in Section 3.3. So we can repeat all steps in the above derivation, arriving again at (4.14) and (4.15).

The other cases of non-simply-laced Lie groups are similar, though less obvious. If $G = \mathsf{G}_2$ or F_4, then again $G^\vee = \mathsf{G}_2$ or F_4. For $G^\vee = \mathsf{G}_2$, we take $G^* = \mathrm{Spin}(8)$ with ζ a triality automorphism, which is of order 3. We can pick ζ to leave fixed $\mathsf{G}_2 \subset G^*$, and a principal $\mathfrak{su}(2)$ subalgebra of G_2 embeds as one of G^*. For $G^\vee = \mathsf{F}_4$, we take $G^* = \mathsf{E}_6$. E_6 admits an outer automorphism ζ of order 2, which we can choose to leave F_4 fixed. Again a principal $\mathfrak{su}(2)$ subalgebra of F_4 embeds as one of E_6. (Proofs of the statements in this paragraph about principal $\mathfrak{su}(2)$ subalgebras have been sketched by B. Kostant.) So we can repeat the above derivation, leading to the same conclusions (4.14) and (4.15).

[25] This actually is clear for the case $G^\vee = \mathsf{G}_2$, $G^* = \mathrm{Spin}(8)$. The complement of the $G^\vee$ Lie algebra in that of G^* is two copies of the irreducible seven-dimensional representation of $G^\vee$. When we embed $G^\vee$ in G^*, the fermion determinant is multiplied by the square of a real determinant associated to the seven-dimensional representation of $G^\vee$, so its sign does not change.

4.4. Ultraviolet completion. Mathematically, the approach to this subject via five-dimensional gauge theory has the great advantage of relying on five-dimensional elliptic differential equations, without needing the full machinery of quantum field theory and string theory. (We have not yet described explicitly the relevant five-dimensional equations and their essential properties; this will be done starting in Section 5.2.) Indeed, this fact is the main reason that the present paper may have some mathematical impact in the short term.

Physicists will generally prefer a starting point based on an ultraviolet-complete quantum field theory. This we will present in Section 5. Some of the drawbacks of relying on five-dimensional supersymmetric Yang–Mills theory were described at the end of Section 4.1.1.

The alternative formulas of (4.15) for the Chern–Simons partition function when G is not simply-laced give an interesting challenge for the six-dimensional approach. In Section 5.5, we will suggest two slightly different six-dimensional starting points that lead to the two formulas.

5. Top-down approach

So far in this paper, we have worked our way up from three to four and then five dimensions. The logical end of this process is the six-dimensional superconformal field theory that provides an ultraviolet completion of five-dimensional super Yang–Mills theory.

In the present section, we begin in six dimensions and deduce the five-dimensional picture that was used in Section 4. We also fill in many key gaps in Section 4, mainly by deriving the explicit form of the relevant elliptic differential equations and describing their key properties.

The six-dimensional starting point in the present section will also bring us closer to the brane constructions that have been used previously in related work [82], [53], [26], [2], and [19].

We began our analysis in Section 2 on a fairly general four-manifold V_4 with boundary W_3. In Section 4, we lifted the analysis to the five-manifold $S^1 \times V_4$. In that context, as was explained in Section 4.1.2, to maintain the bigrading that gives Khovanov homology much of its power, one must specialize[26] to $V_4 = W_3 \times \mathbb{R}_+$, for some W_3, so that the five-dimensional description is based on $M_5 = S^1 \times W_3 \times \mathbb{R}_+$. However, it turns out that this can be generalized. The five-dimensional version of the construction makes sense on $M_5 = M_4 \times \mathbb{R}_+$, with any oriented four-manifold M_4 without boundary, not necessarily of the form $W_3 \times S^1$. (Note that in the important case that $M_5 = S^1 \times W_3 \times \mathbb{R}_+$, M_4 is not the same as V_4; V_4 is $W_3 \times \mathbb{R}_+$ while M_4 is $S^1 \times W_3$.) We will define a four-dimensional topological field theory that will

[26] More generally, one could replace $\mathbb{R}_+$ by another one-manifold, notably a circle, real line, or compact unit interval.

work for an arbitrary M_4. Moreover, M_4 can be endowed with "surface operators," supported on a two-manifold $\Sigma \subset M_4$. Though any M_4 is allowed, this theory is most interesting (for a reason explained in Section 5.2.2 and again involving the bigrading), if the third Betti number of M_4 is positive – a fairly typical example being $M_4 = S^1 \times W_3$. We will also write M_6 for a fairly general six-manifold, although we will soon concentrate on the case $M_6 = M_4 \times D$ for a two-manifold D.

We make one change in notation from the earlier part of this paper. In Section 2, to emphasize that the starting point was a physically sensible, unitary boundary condition for the D3–NS5 system, we started in Lorentz signature and labeled the coordinates of the D3 world-volume as $x^0, \dots, x^3$. After establishing some basics, we then Wick rotated to Euclidean signature (Section 2.1.1), still labeling the coordinates the same way. But in Section 4, we introduced a new coordinate by T-duality, and it is natural to think of this as the time coordinate. To make "room" for labeling the new time coordinate as x^0, we relabel the four "old" coordinates by $x^\mu \to x^{\mu+1}$. The main consequence is that when we do gauge theory on a five-dimensional half-space, starting in Section 5.2, the coordinate normal to the boundary of the half-space will be $y = x^4$, and not x^3 as earlier in this paper.

5.1. Four-dimensional topological field theory from six dimensions

5.1.1. Basics. The basic idea is to construct a four-dimensional topological field theory by twisting of the six-dimensional $(0, 2)$ superconformal field theory associated to a simple and simply-laced Lie group[27] G. (The idea of twisting was briefly described in Section 2.2.1.) The R-symmetry group of this theory is $\mathrm{SO}(5)_R$ or more precisely its double cover $\mathrm{Spin}(5)_R$. As there is no non-trivial homomorphism from $\mathrm{Spin}(6)$ (the structure group of the spin bundle of a generic six-manifold) to $\mathrm{Spin}(5)_R$, there is no way to construct a six-dimensional topological field theory by twisting of the six-dimensional $(0, 2)$ model. However, it is possible to construct topological field theories in dimension five or less.

The specific construction that we want gives a four-dimensional topological field theory. We use the fact that $\mathrm{Spin}(5)_R$ contains a subgroup

$$\mathrm{U} = (\mathrm{Spin}(3) \times \mathrm{Spin}(2))/\mathbb{Z}_2 \subset \mathrm{Spin}(5)_R. \tag{5.1}$$

We specialize to six-manifolds of the form $M_6 = M_4 \times D$, where M_4 is an oriented

[27] To be more precise, the six-dimensional theory is associated to the Dynkin diagram of G rather than to the choice of a specific global form of the group G (such as the adjoint group or its simply-connected cover). In particular, the six-dimensional theory does not distinguish G from $G^\vee$; in the simply-laced case, they are two global forms of the same group. On a six-manifold X, this theory has a family of partition functions labeled by the quantization of a finite Heisenberg group associated to $H^3(X, \mathcal{Z})$; here $\mathcal{Z} = \Gamma^\vee/\Gamma$, with Γ the root lattice of G and $\Gamma^\vee$ its dual. Within this family, one can make a choice that on reduction to five dimensions leads to a desired global form of G; on further reduction to four dimensions, the choices that lead to G or $G^\vee$ are exchanged by S-duality. The details, which are described in [107], will not be important in the present paper.

four-manifold and D is an oriented[28] two-manifold. The structure group of the Riemannian (spin) connection of M_6 reduces to the subgroup

$$\mathrm{V} = (\mathrm{Spin}(4) \times \mathrm{Spin}(2))/\mathbb{Z}_2 \subset \mathrm{Spin}(6). \tag{5.2}$$

Furthermore, we have the exceptional isomorphism

$$\mathrm{Spin}(4) \cong \mathrm{Spin}(3)_\ell \times \mathrm{Spin}(3)_r. \tag{5.3}$$

So it is possible to define a homomorphism

$$\upsilon\colon \mathrm{V} \longrightarrow \mathrm{Spin}(5) \tag{5.4}$$

that annihilates $\mathrm{Spin}(3)_\ell$ and maps $(\mathrm{Spin}(3)_r \times \mathrm{Spin}(2))/\mathbb{Z}_2$ isomorphically onto U. We define a subgroup V' of $\mathrm{Spin}(6) \times \mathrm{Spin}(5)_R$, isomorphic to V:

$$\mathrm{V}' = (1 \times \upsilon)(\mathrm{V}). \tag{5.5}$$

(In the action of V', a spacetime rotation by a group element $\mathrm{v} \in \mathrm{V}$ is combined with an R-symmetry transformation $\upsilon(\mathrm{v})$.)

In a standard fashion, we can define a twisted version of the $(0, 2)$ model on $M_4 \times D$ in which the spin connection couples to the currents that generate V', rather than V. For generic M_4, the unbroken supersymmetries of the twisted model correspond to the V'-invariant supersymmetries that the model has if formulated on $\mathbb{R}^6$. A standard group-theoretic exercise, starting with the fact that the global supersymmetries of the $(0, 2)$ model transform under $\mathrm{Spin}(6) \times \mathrm{Spin}(5)_R$ as $\mathbf{4}_+ \otimes \mathbf{4}_R$ (where $\mathbf{4}_+$ is a positive chirality spinor of Spin(6), and $\mathbf{4}_R$ is a spinor of $\mathrm{Spin}(5)_R$), shows that there is just one V'-invariant supersymmetry generator, which we will call Q. Q transforms as a non-trivial character of $\mathrm{Spin}(2)_R$, and we normalize the generator F of $\mathrm{Spin}(2)_R$ so that

$$[\mathrm{F}, Q] = Q. \tag{5.6}$$

Q also obeys

$$Q^2 = 0; \tag{5.7}$$

indeed, if not zero, Q^2 would be a universally defined Killing vector field on $M_4 \times D$.

Once we restrict to the cohomology of Q, the theory obtained this way is a topological field theory on M_4, but varies holomorphically with the complex moduli of D. One can understand this without detailed computation as follows. First, compactify from six to four dimensions on D, making a $\mathrm{Spin}(2)_R$ twist to preserve supersymmetry. This leads to a four-dimensional theory with $\mathcal{N} = 2$ supersymmetry. The remaining R-symmetry group is the subgroup of $\mathrm{Spin}(5)_R$ that commutes with its $\mathrm{Spin}(2)_R$ subgroup; this is precisely U, which is isomorphic to $(\mathrm{SU}(2) \times \mathrm{U}(1))/\mathbb{Z}_2 = \mathrm{U}(2)$, the usual R-symmetry group of an $\mathcal{N} = 2$ superconformal field theory in four

[28] The orientation of M_4 is necessary to enable us to make a consistent choice of $\mathrm{Spin}(3)_r$ in (5.3). Given this, D must be oriented because the $(0, 2)$ model is only defined on an oriented six-manifold.

dimensions. Indeed, if D is a compact Riemann surface without boundary (possibly with punctures), compactification from six dimensions on D with a supersymmetric twist gives a four-dimensional superconformal gauge theory [36]; the gauge group is semi-simple and the coupling parameters τ_i of its simple factors are the moduli of D.

Now that we are in four dimensions with $\mathcal{N} = 2$ supersymmetry, there is an essentially unique R-symmetry twist, resulting from the identification of $\mathrm{Spin}(3)_r$ with the corresponding subgroup of U. This leads to a four-dimensional topological field theory by the same reasoning as in [101]. The observables of this theory are computed by counting instanton solutions and hence they depend holomorphically on the instanton counting factors $q_i = \exp(2\pi i \tau_i)$, that is, on the moduli of D. Thus, reduction of the six-dimensional theory on $M_4 \times D$ with an R-symmetry twist that preserves supersymmetry gives a theory that is topological on M_4 but varies holomorphically with the moduli of D.

5.1.2. Brane construction. For the case that G is of A or D type, and with favorable choices of M_4 and D, this construction has a realization via M5-branes. Just as in Section 2.2.3, this brane realization is highly informative though not completely general.

We use the fact that the $(0, 2)$-model of type A_{r-1} arises at low energies on a system of r parallel M5-branes supported on $\mathbb{R}^6 \subset \mathbb{R}^{11}$. In this description, the R-symmetry group $\mathrm{Spin}(5)_R$ acts by rotations of the normal bundle to $\mathbb{R}^6$. To construct a topological field theory, we simply replace $\mathbb{R}^6$ by $M_4 \times D$, twisting the normal bundle to maintain supersymmetry. To get the model of type D_r, we make an orbifold version of the same construction, starting with $2r$ M5-branes and dividing by a $\mathbb{Z}_2$-symmetry that acts as -1 on the normal bundle to the M5-branes.

We let X be the total space of the bundle $\Omega^{2,+}(M_4)$ of self-dual two-forms on M_4, and let $Y = T^*D$ be the cotangent bundle of D. Ideally, we would like to endow X and Y with complete metrics of holonomy, respectively, G_2 and $\mathrm{SU}(2)$ – conditions that will maintain supersymmetry. Having done so, we consider M-theory on the product $\mathcal{X} = X \times Y$. Then the low energy limit[29] of r M5-branes wrapped on $M_4 \times D$ will give a realization of the $(0, 2)$ model of type[30] A_r on that manifold with the R-symmetry twist described above. In this description, the R-symmetry twist of Section 5.1.1 arises geometrically from the twisting of the normal bundle to $M_4 \times D$ in $\mathcal{X}$.

Alternatively, we consider M-theory on $\mathcal{X}/\mathbb{Z}_2 = (X \times Y)/\mathbb{Z}_2$, where the non-trivial element of $\mathbb{Z}_2$ leaves fixed $M_4 \times D$ and acts as -1 on the normal bundle to this space. Wrapping $2r$ M5-branes on $M_4 \times D$ and taking the low energy limit, we get now a realization of the $(0, 2)$ model of type D_r.

[29] One reaches this low energy limit by scaling up the metric of $\mathcal{X}$ so that the radius of curvature becomes much greater than the natural M-theory length scale.

[30] Taking account of the center of mass motion of the M5-branes, one actually gets a $\mathrm{U}(r)$ rather than $\mathsf{A}_{r-1} = \mathrm{SU}(r)$ theory; that is, one gets a theory that upon compactification on a circle reduces at low energy to $\mathrm{U}(r)$ gauge theory.

What has just been described is less than a general construction because the desired complete metrics of special holonomy only exist for special choices of M_4 and D. For example (see [15] and [43]), the requisite metrics of G_2 holonomy exist if M_4 is S^4 or $\mathbb{CP}^2$, while for $D = S^2$, the Eguchi–Hansen hyper-Kähler metric is suitable. (In the main example of this paper, D is an open disc with a cigar-like metric and the Taub-NUT metric has the right properties.) Actually, existence of such complete metrics is convenient, but is not necessary for any construction we will make. For one thing, in the M-theory context, all we really care about is the local structure of $\mathcal{X} = X \times Y$ near $M_6 = M_4 \times D$ and any M-theory solution with the appropriate local structure will do. For many choices of M_4 and D, $M_4 \times D$ can be embedded as a supersymmetric cycle in some $X \times Y$ where X and Y are as described above locally near M_4 and D but not globally.

More fundamentally, what we will really study is the six-dimensional $(0, 2)$ model on M_6 with the R-symmetry twist described in Section 5.1.1; this has its own life independently of how it can be embedded in M-theory.

The utility of the M-theory embedding for the present paper is largely that it helps to motivate some constructions and to make obvious the outcome of some field theory computations. We will not consider results that depend on actual existence of an M-theory embedding of $M_4 \times D$. (We do make some arguments that are local on M_4 and use the fact that D can be embedded in a Taub-NUT or Eguchi–Hansen space.) When an M-theory embedding exists, it can lead to further results, as shown strikingly in [82], [75], and [53], by analysis of geometric transitions that do follow from a string/M-theory embedding.

5.1.3. Surface operators. In the twisted $(0, 2)$ model described in Section 5.1.1, we want to include surface operators while preserving the topological symmetry.

The six-dimensional $(0, 2)$ theory has half-BPS surface operators. The simplest example (see [10] and [48]) arises from the fact that an M2-brane can end on a system of parallel M5-branes [93]. (For generalizations, see Section 5.1.4.) As above, we write M_6 for the world-volume of the M5-branes. M_6 is contained in an M-theory spacetime M_{11}. We consider an M2-brane whose world-volume is a three-manifold $P_3 \subset M_{11}$; we assume that the boundary of P_3 is a two-manifold $\Sigma_2 \subset M_6$. P_3 is oriented, so Σ_2 is also. Taking the low energy limit of such a configuration gives us the $(0, 2)$ model of type A or D in the presence of a surface operator. This surface operator depends on the "direction" with which P_3 ends on Σ_2.

Let us specialize to the case $M_6 = M_4 \times D$, embedded in the M-theory spacetime $\mathcal{X} = X \times Y$ as described in Section 5.1.2. For a generic choice of $\Sigma_2 \subset M_6 = M_4 \times D$, the topological supersymmetry of the model is broken. However, it is preserved if we pick $\Sigma_2 = \Sigma_2' \times p$, with Σ_2' an oriented two-manifold in M_4 and p a point in D, and also pick P_3 correctly.

To pick P_3, we proceed as follows (in analogy with the construction in [82] of a Lagrangian brane associated to a knot). Consider a point $q \in \Sigma_2'$. The oriented tangent plane to Σ_2' at q determines a non-zero two-form on M_4 at q, which we can

take to be normalized in a natural metric. Projecting this two-form to its self-dual part, we get a non-zero unit vector $v \in \Omega^{2,+}(M_4)|_q$ (that is, in the fiber at q of the bundle $\Omega^{2,+}(M_4)$ of self-dual two-forms on M_4). But $\Omega^{2,+}(M_4)$ is the normal bundle to M_4 in X, so v determines a ray in the fiber at q of that normal bundle. (What we have just done is to identify the trivial summand ε of (5.59).) The union of all these rays for $q \in \Sigma_2'$ gives a three-manifold $P_3' \subset X$, with boundary Σ_2'. We take the support of our M2-brane to be $P_3 = P_3' \times p$.

The key point is that an M2-brane supported on P_3 does preserve the same supersymmetry as an M5-brane supported on $M_4 \times D$. This can be understood as an exercise in G_2 structures. The tangent space to X at the point q is a copy of $\mathbb{R}^7$, with a G_2 structure defined by a three-form Υ. Choosing on $\mathbb{R}^7$ suitable coordinates x^a, $a = 1, \dots, 7$ and setting $x^{a+7} = x^a$, we have

$$\begin{aligned}\Upsilon &= \sum_{a=1}^{7} \mathrm{d}x^a \wedge \mathrm{d}x^{a+1} \wedge \mathrm{d}x^{a+3} \\ &= \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^4 + \cdots + \mathrm{d}x^3 \wedge \mathrm{d}x^4 \wedge \mathrm{d}x^6 + \cdots .\end{aligned} \tag{5.8}$$

A supersymmetric three-cycle $U_3 \subset \mathbb{R}^7$ is a three-cycle whose volume form coincides with the restriction of Υ; similarly, a supersymmetric four-cycle $R_4 \subset \mathbb{R}^7$ is one whose volume form coincides with the restriction of $\star\Upsilon$. For example, the three-manifold U_3 defined by vanishing of x^3, x^5, x^6, x^7, and so parametrized by x^1, x^2, x^4, is a supersymmetric three-cycle. Similarly, the four-manifold R_4 defined by vanishing of x^3, x^4, x^6, and so parametrized by x^1, x^2, x^5, x^7, is a supersymmetric four-cycle. So branes wrapped on U_3 and R_4 both preserve the supersymmetry that is associated to the G_2 structure. The geometrical relation between U_3 and R_4 is essentially that between P_3' and M_4 as defined earlier. Indeed, setting $M_4 = R_4$, we can identify $X = \mathbb{R}^7$ as $\Omega^{2,+}(M_4)$, and then the G_2 structure coming from Υ coincides with the natural one on $\Omega^{2,+}(M_4)$. In this picture, Σ_2' corresponds to the intersection $U_3 \cap R_4$, and is the subspace of M_4 parametrized by x^1 and x^2. Finally, P_3' is the half-space in U_3 defined by $x^4 \geq 0$.

This ensures that, for any choice of $p \in D$, an M2-brane supported on $P_3 = P_3' \times p$ preserves the same supersymmetry as a system of M5-branes on $M_4 \times D$.

We have presented this construction as if $M_4 \times D$ has an M-theory embedding in $\mathcal{X} = X \times T^*D$. The construction of the half-BPS surface operator does not really depend on this, but only on the section v of $\Omega^{2,+}(M_4)|_{\Sigma_2'}$ that is described above. It is helpful to recall the simplest construction of supersymmetric Wilson operators in $\mathcal{N} = 4$ super Yang–Mills theory. The most simple such operator for a loop K and representation R is

$$\mathrm{Tr}_R \, P \exp \oint_K (A + i\vec{n} \cdot \vec{\varphi} \, \mathrm{d}s), \tag{5.9}$$

where $\vec{\varphi}$ are the adjoint-valued scalar fields of the $\mathcal{N} = 4$ theory, $\vec{n}$ is a unit vector in the space of these scalar fields, and $\mathrm{d}s$ is the geodesic length element along K. The

section v is the analog of $\vec{n}$ in the six-dimensional $(0,2)$ theory, though in this theory one does not have a description by classical fields that would make it possible to write a formula analogous to (5.9).

5.1.4. General construction of surface operators. What we considered in Section 5.1.3 is the most obvious example of a surface operator in the $(0,2)$ model, associated with the boundary of an M2-brane that ends on M5-branes. This gives a surface operator in the $(0,2)$ model of type A. Upon compactification on a circle, if the support of the surface operator wraps the circle, such a surface operator will turn into a Wilson line operator in the fundamental representation of the appropriate A group; in the opposite case, it turns into an 't Hooft operator with minimal nonzero magnetic charge, supported on a two-dimensional surface.[31]

For our applications, we would like to know which Wilson and 't Hooft operators in five-dimensional super Yang–Mills theory (associated with what representations or magnetic charges) arise in this way by compactifying a half-BPS surface operator in six dimensions. In this paper, we will assume that all Wilson and 't Hooft operators arise like that, though this statement goes somewhat beyond what has been established in the literature. In what follows, we indicate some of the known facts.

Large classes of surface operators have been constructed,[32] [17], [20], [77], and [28], in some cases somewhat implicitly, for the models of type A_{N-1}, using the realization of these models via M-theory on $\mathrm{AdS}_7 \times S^4$, with N units of flux on S^4:

$$\int_{S^4} \frac{G}{2\pi} = N. \tag{5.10}$$

Here $G = \mathrm{d}C$ is the curvature of the M-theory three-form field C. These constructions all have better understood and more extensively studied analogs, [45], [110], and [44], for line operators in $\mathcal{N} = 4$ super Yang–Mills theory that are derived from branes in $\mathrm{AdS}_5 \times S^5$.

One basic construction [17] uses an M5-brane supported on $\Theta = \mathrm{AdS}_3 \times S^3 \subset \mathrm{AdS}_7 \times S^4$. (The M5-brane can be regarded as a bound state of several parallel M2-branes, which polarize to an M5-brane via a Myers effect [81]. The support of the surface operator is, as usual, given by the asymptotic behavior of Θ at the boundary of AdS_7.) Here AdS_3 is linearly embedded in AdS_7 in an obvious sense. And S^3 is embedded in S^4 as follows. We view S^4 as the unit sphere in $\mathbb{R}^5$. Then for some unit vector $v \in \mathbb{R}^5$ (v corresponds to the object that was denoted by the same

[31] In any dimension, a Wilson operator is defined by the holonomy of a gauge field, integrated along a curve. So Wilson operators are always supported on curves. By contrast, 't Hooft operators in gauge theory are always supported in codimension three, since an 't Hooft operator is defined, as sketched in Section 3.6.1, by a codimension three singularity. The codimension three singularity is that of a singular Dirac magnetic monopole in the three dimensions normal to the support of the 't Hooft operator. So an 't Hooft operator is supported on a point in three dimensions, a curve in four dimensions, or a two-dimensional surface in five dimensions.

[32] I thank J. Gomis for a guide to this literature and for sharing and giving me permission to summarize some of his insights.

name in Section 5.1.3), we parametrize S^3 by a point $x \in S^4$ that obeys $(v, x) = \kappa$, where $(\ ,\ )$ is the natural inner product in $\mathbb{R}^5$ and κ is a constant.

The constant κ is not arbitrary for the following reason. The M5-brane supports a two-form field whose curvature T equals the restriction to the fivebrane world-volume of C; differently put, C is trivialized when restricted to the fivebrane world-volume. This means that $\int_{S^3} C/2\pi$ must equal an integer, a condition that allows only finitely many choices of κ. Instead of discussing the gauge-dependent field C, it is convenient to let B be a closed four-ball in S^4 of boundary S^3; concretely, we define B by the inequality $(v, x) \leq \kappa$. The condition on C and κ is equivalent to integrality of

$$t = \int_B \frac{G}{2\pi}. \tag{5.11}$$

In $\mathrm{AdS}_7 \times S^4$ compactification, $G/2\pi$ is the volume form of S^4, normalized so its integral over S^4 is N. Its integral over B is positive but less than N. Hence the possible values of t are $1, 2, 3, \ldots, N-1$.

The interpretation, [17] and [20], is that upon compactification on a circle, the surface operator just described reduces to a Wilson operator associated to the t^{th} antisymmetric tensor power of the defining N-dimensional representation. We denote this representation as $\mathcal{R}_t$. The $\mathcal{R}_t$ are known as the *fundamental representations of* $\mathrm{SU}(N)$. In general, every simple Lie group G of rank r has r fundamental representations, associated to the nodes of the Dynkin diagram of G; the highest weights of these representations are called *fundamental weights*. The highest weight of any irreducible representation is a positive integer linear combination of the fundamental weights. Related to this, every irreducible representation of G appears in the algebra of tensor products of fundamental representations provided that we are willing to allow integer linear combinations with coefficients that are not necessarily positive.[33]

For applications to Khovanov homology, one would like to know if the $(0, 2)$ model has additional surface operators such that negative coefficients can be avoided. This will determine whether Khovanov homology groups can be defined for a knot labeled by an arbitrary representation of G, or only for those representations that appear in the tensor algebra of the fundamental representations without negative coefficients. In fact, for the $(0, 2)$ model of type A, there is, [20] and [77], a second construction of half-BPS surface operators with precisely the same half-BPS properties that again is based on M5-branes. The M5-brane world-volume is again $\mathrm{AdS}_3 \times S^3$, but this time $\mathrm{AdS}_3 \times S^3$ is embedded in AdS_7 (as the locus of all points a fixed distance d from an AdS_3 subspace of AdS_7) and is supported at a single point $v \in S^4$ (the same point v that entered the first construction). Surface operators of this type are believed to correspond after compactification on a circle to symmetric tensors of $\mathrm{SU}(N)$, with

[33] For example, let $\boldsymbol{R}$ be an irreducible representation of $\mathrm{SU}(N)$ described as a third rank tensor that is neither completely symmetric nor completely antisymmetric. Then $\boldsymbol{R}$ can be expressed as $\mathcal{R}_1 \otimes \mathcal{R}_2 - \mathcal{R}_3$, since it can be constructed as $\mathcal{R}_1 \otimes \mathcal{R}_2$ with the completely antisymmetric part subtracted out.

a rank determined by[34] the distance d.

More generally, a supergravity analysis [28] of half-BPS solutions of M-theory with $AdS_7 \times S^4$ asymptotics indicates that surface operators exist that are associated to an arbitrary Young tableau (fig. 2 of the paper appears to show the data of a Young tableau), or in other words (after reduction on a circle) to an arbitrary irreducible representation of $SU(N)$.

For the $(0, 2)$ model of type D_r, all of these constructions have analogs, starting with the realization of the model via M-theory on $AdS_7 \times \mathbb{RP}^4$. This may give surface operators that correspond after reduction on a circle to an arbitrary irreducible representation of D_r. Unfortunately, this sort of construction has no close analog for groups of type E.

5.1.5. $U(1)_D$-symmetry. Now we return to our six-dimensional theory on $M_6 = M_4 \times D$. For what follows, we require an action of $U(1)$ on the two-manifold D. Moreover, the theory is ultimately more interesting if the $U(1)$ action on D has a fixed point. If D is to be a complete Riemannian manifold, there are two possible choices. We can take $D = \mathbb{R}^2$, with $U(1)$ acting by rotation around a single fixed point, which we can think of as the origin in $\mathbb{R}^2$. Or we can take $D = S^2$, which admits a $U(1)$ action with two fixed points. We write $U(1)_D$ for the $U(1)$ action on D. We denote its generator as P. (When we reduce back to five dimensions in Section 5.2.1, P will turn into instanton number.)

We can define P in the quantum theory so that it commutes with the unbroken supersymmetry Q. (This condition is needed to define the quantum operator P uniquely; without it, one could add to P a multiple of F.) Thus, recalling (5.6) and (5.7), we have

$$Q^2 = 0, \quad [\mathrm{F}, Q] = Q, \quad [\mathrm{P}, Q] = 0. \tag{5.12}$$

Vanishing of Q^2 implies that one can define a cohomology of Q (on either operators or states). The commutation relations imply that F and P act on this cohomology, so the cohomology of Q is $\mathbb{Z} \times \mathbb{Z}$-graded by the eigenvalues of F and P.

In view of [53] or of arguments given earlier in this paper, we anticipate that Khovanov homology arises from the case $D = \mathbb{R}^2$. (The other choice $D = S^2$ apparently leads to a close relative of Khovanov homology, related to Chern–Simons theory with a complex gauge group; we will not explore this in the present paper.) For $D = \mathbb{R}^2$, it is convenient to endow D with a "cigar-like" metric

$$\mathrm{d}s^2 = \mathrm{d}y^2 + f(y)^2\,\mathrm{d}\psi^2, \tag{5.13}$$

where ψ is an angular variable of period 2π and $f(y)$ is a smooth, increasing function with $f(r) \sim r$ for r small and $f(r) \to$ constant for $r \to \infty$. With a suitable choice

[34] The $AdS_3 \times S^3$ solution for the M5-brane has a nonzero value of $\int_{S^3} T/2\pi$, where T is the selfdual three-form curvature that propagates on the M5-brane world-volume. One expects that Dirac quantization of the flux of T leads to a quantization condition on the possible values of d. This is somewhat analogous to quantization of the parameter t in (5.11).

of f, the cotangent bundle of D can be endowed with a complete hyper-Kähler metric, namely the Taub-NUT metric. This is convenient for the M-theory construction of Section 5.1.2. More importantly, the cigar-like nature of the metric will enable us to reduce to a gauge theory description in Section 5.2. For $D = S^2$, one can similarly regard D as a supersymmetric cycle in a hyper-Kähler manifold (the Eguchi–Hansen manifold).

The remarks of the last paragraph mean that although we cannot use the brane construction of Section 5.1.2 globally along M_4 for arbitrary M_4 (as a general M_4 is not a supersymmetric cycle in a manifold of G_2 holonomy), we can do so globally along D and locally along M_4. Indeed, locally, we approximate M_4 by $\mathbb{R}^4$, which we embed in the flat manifold $\mathbb{R}^7$, whose holonomy (being trivial) is certainly contained in G_2. Thus, to get the model of type A, we consider M-theory on

$$\mathcal{X} = \mathbb{R}^7 \times Y, \tag{5.14}$$

where Y is a hyper-Kähler manifold (Taub-NUT or Eguchi–Hansen if D is $\mathbb{R}^2$ or S^2), with M5-branes wrapped on

$$M_6 = \mathbb{R}^4 \times D, \tag{5.15}$$

D being a supersymmetric cycle in Y. For the model of type D, we similarly wrap M5-branes on $\mathcal{X}/\mathbb{Z}_2$, where $\mathbb{Z}_2$ acts as -1 on the normal bundle to M_6.

If surface operators are present, then as described in Section 5.1.3, we wish to choose them so as to preserve the $\mathrm{U}(1)_D$-symmetry as well as supersymmetry. We do this by taking the support Σ_2 of the surface operator to be $\Sigma_2' \times p$, where Σ_2' is a two-manifold in M_4 and $p \in D$ is a fixed point of the U(1) action. For example, for the case $D = \mathbb{R}^2$, surface operators are required to live at the unique fixed point of $\mathrm{U}(1)_D$, the origin in $\mathbb{R}^2$.

5.1.6. Hamiltonian description. To get Khovanov homology, we go to a Hamiltonian description. For this, we take $M_4 = \mathbb{R} \times W_3$, for some three-manifold W_3. Here $\mathbb{R}$ parametrizes the "time." The overall six-manifold is therefore now $M_6 = \mathbb{R} \times W_3 \times D$.

We write $\mathcal{H}$ for the (infinite-dimensional) physical Hilbert space of the twisted $(0, 2)$ model in this geometry. Actually, we want to consider a generalization with a surface operator included. In order to be able to construct a space of physical states in the presence of a surface operator, we wish the surface operator to have time-independent support. So in the case of a surface operator supported on $\Sigma_2 = \Sigma_2' \times p$, as in Section 5.1.5, we want $\Sigma_2' = \mathbb{R} \times K$, where $K \subset W_3$ is a knot (as usual, one can generalize to a link, that is, a disjoint union of knots) and $\mathbb{R}$ parametrizes the time. The space of physical states in this situation we designate as $\mathcal{H}_K$. We take p to be the fixed point of the $\mathrm{U}(1)_D$ action on D. In this case, $\mathcal{H}_K$ is $\mathbb{Z} \times \mathbb{Z}$-graded, because of the $\mathrm{U}(1)_R \times \mathrm{U}(1)_D$-symmetry.

The operator Q acts on $\mathcal{H}_K$. We write $\mathcal{K}(K)$ or simply $\mathcal{K}$ for the cohomology of Q, acting on $\mathcal{H}_K$. $\mathcal{K}(K)$ inherits the $\mathbb{Z} \times \mathbb{Z}$ grading of $\mathcal{H}_K$. This is the candidate for

the Khovanov homology of K. In Section 5.2, we relate the present six-dimensional description to the gauge theory description that was the basis for Section 4.

Of course, we are not limited to the case that the two-dimensional surface $\Sigma_2' \subset \mathbb{R} \times W_3$ is of the form $\mathbb{R} \times K$ with K a knot or link. A more general case, known mathematically as a *link cobordism,* was already mentioned in Section 1.2. We pick two links L and L' in $\mathbb{R} \times W_3$, and pick Σ_2' to coincide with $\mathbb{R} \times L$ in the past and with $\mathbb{R} \times L'$ in the future. Then we consider the $(0, 2)$ model on $M_6 = \mathbb{R} \times W_3 \times D$ with a surface operator on $\Sigma_2 = \Sigma_2' \times p$. This determines a $\mathrm{U}(1) \times \mathrm{U}(1)$-invariant quantum transition operator from $\mathcal{K}(L)$ to $\mathcal{K}(L')$. In other words, we get a $\mathbb{Z} \times \mathbb{Z}$-graded linear transformation

$$\Phi_{\Sigma_2} \colon \mathcal{K}(L) \longrightarrow \mathcal{K}(L'). \tag{5.16}$$

Link cobordisms can be glued together in an obvious way, and the corresponding linear transformations multiply.

Actually, the sense in which Φ_{Σ_2} is $\mathbb{Z} \times \mathbb{Z}$ graded is a little subtle. It shifts the q-grading in a way that depends on the topology and normal bundle of Σ_2. This is a known result in Khovanov homology, and will be explained from the present point of view in Section 5.4.

5.2. Gauge theory description

5.2.1. Reducing to five dimensions. Our next task is to reduce this six-dimensional description, which rests upon the mysteries of the $(0, 2)$ model, to the five-dimensional gauge theory description of Section 4.

The basic idea is simply to use the $\mathrm{U}(1)_D$-symmetry of the Riemann surface D. By standard arguments, if the metric on $M_4 \times D$ is scaled in a way that we describe momentarily, the $(0, 2)$ model on $M_4 \times D$ has a low energy description via maximally supersymmetric gauge theory on $M_4 \times D/\mathrm{U}(1)_D$.

We consider the case that D is $\mathbb{R}^2$, endowed with the cigar-like metric of (5.13):

$$\mathrm{d}s^2 = \mathrm{d}y^2 + f(y)\mathrm{d}\psi^2, \quad 0 \leq y < \infty,\ 0 \leq \psi \leq 2\pi. \tag{5.17}$$

The $\mathrm{U}(1)_D$-symmetry of D acts by constant shifts of the angular variable ψ.

While keeping fixed the metric on M_4, we multiply the metric of D by a small constant so that the asymptotic value of $f(y)$ for $y \to \infty$ becomes small. In the limit, the $(0, 2)$ model on $M_4 \times D$ has a low energy description in terms of maximally supersymmetric Yang–Mills theory on $M_4 \times \mathbb{R}_+$. Here $\mathbb{R}_+$ is the half-line $D/\mathrm{U}(1)_D$, parametrized by y.

This five-dimensional gauge theory description is actually the same one that we used in Section 4. To see this, consider the description in terms of M5-branes wrapped on $M_4 \times D \subset X \times \mathrm{TN}$, where TN is a Taub-NUT manifold in which D is embedded. $\mathrm{U}(1)_D$ acts on TN, with a unique fixed point p (which coincides with the fixed point at $y = 0$ in the action of $\mathrm{U}(1)_D$ on $D \subset \mathrm{TN}$). In the limit that the $\mathrm{U}(1)_D$ orbits are small, M-theory on $X \times \mathrm{TN}$ reduces to Type IIA superstring theory on $X \times \mathrm{TN}/\mathrm{U}(1)_D$. The

quotient $\mathrm{TN}/\mathrm{U}(1)_D$ is simply a copy of $\mathbb{R}^3$, but with a key subtlety, [95] and [49]: in the Type IIA description based on this quotient, there is a D6-brane supported on $X \times p$.

Additionally, when we reduce from M-theory to Type IIA, the M5-branes wrapped on $M_4 \times D$ become D4-branes wrapped on $M_4 \times \mathbb{R}_+$, where $\mathbb{R}_+ = D/\mathrm{U}(1)_D$ is a half-line in $\mathbb{R}^3$ that ends at p. What we have arrived at is a D4–D6 system, with D4-branes supported on $M_4 \times \mathbb{R}_+$ and ending on a D6-brane. But this is precisely the system that was investigated in Section 4. The advantage of deducing this description from a reduction of the $(0, 2)$ model in six dimensions is that the latter provides an ultraviolet completion of five-dimensional super Yang–Mills theory.

To be consistent with the notation used in Section 4.1.1 and earlier in this paper, we will denote as $G^\vee$ the gauge group of the five-dimensional description that arises by reducing on the $\mathrm{U}(1)_D$ orbits. As explained in footnote 27, it is a little subtle how the global form of $G^\vee$ (as opposed to its Lie algebra) is encoded in the six-dimensional theory. The details of this will not be important in the present paper.

5.2.2. The symmetry group. Now we have to ask how the $\mathrm{U}(1) \times \mathrm{U}(1)$-symmetry generated by P and F is realized in the gauge theory description.

Let us first consider the generator P of rotations of D. In general, when the $(0, 2)$ model is reduced on a circle, the momentum around the circle becomes instanton number in the description by five-dimensional gauge theory. (This is clear in the M-theory description. Momentum around the circle turns into D0-brane charge in Type IIA superstring theory. But, in the gauge theory of a system of Type IIA D4-branes, D0-brane charge is carried by instantons.) So P corresponds in the gauge theory description to instanton number. In the earlier part of this paper, this result was found in another way (in this other approach, the coupling of the theta-angle to instanton number in the D3–NS5 system was converted after some dualities to instanton number as a conserved charge in the D4–D6 system).

Perhaps we should clarify the precise meaning of the statement that P corresponds to instanton number. Instanton number is associated to the closed four-form $\mathrm{Tr}\, F \wedge F$, which in five dimensions is dual to a conserved current. The claim is that this is the conserved current that generates $\mathrm{U}(1)_D$-symmetry. Its integral over an initial value surface, such as a surface of fixed time in $M_5 = \mathbb{R} \times W_3 \times \mathbb{R}_+$, is a conserved quantity P. Actually in making this claim, we have to be careful, just as in Section 3.5, with the behavior at both $y = 0$ and $y = \infty$. That behavior will be analyzed in Section 5.4, and has some significant consequences. But the conserved instanton number current does lead to a $\mathbb{Z}$-grading that hopefully corresponds to the q-grading of Khovanov homology.

The topological field theory derived from twisting the $(0, 2)$ model on $M_4 \times D$ can be defined on any (oriented) M_4, but it is probably more interesting if M_4 has a suitable[35] three-cycle, leading to a four-cycle in $M_4 \times \mathbb{R}_+$. In the absence of such

[35] This three-cycle may be non-compact, as in our main example $M_4 = \mathbb{R} \times \mathbb{R}^3$, in which the three-cycle

a four-cycle, we effectively lose the grading associated with instanton number. But Khovanov homology loses much of its power if we forget the q-grading; this would be analogous roughly to taking the classical limit $q = 1$ in Chern–Simons theory.

The other conserved quantity F of the $(0, 2)$ model is the generator of an R-symmetry that is left unbroken by the twisting procedure. It has the same type of interpretation in the description by five-dimensional gauge theory.

5.2.3. Details of notation. Our next goal is to fill a major gap from Section 4 and identify the elliptic partial differential equations that are associated with supersymmetry in this problem.

Some notational preliminaries will be helpful. It is convenient to formulate maximally supersymmetric Yang–Mills theory in five dimensions via dimensional reduction from ten dimensions. This means that we combine the five components of the five-dimensional gauge field, together with five scalars in the adjoint representation, and regard them as components of a ten component "gauge field" A_I. (A_I has ten components, but they depend only on the five coordinates of $M_5 = M_4 \times \mathbb{R}_+$.) We label the five coordinates of $M_5 = M_4 \times \mathbb{R}_+$ as $x^0, x^1, \dots, x^4$, where $x^0, \dots, x^3$ parametrize M_4 and $x^4 = y$. When we specialize to $M_4 = \mathbb{R} \times W_3$, with a three-manifold W_3, we will take x^0 to parametrize $\mathbb{R}$ and call it the "time" coordinate. As for the scalars, we call them φ_I where $I = \dot{1}, \dot{2}, \dot{3}, \dot{4}, \dot{5}$. (We do not label any of the scalars as $\dot{0}$, since none will have "timelike" properties.) The curvature is defined as $F_{IJ} = [D_I, D_J]$, where D_I is a covariant derivative if $I = 0, 1, 2, 3, 4$ and otherwise D_I is one of the scalar fields φ_I.

The fermions fields λ of maximally supersymmetric Yang–Mills theory can be regarded as a positive chirality spinor field of $\mathrm{SO}(1, 9)$ with values in the adjoint representation. We write Γ^I for the gamma matrices of $\mathrm{SO}(1, 9)$; again I takes values $0, 1, 2, 3, 4$ and $\dot{1}, \dot{2}, \dot{3}, \dot{4}, \dot{5}$. Both λ and the supersymmetry generator ε obey a chirality condition. In Euclidean signature, we can take this condition to be

$$\overline{\Gamma}\lambda = -i\lambda, \quad \overline{\Gamma}\varepsilon = -i\varepsilon, \tag{5.18}$$

with $\overline{\Gamma} = \Gamma_0\Gamma_1 \dots \Gamma_4\Gamma_{\dot{1}}\Gamma_{\dot{2}} \dots \Gamma_{\dot{5}}$.

5.2.4. The boundary condition. Now we want to consider this theory on a half-space $\mathbb{R}^4 \times \mathbb{R}_+$, where $\mathbb{R}_+$ is the half-line $y \geq 0$, and we want the boundary condition at $y = 0$ that corresponds to reduction on $\mathrm{U}(1)_D$ orbits of a system of M5-branes on $\mathbb{R}^4 \times D$. In particular, this boundary condition will break the R-symmetry group $\mathrm{Spin}(5)_R$ to $(\mathrm{Spin}(3) \times \mathrm{Spin}(2))/\mathbb{Z}_2$. The Spin(3)-symmetry will later be used in maintaining some supersymmetry when $\mathbb{R}^4$ is replaced by an arbitrary four-manifold M_4.

The scalar fields φ_I represent normal fluctuations in the D4-brane position. In the context of the D4–D6 system, they play quite different roles. Three scalars,

is $\{0\} \times \mathbb{R}^3$, with $\{0\}$ a point in $\mathbb{R}$.

which we will call $\varphi_{\dot 1}$, $\varphi_{\dot 2}$, $\varphi_{\dot 3}$, describe fluctuations in the D4-brane position along the D6-brane. And the remaining two scalars, which we will call $\varphi_{\dot 4}$ and $\varphi_{\dot 5}$, describe fluctuations normal to the D6-brane.

The normal fluctuations must vanish at $y = 0$ where the D4-brane ends on the D6-brane, so the boundary conditions for the last two scalars at $y = 0$ are $\varphi_{\dot 4} = \varphi_{\dot 5} = 0$. We combine these two fields to a complex scalar field

$$\sigma = \frac{\varphi_{\dot 4} - i\varphi_{\dot 5}}{\sqrt{2}}. \tag{5.19}$$

We define a $\mathrm{Spin}(2)_R$ subgroup of the Spin(5) R-symmetry group of the theory that rotates $\varphi_{\dot 4}$ and $\varphi_{\dot 5}$ and acts trivially on the other scalars. We define the generator F of $\mathrm{Spin}(2)_R = \mathrm{U}(1)_R$ so that σ has $\mathrm{F} = 2$; the fermions then have U(1) charges ± 1. When we eventually define a topological field theory by picking a supercharge Q that obeys $Q^2 = 0$, Q will have $\mathrm{F} = 1$. The field σ will then inevitably be Q-invariant:

$$[Q, \sigma] = 0. \tag{5.20}$$

Indeed, the quantum numbers of $[Q, \sigma]$ (it has spin $1/2$, $\mathrm{F} = 3$, and dimension 3/2, and transforms in the adjoint representation of the gauge group) do not coincide with those of any elementary or composite fermion field of five-dimensional super Yang–Mills theory.

The three scalar fields that describe the motion of the D4-branes along the D6-brane have a polar behavior at $y = 0$. This polar behavior is a general property of the Dp-D$(p+2)$ system for any p and was described in the context of the D3–D5 system in Section 3.3. The polar behavior is that

$$\varphi_{\dot k} = \frac{\xi(t_k)}{y} + \cdots, \quad k = 1, 2, 3, \tag{5.21}$$

where the t_k are a standard set of $\mathfrak{su}(2)$ generators and $\xi\colon \mathfrak{su}(2) \to \mathfrak{g}$ is a principal embedding. We will combine the $\varphi_{\dot k}, k = 1, 2, 3$ to a three-vector $\vec\varphi$. (For the moment, this three-vector lives in an abstract space; it will be reinterpreted in (5.28).) One can define a subgroup Spin(3) of the R-symmetry group that rotates $\vec\varphi$. It preserves the boundary condition when combined with a gauge transformation. As expected, the boundary condition has reduced the R-symmetry group from Spin(5) to $(\mathrm{Spin}(3) \times \mathrm{Spin}(2))/\mathbb{Z}_2$.

The polar behavior of $\vec\varphi$ preserves half of the supersymmetry of the model. To describe which half, we recall that the supersymmetry transformation law for fermions is

$$\delta\lambda = \frac{1}{2}\Gamma^{IJ} F_{IJ}\varepsilon, \tag{5.22}$$

where ε is the supersymmetry generator. (As usual a symbol such as $\Gamma_{I_1 \ldots I_k}$ vanishes if two indices are equal and otherwise equals the product of the indicated gamma matrices.) Nahm's equations (3.17) for the scalar fields $\varphi_{\dot 1}$, $\varphi_{\dot 2}$, $\varphi_{\dot 3}$ can be regarded as

a selfduality condition in the four-dimensional subspace corresponding to directions $4\dot{1}\dot{2}\dot{3}$. Writing Γ_y for Γ_4, Nahm's equations preserve those supersymmetries whose generator obeys

$$\Gamma_{y\dot{1}\dot{2}\dot{3}}\varepsilon = \varepsilon. \tag{5.23}$$

The solution (5.21) of Nahm's equations preserves the supersymmetry of (5.23) for any choice of homomorphism $\xi\colon \mathfrak{su}(2) \to \mathfrak{g}$. However, in the case of D4-branes ending on a single D6-brane, the appropriate choice is that ξ is a principal embedding. More general choices of ξ correspond to D4–D6 systems with multiple D6-branes; this has been described in detail in [39]. In terms of the six-dimensional $(0, 2)$ theory, these more general choices correspond to formulating that theory on $M_4 \times D$ with a suitable defect operator (of a type considered in [36]) supported on $M_4 \times p$. These more general choices can be analyzed by methods similar to those of the present paper; they do not lead precisely to Khovanov homology, but to an interesting analog of it.

5.2.5. Twisting along M_4. So far we have described the boundary condition at $y = 0$ that breaks half of the supersymmetry and reduces the R-symmetry group to $(\mathrm{Spin}(3) \times \mathrm{Spin}(2))/\mathbb{Z}_2$. As explained in Section 5.1.1, the next step is to twist along M_4, making a Spin(3) twist so that one supersymmetry remains unbroken for an arbitrary M_4.

It is straightforward to describe this one unbroken supersymmetry. The Spin(4)-symmetry of $\mathbb{R}^4$ is generated by operators $\Gamma_{\mu\nu} = \frac{1}{2}[\Gamma_\mu, \Gamma_\nu]$ acting on spinors. When we decompose $\mathrm{Spin}(4) = \mathrm{Spin}(3)_\ell \times \mathrm{Spin}(3)_r$, the two factors are generated by the anti-selfdual and selfdual parts of $\Gamma_{\mu\nu}$, respectively. According to Section 5.1.1, the desired supersymmetry generator ε is invariant under $\mathrm{Spin}(3)_\ell$ and under a diagonal combination of $\mathrm{Spin}(3)_r$ and a group of R-symmetries; we will denote this combination as $\mathrm{Spin}(3)'_r$. The condition that ε is invariant under $\mathrm{Spin}(3)_\ell$ is that

$$(\Gamma_{01} - \Gamma_{23})\,\varepsilon = 0, \tag{5.24}$$

along with similar statements that follow by cyclic permutation of indices 123. The condition that ε is also invariant under $\mathrm{Spin}(3)'_r$ is

$$(\Gamma_{12} + \Gamma_{\dot{1}\dot{2}})\varepsilon = 0, \tag{5.25}$$

again with similar statements obtained by simultaneous cyclic permutations of indices 123 and $\dot{1}\dot{2}\dot{3}$.

The conditions (5.23), (5.24), and (5.25) have a one-dimensional space of solutions, which corresponds to the unbroken supersymmetry of the twisted model on a general M_4. For practice, let us use these conditions to determine how ε transforms under the $\mathrm{U}(1)_R = \mathrm{Spin}(2)_R$ group that commutes with the Nahm pole. Taking the generator of this symmetry to be $\mathrm{F} = i\,\Gamma_{\dot{4}\dot{5}}$, we use (5.24), (5.18), and (5.23) to deduce that

$$\mathrm{F}\varepsilon = -\varepsilon, \tag{5.26}$$

implying that the corresponding supercharge Q has $\mathrm{F} = +1$.

By standard arguments, any quantum computation in the twisted model can be localized on fields that are invariant under the topological supersymmetry. As in other models of this type, such as the twisted version of $\mathcal{N} = 2$ super Yang–Mills theory that is related to Donaldson theory, there will be equations – generalizing the instanton equations of Yang–Mills theory – that characterize what fields are invariant under this supersymmetry. The necessary condition is that the supersymmetry variations of the fermions – given in (5.22) – should vanish. In other words, we want

$$0 = \Gamma^{IJ} F_{IJ} \varepsilon. \tag{5.27}$$

Having characterized ε, we can work out the consequences of this condition.

The analysis will lead to differential equations on $M_5 = M_4 \times \mathbb{R}_+$ that will have only four-dimensional symmetry. Because of this, we introduce some notation that uses the product structure of M_5. It will be convenient to write $\Omega^{2,+}(M_4)$ for the bundle of self-dual two-forms on M_4, pulled back to M_5. An important preliminary point is that in the twisted theory, the scalar fields $\varphi_{\dot{1}}, \varphi_{\dot{2}}, \varphi_{\dot{3}}$ are best understood as a section of $\Omega^{2,+}(M_4)$, with values in the adjoint bundle $\mathrm{ad}(E)$ (derived from the underlying $G^\vee$ bundle $E \to M_5$). Thus, we define a self-dual antisymmetric tensor field B by

$$B_{0i} = \varphi_{\dot{i}}, \quad B_{ij} = \epsilon_{ijk}\varphi_{\dot{k}}, \quad i, j, k = 1, \dots, 3. \tag{5.28}$$

We regard B as a section of $\Omega^{2,+}(M_4) \otimes \mathrm{ad}(E)$. A useful fact is that $\Omega^{2,+}(M_4)$ is of rank 3, which ensures that there is a "cross product" operation on sections of $\Omega^{2,+}(M_4) \otimes \mathrm{ad}(E)$; this operation is inherited from the usual cross product for vectors in $\mathbb{R}^3$, along with the Lie algebra structure of $\mathrm{ad}(E)$. Explicitly, given B, we define a new section $B \times B$ of $\Omega^{2,+}(M_4) \otimes \mathrm{ad}(E)$ by

$$(B \times B)_{\mu\nu} = \sum_\tau [B_{\mu\tau}, B_{\nu\tau}], \tag{5.29}$$

where on the right hand side $[\ ,\ ]$ is the commutator in the Lie algebra. The right hand side of (5.29) is selfdual if B is, so in particular $B \times B$ is valued in $\Omega^{2,+}(M_4) \otimes \mathrm{ad}(E)$, as promised. One final preliminary is that given a two-form F on M_4 – such as the gauge curvature F – we define its selfdual projection $F^+ = (1 + \star)F/2$, with $\star$ the Hodge star (defined so $\star(\mathrm{d}x^0 \wedge \mathrm{d}x^1) = \mathrm{d}x^2 \wedge \mathrm{d}x^3$).

We consider first the part of (5.27) with $\mathrm{F} = -1$. It is convenient to observe that the spinors with $\mathrm{F} = -1$ transform under $\mathrm{Spin}(3)_\ell \times \mathrm{Spin}(3)'_r$ as $(1/2, 1/2) \oplus (0, 1) \oplus (0, 0)$. The $(0, 0)$ part of the equation is satisfied identically. The $(0, 1)$ part of the equation is

$$\Big(\sum_{\mu,\nu=0}^{3} \Gamma^{\mu\nu} F_{\mu\nu} + 2 \sum_{\dot{i}=1}^{3} \Gamma^{y} \Gamma^{\dot{i}} D_y \varphi_{\dot{i}} + \sum_{\dot{i},\dot{j}=1}^{3} \Gamma^{\dot{i}\,\dot{j}} [\varphi_{\dot{i}}, \varphi_{\dot{j}}] \Big) \varepsilon = 0. \tag{5.30}$$

Using the conditions obeyed by ε, the condition for this to vanish is

$$F^+ - \frac{1}{4}B \times B - \frac{1}{2}D_y B = 0. \tag{5.31}$$

To derive this formula, it is convenient to look at a particular component, say the 01 component. A part of equation (5.30) is

$$(\Gamma^{01}F_{01} + \Gamma^{23}F_{23} + \Gamma^{y\dot{1}}D_y\varphi_{\dot{1}} + \Gamma^{\dot{2},\dot{3}}[\varphi_{\dot{2}}, \varphi_{\dot{3}}])\varepsilon = 0. \tag{5.32}$$

Using (5.24), we can replace Γ^{01} by Γ^{23}; using (5.23), we can replace $\Gamma^{y\dot{1}}$ by $-\Gamma^{\dot{2}\dot{3}}$; and using (5.25), we can replace $\Gamma^{\dot{2}\dot{3}}$ by $-\Gamma^{23}$. At this stage the gamma matrices drop out and we find the equation $F_{01} + F_{23} - D_y\varphi_{\dot{1}} - [\varphi_{\dot{2}}, \varphi_{\dot{3}}] = 0$. Using the definitions of B and $B \times B$, this is equivalent to $F^+_{01} - \frac{1}{2}D_y B_{01} - \frac{1}{4}(B \times B)_{01} = 0$, which is a component of (5.31). The equation of type $(1/2, 1/2)$ can be written

$$\left(\Gamma^y\Gamma^\mu F_{y\mu} + \Gamma^\mu \sum_{k=1,2,3} \Gamma^{\dot{k}} D_\mu \varphi_{\dot{k}}\right)\varepsilon = 0. \tag{5.33}$$

Reducing this equation in a similar way to what has just been described, we arrive at

$$F_{y\mu} + \sum_{\nu=0}^{3} D^\nu B_{\nu\mu} = 0, \quad \mu = 0, \dots, 3. \tag{5.34}$$

We also need to analyze the part of (5.27) with F $= 1$. This, however, is more straightforward. We simply learn that

$$D_\mu\sigma = D_y\sigma = [B, \sigma] = 0, \tag{5.35}$$

where σ was defined in (5.19). Equation (5.35) says that a gauge transformation generated by the adjoint-valued field σ is a symmetry of the solution. Since our boundary condition at $y = 0$ forces the solution to be irreducible (and even if we relax the assumption that $\xi\colon \mathfrak{su}(2) \to \mathfrak{g}$ is a regular embedding, supersymmetry requires that $\sigma = 0$ at $y = 0$), these conditions force σ to vanish.

Having reinterpreted $\vec{\varphi}$ in the twisted theory as a section B of $\Omega^{2,+}(M) \otimes \mathrm{ad}(E)$, we should reconsider the boundary conditions at $y = 0$ that were described in Section 5.2.4. This will be done in Section 5.3.4.

5.2.6. What are these equations good for? According to (5.31) and (5.34), the equations for a supersymmetric field configuration in this theory read

$$\begin{aligned} F^+ - \frac{1}{4}B \times B - \frac{1}{2}D_y B &= 0, \\ F_{y\mu} + D^\nu B_{\nu\mu} &= 0, \end{aligned} \tag{5.36}$$

along with $\sigma = 0$. We will call these simply the supersymmetric equations.

What is one supposed to do with these equations? This question was answered in Section 4.2. Time-independent solutions of these equations supply a basis for a space $\mathcal{K}_0$ of approximate supersymmetric ground states. The actual space $\mathcal{K}$ of supersymmetric ground states is found by constructing the supercharge Q as a linear transformation of $\mathcal{K}_0$ and computing its cohomology. Concretely, Q is constructed as in (4.3) by counting time-dependent solutions of the equations (5.36) that interpolate between specified limits in the far past and future. Both $\mathcal{K}_0$ and $\mathcal{K}$ are $\mathbb{Z}\times\mathbb{Z}$-graded by the action of P and F. The eigenvalue of P is given by the classical instanton number; that of F is found by computing the charge of the filled Dirac sea of negative energy states. That is why we refer to F as fermion number, though in the full supersymmetric gauge theory it is carried by some bosons (notably σ) as well as fermions.

In Section 5.3, we will describe some useful properties of the supersymmetric equations (5.36). But it may be well to mention here their most basic property, without which the counting of solutions outlined in Section 4.2 would not make sense: they are elliptic modulo the action of the gauge group. This actually follows from the relation of these equations to the underlying super Yang–Mills theory, as we will explain in Section 5.3.3.

5.3. Some properties of the equations

5.3.1. Reductions to four dimensions. Another basic property of the equations is that they can be specialized to more familiar equations in lower dimensions.

We begin with the most obvious specialization. We can look for solutions on $M_4\times\mathbb{R}_+$ that are independent of y. We do not assume that the solution is a pullback from M_4; rather, we replace the covariant derivative D/Dy with the commutator with an adjoint-valued scalar field C. So the equations become

$$\begin{aligned} F^+ - \frac{1}{4}B\times B - \frac{1}{2}[C,B] &= 0,\\ -D_\mu C + D^\nu B_{\nu\mu} &= 0. \end{aligned} \tag{5.37}$$

These equations have been obtained previously [99] by topological twisting of four-dimensional $\mathcal{N}=4$ super Yang–Mills theory. For our purposes, we do not want to study solutions that are independent of y everywhere, because our boundary condition at $y=0$ does not allow this. However, it is natural on $M_4\times\mathbb{R}_+$ to consider solutions that are y-independent for $y\to\infty$, and thus we define our boundary condition at $y=\infty$ by specifying a solution of the equations (5.37). In the important case that $M_4=\mathbb{R}\times W_3$, we are primarily interested in boundary conditions at $y=\infty$ that are invariant under time translations. In the time-independent case, equations (5.37) describe complex-valued flat connections $\mathcal{A}=\sum_i(A_i+iB_{0i})\mathrm{d}x^i$. (This will be clear from another reduction that we describe momentarily.) So, as in most of this paper,

we define the boundary condition by specifying a complex-valued flat connection at infinity.

It is not hard to see why our supersymmetric equations (5.36), for fields that are independent of y, give an equation that can be derived from $\mathcal{N} = 4$ super Yang–Mills theory. Suppose that for our starting point, we had taken the $(0, 2)$ model on $M_6 = M_4 \times D$, with now D equal to a two-torus $\widetilde{S}^1 \times S^1$ rather than $\mathbb{R}^2$. Then, upon reducing on S^1, the same derivation would lead to the same supersymmetric equations (5.36) on $M_4 \times \widetilde{S}^1$, with y now an angular variable parametrizing $\widetilde{S}^1$. It now makes sense to take the solutions to be independent of y, and this leads to (5.37). The two-step process of reducing on first one circle and then the other amounts to the usual two-torus compactification from the $(0, 2)$ model in six dimensions to $\mathcal{N} = 4$ super Yang–Mills in four dimensions. So naturally it leads to equations that can be obtained by a topological twist of the $\mathcal{N} = 4$ theory.

There is another reduction of (5.36) that is more surprising if one simply starts with those equations, though it is obvious from the derivation we have given. This comes if we specialize to the case $M_4 = \mathbb{R} \times W_3$, for some W_3, and ask for a solution of (5.36) on $M_4 \times \mathbb{R}_+$ that is time-independent, that is invariant under translations of $\mathbb{R}$. This process amounts to undoing the lift from four to five dimensions which was the first step in Section 4. Starting with the supersymmetric equations of the D4–D6 system, if we drop the dependence on time we will get the corresponding supersymmetric equations of the D3–D5 system. We already know what these equations are, from (3.6). They are the familiar equations

$$F - \varphi \wedge \varphi + \star \mathrm{d}_A \varphi = 0 = \mathrm{d}_A \star \varphi \tag{5.38}$$

for a pair (A, φ) where A is a connection on a G-bundle $E \to W_3 \times \mathbb{R}_+$ and φ is an $\mathrm{ad}(E)$-valued one-form on $W_3 \times \mathbb{R}_+$.

To actually get these equations by a time-independent reduction of our five-dimensional ones, we proceed as follows. First of all, parametrize $\mathbb{R}$ by a time coordinate x^0 and W_3 by local coordinates x^i, $i = 1, \dots, 3$. As in the case already considered, we look for a solution on $\mathbb{R} \times W_3 \times \mathbb{R}_+$ that is invariant under translations of x^0, but we do not assume that the solution is a pullback from $W_3 \times \mathbb{R}_+$. In particular, we do not assume that A_0, the component of the connection in the x^0 direction, vanishes. Now we define an adjoint-valued one-form on $W_3 \times \mathbb{R}_+$ by

$$\varphi = \sum_{k=1}^{3} B_{0k}\, \mathrm{d}x^k - A_0\, \mathrm{d}y. \tag{5.39}$$

Notice that A_0, which was the component of the connection A in the x^0 direction, has been reinterpreted (apart from a minus sign) as what we might call φ_y, the component of the one-form φ in the y direction. Of course, this only makes sense because both the x^0 direction and the y direction have been factored out in $M_5 = M_4 \times \mathbb{R}_+ = \mathbb{R} \times W_3 \times \mathbb{R}_+$.

It is a short calculation, starting with the five-dimensional supersymmetric equations (5.36) and the definition (5.39), to arrive at the four-dimensional supersymmetric equations (5.38). The reason that this result is important is that, as explained in Section 4.2, the time-independent solutions of the supersymmetric equations (5.36) are the basis for the classical approximation $\mathcal{K}_0$ to the space $\mathcal{K}$ of supersymmetric ground states. Understanding these time-independent solutions is the starting point in studying Khovanov homology via five-dimensional gauge theory in the way described here.

Even though we had a good reason to expect the above results and they are not difficult to prove, they should give us a renewed appreciation for the fact that the five-dimensional equations (5.36) actually are elliptic. These equations can be obtained in either of two ways from an elliptic equation in four dimensions by replacing a field with a covariant derivative. We start with (5.37) and substitute $C \to D/Dy$, or we start with (5.38) and substitute $\varphi_y \to -D/Dx^0$. It is quite exceptional that starting with an elliptic differential equation and replacing one of the fields by the derivative with respect to a new variable, one arrives at an elliptic differential equation in one dimension more. However, equations (5.37) and (5.38) both have this property. From the point of view developed in the present paper, the fact that the four-dimensional equations (5.38) can be "lifted" in this sense to five dimensions is part of the reason that Chern–Simons gauge theory can be "categorified," which is just a fancy way to say that it can be derived from a theory in one dimension higher. Similarly, the fact that the four-dimensional equations (5.37) can be lifted to five dimensions means that the four-dimensional invariant given by counting solutions of those equations can be categorified. Modulo a certain vanishing theorem, this four-dimensional invariant is the Euler characteristic of instanton moduli space [99], and its categorification is, modulo the vanishing theorem and various technicalities involving the noncompactness of the moduli space, the cohomology of instanton moduli space.

5.3.2. Relation to Morse theory. The twisted version of super Yang–Mills theory that we are studying here has in general one supercharge Q when formulated on $M_5 = M_4 \times \mathbb{R}_+$. However, when we specialize to $M_4 = \mathbb{R} \times W_3$, for some W_3, the theory becomes unitary and a second supercharge appears, namely the adjoint of Q. Supersymmetric quantum mechanics with two supercharges is commonly related to Morse theory [100], and as we will now show, this is the case here.

In general, on a manifold Z, with local coordinates u^i, a metric tensor γ_{ij}, and a Morse function S, the flow equations of Morse theory read

$$\frac{\mathrm{d}u^i}{\mathrm{d}t} = -\gamma^{ij}\frac{\partial S}{\partial u^j}\,. \tag{5.40}$$

We wish to show that in the gauge $A_0 = 0$, our supersymmetric equations (5.36) can be written as such flow equations, if we pick a suitable metric on the space of fields and a suitable Morse function.

This is actually a straightforward exercise. We endow $W_3 \times \mathbb{R}_+$ with a metric $g_{ij}\mathrm{d}x^i\,\mathrm{d}x^j + \mathrm{d}y^2$. On the space of fields on $W_3 \times \mathbb{R}_+$, we define the metric

$$\mathrm{d}s^2 = -\int_{W_3\times\mathbb{R}_+} \mathrm{d}^3x\,\mathrm{d}y\,\sqrt{g}\,\mathrm{Tr}(g^{ij}\delta A_i\delta A_j + \delta A_y\delta A_y + g^{ij}\delta B_{0i}\delta B_{0j}). \quad (5.41)$$

And then we define the Morse function

$$\begin{aligned} S = -\int_{W_3\times\mathbb{R}_+} \mathrm{d}^3x\,\mathrm{d}y\,\mathrm{Tr}\Big(&\sqrt{g}g^{ij}F_{yi}B_{0j} \\ &+ \frac{1}{2}\epsilon^{ijk}\Big(A_i\partial_j A_k + \frac{2}{3}A_iA_jA_k - B_{0i}D_jB_{0k}\Big) + \sqrt{g}w\Big), \end{aligned} \quad (5.42)$$

with w a constant chosen so that the integral converges for $y \to \infty$. (The required constant of course depends on which $G^\vee_{\mathbb{C}}$-valued flat connection $\mathcal{A} = (A_i + iB_{0i})\mathrm{d}x^i$ is used to define the boundary conditions at $y = \infty$.) A straightforward computation shows that the supersymmetric equations (5.36), in the gauge $A_0 = 0$, are indeed the flow equations with S as a Morse function.

What we have just described is really the proper input for Section 4.2, in which we sketched the use of Morse theory (as extended to field theory problems in [31]) to describe the space $\mathcal{K}$ of supersymmetric ground states. The starting point is a knowledge of the time-independent solutions of the supersymmetric equations. These correspond to critical points of the Morse function S, and they furnish a basis of a space $\mathcal{K}_0$ of approximate quantum ground states. One then realizes Q as a linear transformation of $\mathcal{K}_0$ via the formula (4.3); the main step in constructing this formula is to count, with appropriate signs, the solutions of the Morse theory flow equations (5.40) that connect two given critical points. The cohomology of Q gives then the space $\mathcal{K}$ of exact supersymmetric ground states.

Because of the connection with Morse theory, the value of F associated to a given critical point has an interesting interpretation: it is the regularized Morse index of that critical point. In the case of two critical points on bundles of the same topological type (that is, two critical points with the same value of P), the difference of F at the two critical points can be computed by spectral flow. To evaluate this spectral flow, one counts the fermion states of $\mathrm{F} = 1$ or $\mathrm{F} = -1$ that pass through zero energy when one interpolates between the two critical points.

The attentive reader might notice an apparent clash between what we have said in Section 5.3.1 about time-independent solutions of the supersymmetric equations and what we have just described. In interpreting the time-independent solutions as Morse theory flow equations, the first step was to go to the gauge $A_0 = 0$. On the other hand, in Section 5.3.1, we carefully did not set A_0 to zero, and instead gave it a new name $-\varphi_y$. The resolution of this puzzle is that (5.38) is actually subject to a vanishing theorem: in a solution on $W_3 \times \mathbb{R}_+$ with the boundary conditions of interest to us, φ_y vanishes; see[36] the analysis of (4.13) in [108]. The claim that time-independent solutions of our supersymmetric equations (5.36) correspond to critical

[36] In brief, after squaring the equations, integrating, and integrating by parts, one finds that φ_y is

points depends on this vanishing theorem. Equations (5.38) are covariant and elliptic with φ_y included. If one uses the vanishing theorem to set φ_y to zero, the equations are of course no longer covariant in four dimensions; they also are not elliptic modulo the gauge group (but, assuming that one is expanding around a classical solution, they can be embedded in a larger elliptic complex). However, setting φ_y to zero makes the Morse theory interpretation of these equations clearer. This is so both for the five-dimensional equations (5.36) and for the four-dimensional equations (5.38) that were related to Morse theory in a similar way in [108].

The vanishing theorem that we just encountered has a perhaps more familiar analog for Floer theory of the space of connections on a three-manifold. If on a four-manifold of the form $\mathbb{R} \times W_3$, one looks for time-independent solutions of the instanton equation $F^+ = 0$, one gets in three dimensions the Bogomolny equations $F + \star DA_0 = 0$. These equations are the analog of (5.38); they are elliptic modulo the action of the gauge group, and they do not correspond directly to the critical points of any Morse function. However, assuming that W_3 is compact and we want nonsingular and irreducible solutions, one can deduce from the Bogomolny equations a vanishing theorem $A_0 = 0$. (The proof is made by the same sort of argument as in footnote 36.) From this vanishing theorem, one learns that the time-independent solutions of the instanton equation actually correspond to flat connections on W_3. These are the critical points of a Morse function, namely the Chern–Simons function $\mathrm{CS}(A)$. The equation $F = 0$ that we get after using the vanishing theorem is not elliptic modulo the gauge group, but it is part of a larger elliptic complex.

The Chern–Simons function $\mathrm{CS}(A)$ of standard Floer theory is not quite well defined as a real-valued function on the space of gauge fields modulo gauge transformations (but only as a circle-valued function); because of this, Floer theory is ultimately not $\mathbb{Z}$-graded by the Morse index of a critical point, but $\mathbb{Z}/4h\mathbb{Z}$-graded, where h is the dual Coxeter number of the gauge group. By contrast, in our present problem, the Morse function S is actually a well-defined real-valued function, and hence the grading by the fermion number F is an actual $\mathbb{Z}$-grading, as we have asserted throughout this paper. To verify that S is well defined, a slightly subtle point is the following. One contribution in the definition (5.42) of S is the integral over $W_3 \times \mathbb{R}_+$ of a Chern–Simons three-form (times $\mathrm{d}y$). This contribution may look dangerous since the Chern–Simons integral is not quite well defined as a real number, but we pick the constant w to cancel the limiting value of the Chern–Simons integral at $y = \infty$, and then that integral causes no further difficulties.

5.3.3. The action. By analogy with familiar facts about the equations for Yang–Mills instantons, we anticipate that the first-order supersymmetric equations (5.36) imply the second order Euler–Lagrange equations of supersymmetric Yang–Mills theory. In many examples, an efficient way to establish such a result is to square the

annihilated by a strictly positive operator. This implies vanishing of φ_y, a result that was also used in Section 3.6.2 above. Note that our φ_y is called φ_t in [108].

first-order equations, integrate over spacetime, and compare the result to the action of the underlying physical theory.

In the case at hand, setting

$$Y_{\mu\nu} = (F^+ - \frac{1}{4}B \times B - \frac{1}{2}D_y B)_{\mu\nu}, \quad Z_\mu = F_{y\mu} + D^\sigma B_{\sigma\mu}, \tag{5.43}$$

so that the supersymmetric equations are $Y = Z = 0$, we find the following identity

$$\begin{aligned} -\int_{M_4\times\mathbb{R}_+} \mathrm{d}^4x\,\mathrm{d}y\,\sqrt{g}\,\mathrm{Tr}(Y_{\mu\nu}Y^{\mu\nu} + Z_\mu Z^\mu) \\ = -\int_{M_4\times\mathbb{R}_+} \mathrm{d}^4x\,\mathrm{d}y\,\sqrt{g}\,\mathrm{Tr}\Big(\frac{1}{2}F_{\mu\nu}F^{\mu\nu} + F_{y\mu}F^{y\mu} + \frac{1}{4}(D_y B_{\mu\nu})^2 \\ + \frac{1}{4}(D_\alpha B_{\mu\nu})^2 + \frac{1}{16}(B\times B)_{\mu\nu}(B\times B)^{\mu\nu} \\ + \frac{R}{8}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}R_{\lambda\nu\mu\tau}B^{\lambda\nu}B^{\mu\tau}\Big) + \cdots . \end{aligned} \tag{5.44}$$

Here $R_{\lambda\nu\mu\tau}$ and R are the Riemann tensor and Ricci scalar of M_4; these curvature couplings are dictated by supersymmetry when M_4 becomes curved. In (5.44), the ellipses represent the omission of certain terms whose local variations vanish – both surface terms and a multiple of the instanton number evaluated on M_4. In fact, with our boundary conditions, both the volume integral on the right hand side of (5.44) and the omitted terms are divergent. Because their local variations vanish, the omitted terms do not affect the argument below.

The right hand side of (5.44) is essentially the bosonic part of the action of maximally supersymmetric Yang–Mills theory in five dimensions.[37] What do we learn from this relationship? If $Y = Z = 0$, then the left hand side of (5.44) is certainly stationary. So the right hand side is also. It follows, then, that the Euler–Lagrange equations derived from the right hand side of (5.44) are consequences of the first order supersymmetric equations. Those Euler–Lagrange equations are essentially the usual field equations of super Yang–Mills theory (with some scalar fields twisted to the two-form B, with fermions and σ omitted, and with some curvature couplings added).

We can use this relation between the first order and second order equations to show that the first order equations in question are elliptic. Linearization and gauge-fixing[38] of the equations $Y = Z = 0$ gives a linear differential operator that we may denote as $\mathcal{D}$. The *leading symbol* of $\mathcal{D}$ is given by the highest order part of $\mathcal{D}$, written in momentum space. Let us denote this leading symbol as σ. In the present example,

[37] To be more precise, (5.44) can be obtained from the super Yang–Mills action by setting two of the five scalar fields to zero, twisting the other three to a selfdual two-form B, and adding some curvature couplings that are needed to preserve some supersymmetry when M_4 is curved.

[38] It is convenient to use a "background field" version of Landau gauge, in which the fluctuation δA of the gauge field A is constrained to obey $\mathrm{d}_A \star \delta A = 0$.

$\mathcal{D}$ is a first order operator and σ is a matrix-valued linear function of the momentum. Ellipticity of a system of equations means that the leading symbol of the linearization is invertible for any nonzero (real) momentum. Letting σ^t denote the transpose of σ, certainly σ is invertible if $\sigma^t\sigma$ is. But the relation (5.44), or more exactly the relation between first order and second order equations that it implies, means that $\sigma^t\sigma$ is the leading symbol of the equations obtained by linearizing the second order equations of super Yang–Mills theory. Those equations are certainly elliptic; indeed (in the gauge mentioned in footnote 38), their leading symbol is the identity matrix multiplied by the leading symbol of the Laplace operator on scalars. That symbol is simply the function of a momentum vector p given by $f(p) = p^2$; it is nonzero for real nonzero p.

5.3.4. The boundary condition after twisting. Finally, we should reconsider the boundary conditions at $y = 0$ for the supersymmetric equations (5.36) on $M_4 \times \mathbb{R}_+$. For the special case $M_4 = \mathbb{R}^4$ without surface operators, these boundary conditions have already been described in Section 5.2.4: $\vec{\varphi}$ has a regular Nahm pole at $y = 0$. What happens now that we have reinterpreted $\vec{\varphi}$ in the twisted theory as a section B of $\Omega^{2,+}(M_4) \otimes \mathrm{ad}(E)$?

In fact, what happens is quite similar to what we have already described in one dimension less in Section 3.4. The field $\vec{\varphi}$, which was a section of TW_3 for a three-manifold W_3, has been promoted to a self-dual two-form B on a four-manifold M_4. With this change, all of the previous statements have close analogs.

Since we are interested in what happens at $y = 0$, let us write simply E for the restriction of the gauge bundle E to $M_4 \times \{y = 0\}$. Suppose first that $G^\vee = \mathrm{SO}(3)$. Let us write $B = b/y + \cdots$ near $y = 0$. Then, by virtue of the vanishing of the terms of order $1/y^2$ in the supersymmetric equations (5.36), b establishes an isomorphism between $\Omega^{2,+}(M_4)$ and $\mathrm{ad}(E)$, and this isomorphism identifies the metric on $\Omega^{2,+}(M_4)$ with that of $\mathrm{ad}(E)$. In Section 3.4, we used analogous statements, which were deduced in the same way, to identify the polar residue of $\vec{\varphi}$ with the vierbein e. Here the analogous statement is that b can be identified with the selfdual part of $e \wedge e$. Moreover, the vanishing of the term of order $1/y$ in the supersymmetric equations implies that $\mathrm{d}_A b = 0$. And this in turn implies[39] that the identification between $\Omega^{2,+}(M_4)$ and $\mathrm{ad}(E)$ given by b is covariantly constant, meaning that the restriction to M_4 of the $G^\vee$ connection A is simply the Riemannian connection on $\Omega^{2,+}(M_4)$. So just as in Section 3.4, the restriction to the boundary of the bundle E and the connection A are directly determined by the Riemannian geometry.

For any $G^\vee$, there is a similar story making use of a principal $\mathfrak{su}(2)$ embedding $\xi\colon \mathfrak{su}(2) \to \mathfrak{g}^\vee$. The restrictions of $\mathrm{ad}(E)$ and A to the boundary are obtained from $\Omega^{2,+}(M_4)$ and the Riemannian connection on it via the homomorphism ξ. (In general, depending on the global form of $G^\vee$, the construction of E itself as

[39] Once one knows that b is the selfdual part of $e \wedge e$, the analysis of the condition $\mathrm{d}_A b = 0$ to show that A is the Riemannian connection on $\Omega^{2,+}(M_4)$ is a problem that has been considered in the context of canonical quantum gravity [4].

opposed to its adjoint form may require a lift of the structure group of $\Omega^{2,+}(M_4)$ from SO(3) to Spin(3).) Similarly the polar part of B establishes an isomorphism between $\Omega^{2,+}(M_4)$ and a subbundle of $\mathrm{ad}(E)$ corresponding to $\xi(\mathfrak{su}(2)) \subset \mathfrak{g}^\vee$.

It is illuminating to consider the case that $M_4 = S^1 \times W_3$ (or $\mathbb{R} \times W_3$) with a product metric, and to look for solutions on $M_4 \times \mathbb{R}_+$ that are pulled back from $W_3 \times \mathbb{R}_+$. Equations (5.36) then reduce, according to Section 5.3.1, to the four-dimensional equations whose boundary conditions were considered in Section 3.4. And, as $\Omega^{2,+}(M_4)$ is the pullback to M_4 of TW_3, the boundary conditions that we have just described in the five-dimensional case do reduce to the four-dimensional boundary conditions of Section 3.4.

So far we have described the appropriate boundary condition away from surface operators. In the presence of surface operators, we proceed just as we did in Section 3.6. We first look at a local problem with a surface operator supported on $\Sigma_2 = \mathbb{R}^2$ linearly embedded in $M_4 = \mathbb{R}^4$. For this local problem, we find a model solution on $M_4 \times \mathbb{R}_+$ that is invariant under translations along Σ_2 and has a singularity in the normal plane to Σ_2 that is associated to a given irreducible representation R of G. Since the solution is invariant under translations of Σ_2, it is the pullback to $M_4 \times \mathbb{R}_+$ of a solution of reduced three-dimensional equations on $\mathbb{R}^2_\perp \times \mathbb{R}_+$, where $\mathbb{R}^2_\perp$ is the normal plane. But in fact, the relevant reduced equations coincide with the ones already analyzed in Section 3.6. This again follows from the statements in Section 5.3.1 about dimensional reduction. So in particular, for $G^\vee = \mathrm{SO}(3)$ or SU(2), the relevant model solutions have been fully described in Section 3.6.4.

Once the model solutions are known, a surface operator supported on a general embedded oriented two-manifold $\Sigma_2 \subset M_4$ and labeled by a representation R is defined rather as in Section 3.6: we define a boundary condition for the supersymmetric equations such that near a generic boundary point, the singular behavior is that of the regular Nahm pole, while along Σ_2 the singular behavior is that of the relevant model solution.

There is one important phenomenon that does not quite have an analog in one dimension less: the topology of Σ_2 and of its normal bundle influence the q-grading of Khovanov homology. This we consider next.

5.4. Surface operators and q-grading. In general, suppose that in five dimensions one is given a conserved current J. Then the four-form $\star J$ is a conserved charge density, and given an initial value surface Ω, we define the conserved charge

$$q = \int_\Omega \star J. \tag{5.45}$$

We are interested in the case that $\star J$ is the instanton current:

$$\star J = \frac{1}{32\pi^2} \epsilon^{\mu\nu\alpha\beta} \,\mathrm{Tr}\, F_{\mu\nu} F_{\alpha\beta}. \tag{5.46}$$

We have normalized the instanton current so that, for any simply-connected $G^\vee$, the conserved charge q takes integer values if Ω is compact and without boundary.

Let us now specialize to $M_4 = \mathbb{R} \times W_3$ and thus $M_5 = \mathbb{R} \times W_3 \times \mathbb{R}_+$. Given a conserved current J, we define a charge at time $t \in \mathbb{R}$ by integration of this four-form over the initial value surface $\{t\} \times W_3 \times \mathbb{R}_+$:

$$q(t) = \int_{\{t\}\times W_3 \times \mathbb{R}_+} \star J. \tag{5.47}$$

Is $q(t)$ independent of time? Conservation of J is not quite enough to ensure this, since current might disappear at the ends $y = 0$ and $y = \infty$. In general the change in q between initial and final times t_i and t_f is

$$q(t_f) - q(t_i) = \int_{\Delta_0(t_f,t_i)} \star J - \int_{\Delta_\infty(t_f,t_i)} \star J. \tag{5.48}$$

Here $\Delta_0(t_f, t_i)$ is defined by $y = 0$, $t_f \geq t \geq t_i$, and $\Delta_\infty(t_f, t_i)$ by $y = \infty$, $t_f \geq t \geq t_i$. Taking $t_f \to +\infty$, $t_i \to -\infty$ and writing just Δ_0 and Δ_∞ for the boundaries at $y = 0$ and $y = \infty$, the total change in the charge is

$$\Delta q = \int_{\Delta_0} \star J - \int_{\Delta_\infty} \star J. \tag{5.49}$$

In the case of the instanton current, naively the conserved charge is the instanton number

$$\mathrm{P}(t) = \frac{1}{32\pi^2} \int_{\{t\}\times W_3 \times \mathbb{R}_+} \epsilon^{\mu\nu\alpha\beta}\, \mathrm{Tr}\, F_{\mu\nu} F_{\alpha\beta}. \tag{5.50}$$

Actually, as in (3.33), to eliminate a dependence on the metric of W_3 (replacing it with a dependence on a framing of W_3), we should subtract from P a multiple of the gravitational Chern–Simons function $\mathrm{CS}_{\mathrm{grav}}$, replacing P with

$$\widehat{\mathrm{P}} = \mathrm{P} - \frac{v\, \mathrm{CS}_{\mathrm{grav}}}{8\pi}. \tag{5.51}$$

Since we will take the metric on W_3 to be time-independent, this correction term is time-independent. So the total change in $\widehat{\mathrm{P}}$ between the far past and the far future is the same as the change in P. From (5.49), it is the sum of two contributions given by the fluxes of the conserved current at $y = 0$ and $y = \infty$. In the present context, those two terms are the instanton numbers of the $G^\vee$ bundle E, restricted to $y = 0$ or $y = \infty$. We write $\mathrm{P}(y = 0)$ and $\mathrm{P}(y = \infty)$ for the instanton number evaluated at $y = 0$ or at $y = \infty$, so

$$\Delta\widehat{\mathrm{P}} = \Delta\mathrm{P} = \mathrm{P}(y = 0) - \mathrm{P}(y = \infty). \tag{5.52}$$

We want to apply this to Khovanov homology, meaning that the boundary condition at $y = \infty$ is that the connection A approaches a fixed, time-independent flat connection. This ensures that $\mathrm{P}(y = \infty) = 0$. Likewise, $\mathrm{P}(y = 0)$ will vanish if

the boundary condition at $y = 0$ is time-independent. This will happen if there are no knots at $y = 0$, since then the boundary condition says that the restriction of the connection to $y = 0$ is the pullback of the Riemannian connection on W_3. More generally, this will happen if all knots are static and time-independent, for then the boundary condition still identifies the restriction of the connection to $y = 0$ with a pullback from W_3.

We want to allow time-dependence by including a surface operator supported on a possibly time-dependent two-manifold $\Sigma_2 \subset \mathbb{R} \times W_3$. Such surface operators are associated to the knot cobordisms of Khovanov homology. To describe a transition from the Khovanov homology of a link L in the far past to the Khovanov homology of another link L' in the far future, we require that in the past Σ_2 looks like $\mathbb{R} \times L$ and in the future it looks like $\mathbb{R} \times L'$. We assume in addition that Σ_2 is an oriented, embedded surface without boundary and with no other ends apart from the ones just described. Otherwise, Σ_2 may have an arbitrary time-dependence. The quantum transition amplitude in this situation from an initial state in $\mathcal{K}(L)$ to a final state in $\mathcal{K}(L')$ will give a linear map $\Phi_{\Sigma_2}\colon \mathcal{K}(L) \to \mathcal{K}(L')$. This linear map is, in mathematical language, the morphism of Khovanov homology associated to the link cobordism Σ_2.

Including Σ_2 makes the boundary condition at $y = 0$ time-dependent, so there is no reason for $\Delta\widehat{\mathrm{P}}$ to vanish. Instead, $\Delta\widehat{\mathrm{P}}$ will simply equal $\mathrm{P}(y = 0)$, the instanton number of the bundle E restricted to $y = 0$. $\Delta\widehat{\mathrm{P}}$ is equal to the amount by which the quantum transition amplitude Φ_{Σ_2} shifts the q-grading of Khovanov homology.

The fundamental case to understand is the case that Σ_2 is compact and L and L' are empty. After treating this case in Section 5.4.1, we will reintroduce the knots in Section 5.4.2.

The problem we consider in Section 5.4.1 is somewhat like the one studied for framing of knots in Section 3.7, but it is simpler because we will be computing a characteristic class (the instanton number) rather than a secondary characteristic class (the Chern–Simons function). We will see in Section 5.4.2 that the simpler computation we do here actually implies the result of Section 3.7.

5.4.1. Compactly supported surface operator. In the following, we consider a surface operator of compact support in an arbitrary four-manifold M_4, which we regard as the boundary at $y = 0$ of $M_5 = M_4 \times \mathbb{R}_+$. We write simply Σ, rather than Σ_2, for the support of the surface operator, and we write simply E for the restriction of the gauge bundle E to M_4, that is, to $y = 0$. As in our study of knot framings, we will do this analysis for $G^\vee = \mathrm{SO}(3)$. The instanton number of E is 1/4 times the first Pontryagin class of $\mathrm{ad}(E)$:

$$\mathrm{P}(y = 0) = \frac{1}{4}\int_{M_4} p_1(\mathrm{ad}(E)). \tag{5.53}$$

(The factor of $1/4$, which corresponds to $1/2h^\vee$ in (3.11), comes from the ratio of the trace of the four-form $F \wedge F$ in the two-dimensional and three-dimensional

representations of SU(2).)

In the absence of a surface operator, ad(E) is simply $\Omega^{2,+}(M_4)$, so P$(y=0)$ can be expressed in terms of the Euler characteristic and signature of M_4. We want to determine the shift in P$(y=0)$ due to the presence of the surface operator:

$$\Delta \mathrm{P}(y=0) = \frac{1}{4}\int_{M_4}(p_1(E) - p_1(\Omega^{2,+}(M))). \tag{5.54}$$

Let us first describe the restriction to Σ of $\Omega^{2,+}(M_4)$. At a point $p \in \Sigma$, we pick an orthonormal basis of one-forms e_1, e_2 and f_1, f_2, such that the e_i are tangent to Σ and the f_j are normal to Σ. Also we orient them so that $e_1 \wedge e_2$ and $f_1 \wedge f_2$ determine the orientations of the tangent bundle $T\Sigma$ to Σ and its normal bundle $N\Sigma$, respectively, and hence the orientation of M_4 corresponds to $e_1 \wedge e_2 \wedge f_1 \wedge f_2$.

Now let us simply write down an orthonormal basis of self-dual two-forms at p. We can take one such form to be $w_1 = e_1 \wedge e_2 + f_1 \wedge f_2$. For the other two such forms, we write

$$w_2 + iw_3 = (e_1 + ie_2) \wedge (f_1 + if_2) \tag{5.55}$$

or

$$w_2 = e_1 \wedge f_1 - e_2 \wedge f_2, \quad w_3 = e_1 \wedge f_2 + e_2 \wedge f_1. \tag{5.56}$$

Clearly, w_1, w_2, and w_3 are indeed selfdual and (in a natural inner product) orthonormal.

The definition of w_1 was completely natural, so $\Omega^{2,+}(M_4)|_\Sigma$ contains a one-dimensional trivial real summand that we will call ε. As for $w_2 + iw_3$, it is best understood as lying in the fiber at $p \in \Sigma$ of a complex line bundle $\mathcal{M} \to \Sigma$. To construct this line bundle, we view $T^*\Sigma$ and $N^*\Sigma$ (the duals of $T\Sigma$ and $N\Sigma$) as rank one complex line bundles, placing on them the complex structures that act by

$$I(e_1 + ie_2) = i(e_1 + ie_2), \quad J(f_1 + if_2) = i(f_1 + if_2). \tag{5.57}$$

Evidently, $\mathcal{M} \cong T^*\Sigma \otimes_{\mathbb{C}} N^*\Sigma$, since $e_1 + ie_2$ takes values in $T^*\Sigma$ and $f_1 + if_2$ in $N^*\Sigma$. So the restriction of $\Omega^{2,+}(M_4)$ to Σ is

$$\Omega^{2,+}(M_4)|_\Sigma = \varepsilon \oplus \mathcal{M}, \tag{5.58}$$

where $\mathcal{M}$ is regarded as a real vector bundle of rank 2.

As a real bundle of rank 2, $\mathcal{M}$ is equivalent to its dual. (In fact, the Riemannian metric on M_4 gives a natural identification between them.) This means that in (5.58), we can replace $\mathcal{M}$ by $\mathcal{L} = \mathcal{M}^{-1}$. Here $\mathcal{L} = T\Sigma \otimes_{\mathbb{C}} N\Sigma = \mathcal{T} \otimes \mathcal{N}$, where we write simply $\mathcal{T}$ and $\mathcal{N}$ for $T\Sigma$ and $N\Sigma$ regarded as complex line bundles. Thus (5.58) is equivalent to $\Omega^{2,+}(M_4)|_\Sigma = \varepsilon \oplus \mathcal{L}$. A small neighborhood $\mathcal{U}$ of Σ is contractible onto Σ, and this isomorphism automatically extends over $\mathcal{U}$:

$$\Omega^{2,+}(M_4)|_{\mathcal{U}} \cong \varepsilon \oplus \mathcal{L}. \tag{5.59}$$

Now we want to modify $\Omega^{2,+}(M_4)$ along Σ by gluing in along Σ an 't Hooft operator supported on Σ and dual to the spin j representation of $G = \mathrm{SU}(2)$. In the full five-dimensional description, the support of the 't Hooft operator is on $\Sigma\times\{y = 0\} \subset M_4 \times \mathbb{R}_+$, so it is of codimension three as expected for 't Hooft operators. We denote the modified bundle as $E_{(j)}$. We can understand the structure of $E_{(j)}$ from the model solution described in Section 3.6.4 – lifted now to five dimensions rather than to four as assumed in Section 3.6. The gauge field of the model solution is $\mathrm{u}(1)$-valued (though the full model solution including the other fields is irreducible). In the context of a knot K in a three-manifold W_3, the $\mathrm{U}(1)$ in question acts on the normal bundle to K. When we lift to a surface Σ in a four-manifold M_4, the $\mathrm{U}(1)$ in question acts on the subbundle of $\Omega^{2,+}(M_4)|_\Sigma$ that is orthogonal to ε. In other words, it acts on $\mathcal{L}$.

To construct $E_{(j)}$, we are supposed to glue in $2j$ units of flux in this $\mathrm{U}(1)$ subgroup. This means that $E_{(j)}$ restricted to Σ will have the form $\varepsilon \oplus \mathcal{S}$ where $\mathcal{S}$ is a complex line bundle with the following properties: (1) Away from Σ, $\mathcal{S}$ is isomorphic to $\mathcal{L}$, ensuring that $E_{(j)}$ is equivalent to $\Omega^{2,+}(M_4)$. (2) The isomorphism between $\mathcal{L}$ and $\mathcal{S}$ has a zero along Σ of degree $2j$. This second condition captures the idea that E_j is obtained from $\Omega^{2,+}(M_4)$ by adding $2j$ units of flux in the normal direction.

The two conditions have a simple and unique solution. In general, if Σ is a Riemann surface, there is no natural way to pick a section of a complex line bundle $\mathcal{S} \to \Sigma$. But let X be the total space of the line bundle $\mathcal{S} \to \Sigma$ and let $\pi : X \to \Sigma$ be the natural projection, and pull back $\mathcal{S}$ to a line bundle $\pi^*\mathcal{S} \to X$. Then $\pi^*\mathcal{S}$ does have a natural section, which moreover has a simple zero along $\Sigma \subset X$. This section is defined as follows: for $q \in X$, define $p \in \Sigma$ by $p = \pi(q)$. Then q lies in $\mathcal{S}_p$, the fiber of $\mathcal{S}$ over p. But by the definition of pullback, $\mathcal{S}_p$ is naturally isomorphic to the fiber of $\pi^*\mathcal{S}$ over q. This isomorphism maps q to an element $s(q)$ of this fiber, and the map $q \to s(q)$ is the desired section of $\pi^*\mathcal{S} \to X$.

The most familiar example of this construction is the case that $\mathcal{S}$ is the canonical bundle K_Σ of Σ; K_Σ has no natural section, but its pullback to the total space of the fibration $K_\Sigma \to \Sigma$ does have a natural section, usually written as $p\,\mathrm{d}x$, where x is a local coordinate on Σ and p is a fiber coordinate. We note that $p\,\mathrm{d}x$ has indeed a simple zero at $p = 0$, that is, along Σ, and is nonzero for $p \neq 0$.

If s is a section of $\pi^*\mathcal{S} \to X$ with a simple zero along Σ, then s^{2j} is a section of $(\pi^*\mathcal{S})^{2j} \to X$ with a zero along Σ of degree $2j$ and no other zeroes. Moreover, up to isomorphism, $(\pi^*\mathcal{S})^{2j}$ and s^{2j} are the unique line bundle and section with these properties.

To apply this to our problem, we observe that a small neighborhood $\mathcal{U}$ of $\Sigma \subset M_4$ can be identified, in a way that is unique up to homotopy, with a neighborhood of the zero section in the total space of the fibration $\pi^*\mathcal{N} \to \Sigma$. So a line bundle over $\mathcal{U}$ that has a section vanishing in degree $2j$ along Σ and nowhere else is the pullback to $\mathcal{U}$ of $(\pi^*\mathcal{N})^{2j}$. More informally, we call this line bundle simply $\mathcal{N}^{2j}$.

So a line bundle that is isomorphic to $\mathcal{L}$ away from Σ by an isomorphism that has a zero of degree $2j$ along Σ is simply $\mathcal{L} \otimes \mathcal{N}^{2j}$. We thus arrive at a description

of $E_{(j)}$. In a neighborhood of Σ it is

$$E_{(j)}|_{\mathcal{U}} = \varepsilon \oplus \mathcal{L} \otimes \mathcal{N}^{2j} = \varepsilon \oplus \mathcal{T} \otimes \mathcal{N}^{2j+1}. \tag{5.60}$$

In general, if E is a rank three real vector bundle that is given globally as $\varepsilon \oplus \mathcal{R}$, where ε is a trivial real line bundle and $\mathcal{R}$ is a complex line bundle that we view as a real vector bundle of rank two, then $p_1(E) = c_1(\mathcal{R})^2$. So from (5.54), if the formulas (5.59) and (5.60) are valid globally on M_4, not just in a neighborhood of Σ, then the change in the instanton number due to the surface operator is

$$\Delta \mathrm{P}(y=0) = \frac{1}{4}\int_{M_4} (c_1(\mathcal{T} \otimes \mathcal{N}^{2j+1})^2 - c_1(\mathcal{T} \otimes \mathcal{N})^2). \tag{5.61}$$

It is possible for (5.59) and (5.60) to be valid globally, if $\mathcal{T}$ and $\mathcal{N}$ are suitably extended over M_4. This happens if M_4 is a complex manifold and Σ is a complex submanifold. In this case, $\Omega^{2,+}(M_4) = \varepsilon \oplus K_{M_4}$, where K_{M_4} is the canonical line bundle of M_4. As a real bundle of rank two, K_{M_4} is equivalent to the anticanonical bundle $K_{M_4}^{-1}$. When restricted to Σ, $K_{M_4}^{-1} \cong \mathcal{T} \otimes \mathcal{N}$, showing that (5.59) holds globally. Similarly (5.60) holds, with $\mathcal{N}$ interpreted as the line bundle $\mathcal{O}(\Sigma)$ whose holomorphic sections are meromorphic functions that may have a simple pole along Σ. Not only is it possible for (5.59) and (5.60) to hold globally, but this can be the case with no restriction on the topology of Σ or its normal bundle. So cases of this type must suffice to determine the general result.

Actually, one can justify (5.61) more directly without reference to the question of whether (5.59) and (5.60) may hold globally. The formal difference $E \ominus \Omega^{2,+}(M_4)$ represents a class in the K-theory of $\mathcal{U}$ with compact support (since E and $\Omega^{2,+}(M_4)$ are isomorphic on the complement of Σ). The difference between the formulas (5.59) and (5.60) is a valid formula in this K-theory with compact support, and this is enough to justify (5.61), which involves only the first Pontryagin class of $E \ominus \Omega^{2,+}(M_4)$.

As for the actual evaluation of the right hand side of (5.61), all that one needs to know is that the integral of $c_1(\mathcal{N})^2$ is $\Sigma \cap \Sigma$, the self-intersection number of Σ, and that the integral of $c_1(\mathcal{N}) \cdot c_1(\mathcal{T})$ is $\chi(\Sigma)$, the Euler characteristic of Σ. Both statements follow from the fact that $\mathcal{N}$ has a section with a simple zero along Σ. So finally the shift in the q-grading due to the surface operator is

$$\Delta \mathrm{P}(y=0) = j\ \chi(\Sigma) + j(j+1)\ \Sigma \cap \Sigma. \tag{5.62}$$

5.4.2. Transitions between knots. Now let us consider link cobordisms. For brevity in the exposition, let us assume that there are no knots in the past and there is a single knot K in the future. The generalization to arbitrary links in the past and future does not change much; the remarks that follow apply to each boundary component separately. So we take Σ to be compact toward the past and to have an end toward the future that looks like $K \times \mathbb{R}_+$. (This $\mathbb{R}_+$ is future-pointing and does not coincide with the usual $\mathbb{R}_+$ that is parametrized by y.)

Nothing changes in the above derivation provided the line bundles $\mathcal{T}$ and $\mathcal{N}$ are trivialized near the noncompact end of Σ. $\mathcal{T}$ has a natural trivialization near $t = \infty$ associated with a vector field that generates time translations along $K \times \mathbb{R}_+$. One can think of this as the reason that there is no problem to define the Euler characteristic of a noncompact Riemann surface like Σ. However, a time-independent trivialization of $\mathcal{N}$ near $t = \infty$ corresponds to a framing of Σ. If the framing of K is shifted by 1 unit, then $\Sigma \cap \Sigma$, defined relative to this trivialization, shifts by 1 unit. This shifts $\Delta P(y = 0)$ by $j(j + 1)$, so the q-grading of the final state in $\mathcal{K}(K)$ is also shifted by $j(j + 1)$. This is consistent with the fact that the expectation value of a Wilson operator supported on K in Chern–Simons theory is multiplied by $q^{j(j+1)}$ under a unit shift in framing of K, a fact that we have also explained in another way in Section 3.7.

Another interesting effect results from the term in (5.62) proportional to $\chi(\Sigma)$. For a closed Riemann surface Σ, χ is even, but for a Riemann surface ending on a single knot, χ is odd. It follows then that if j is half-integral, $\Delta P(y = 0)$ is also half-integral and the shift in q-grading in a transition from the vacuum (no knots) to a state in the Khovanov homology of a single knot is half-integral. This gives a new explanation of why the Jones polynomial of a knot (the invariant associated to $j = 1/2$) is $q^{1/2}$ times a series in (positive and negative) integer powers of q. More generally, by the same reasoning, the Jones polynomial of a link with ν components is $q^{\nu/2}$ times a series in integer powers of q.

5.5. Gauge groups that are not simply-laced

5.5.1. Preliminaries. Starting with Section 5.1.1, the groups G and $G^\vee$ have been simply-laced, for the simple reason that our main tools, the $(0, 2)$ models in six dimensions, are associated to simply-laced groups. Nonetheless, it is possible to deduce S-duality in four dimensions for a gauge group G that is not simply-laced by starting [98] with the six-dimensional model of a simply-laced group G^*. The relation between G and G^* is the same as it was in Section 4.3: G^* has an outer automorphism ζ, such that the subgroup of G^* that commutes with ζ is $G^\vee$, the dual of G. As we have seen in Section 4.3, when G is not simply-laced, there are two different Khovanov-like formulas, both presented in (4.15), that express the knot invariants of G Chern–Simons theory as traces in some space akin to Khovanov homology. Our goal here is to identify two six-dimensional constructions, starting with the $(0, 2)$ theory of type G^*, that lead to these two formulas.

The first basic fact that one needs to know is that for every pair (G^*, ζ) that appeared in Section 4.3, the $(0, 2)$ model of type G^* has ζ as a global symmetry. One way to see this is to use the unified description [104] of $(0, 2)$ models for all $\mathsf{A} - \mathsf{D} - \mathsf{E}$ groups in terms of Type IIB superstring theory at the corresponding $\mathsf{A} - \mathsf{D} - \mathsf{E}$ singularity. In all cases, ζ acts as a hyper-Kähler automorphism of the singularity of type G^* (this fact was first used in string theory in [5]) and hence as a

symmetry of the corresponding $(0,2)$ model.[40]

Before generalizing to include the automorphism ζ, let us recall the standard claim about compactification of the $(0,2)$ model of type G^* on a two-torus $\widetilde{S}^1 \times S^1$. If one formulates the $(0,2)$ model on $M_4 \times \widetilde{S}^1 \times S^1$ for some M_4, and scales down the metric of $\widetilde{S}^1$, then it reduces to supersymmetric gauge theory on $M_4 \times S^1$. The gauge group in this description is a global form of the group G^*. Which global form arises depends on a subtle choice one makes in defining the theory in six dimensions; see footnote 27. If instead one reduces on S^1, one gets a five-dimensional gauge theory based on a possibly different global form of G^* – the Langlands or GNO dual form. (This duality exchanges the center of G^* with its fundamental group, so for instance the adjoint form of the group is dual to the simply-connected form.)

Now let us repeat this discussion with ζ included. We consider the $(0,2)$ model of type G^* on $M_4 \times \widetilde{S}^1 \times S^1$, but now with a twist by ζ in going around one of the two circles. Again, we consider what happens when $\widetilde{S}^1$ is scaled down. There are two cases:

(i) If the twist is made around S^1, then the reduction on $\widetilde{S}^1$ gives five-dimensional G^* gauge theory on $M_4 \times S^1$, just as if there were no twist. But in this gauge theory description, one sees a twist by ζ in going around S^1. The twist breaks G^* down to $G^\vee$, so in four dimensions one gets $G^\vee$ gauge symmetry.

(ii) If instead the twist is made around $\widetilde{S}^1$, one gets in five dimensions gauge theory on $M_4 \times S^1$ with gauge group G, the dual of $G^\vee$. Since there is no twist around S^1, the compactification on S^1 does not affect the gauge group observed in four-dimensions at scales large compared to the radius of S^1.

Statements (i) and (ii) are related by electric-magnetic duality in four dimensions, since obviously exchanging the two circles (which is the basic operation of electric-magnetic duality) is equivalent to changing the circle around which the twist is made. Statement (ii) is used in the literature as a way to generate non-simply-laced gauge symmetry starting from M-theory or Type II superstring theory.

We need to know one more fact about the $(0,2)$ model of type G^*, beyond the fact that it admits ζ as a global symmetry. This model admits a half-BPS defect consisting of a codimension two submanifold around which all fields undergo the automorphism ζ. This fact has been briefly mentioned in [29] and exploited in [94].

[40] As has been pointed out by the author of [94], it is not true that all outer automorphisms of simply-laced groups act as hyper-Kähler automorphisms of the corresponding singularity. Rather, this is so precisely for the pairs (G^*, ζ) that are associated to groups $G^\vee$ that are not simply-laced. These pairs are $G^* = \mathsf{A}_{2n-1}$ with the automorphism of complex conjugation combined with a suitable inner automorphism (related to $G^\vee = \mathsf{C}_n = \mathrm{Sp}(2n)$), $G^* = \mathsf{D}_{2n}$ with the automorphism a reflection of one variable (related to $G^\vee = \mathsf{B}_{n-1} = \mathrm{SO}(2n-1)$), $G^* = E_6$ with its outer automorphism (related to $G^\vee = \mathsf{F}_4$), and $G^* = \mathsf{D}_4$ with an outer automorphism of order 3 (related to $G^\vee = \mathsf{G}_2$). A concise way to state the relation between these pairs is that (by the usual duality that exchanges long and short roots of the Dynkin diagram) the loop group of $G^\vee$ is GNO or Langlands dual to the ζ-twisted loop group of G^*. The example of an outer automorphism that does not arise as a hyper-Kähler symmetry of the appropriate singularity and is not related to a non-simply-laced Lie group is A_{2n} with the automorphism of complex conjugation.

5.5.2. Two constructions. Using these facts, we can now describe two six-dimensional constructions that are related to the two formulas presented in (4.15) for the knot invariants derived from Chern–Simons theory of a simple but not simply-laced Lie group G. In explaining these formulas, as in Section 4.3, G^* will be a simply-laced Lie group that possesses an outer automorphism ζ that leaves fixed $G^\vee$, the dual of G. Now, however, we will also need a simply-laced Lie group $G^\diamond$ that is related to G the way G^* is related to $G^\vee$. Thus, $G^\diamond$ admits an outer automorphism ζ' that leaves fixed G. If G is of type G_2 or F_4, then $G = G^\vee$ and $G^\diamond = G^*$. The case that $G^\diamond$ and G^* are different is that $G = \mathrm{Sp}(2n)$ and $G^\vee = \mathrm{SO}(2n+1)$ (or *vice versa*); then $G^* = \mathrm{SO}(2n+2)$ and $G^\diamond = \mathrm{SU}(2n)$.

Now we consider two constructions that will lead to the two formulas in (4.15):

(1) The first construction is familiar. We consider the $(0,2)$ model of type G^* on $M_6 = \mathbb{R} \times W_3 \times D$. We write $\mathcal{K}^*$ for its space of physical ground states. After reducing on the $\mathrm{U}(1)_D$ orbits, $\mathcal{K}^*$ can be computed by solving the supersymmetric equations (5.36) in G^* gauge theory.

(2) In the second construction, we start with the $(0,2)$ model of type $G^\diamond$, again on $M_6 = \mathbb{R} \times W_3 \times D$. Now, however, we include a defect operator associated to the outer automorphism ζ' and supported on $\mathbb{R} \times W_3 \times p$, where $p \in D$ is the $\mathrm{U}(1)_D$ fixed point. Reducing on the $\mathrm{U}(1)_D$ orbits, we get a description by supersymmetric gauge theory on $\mathbb{R} \times W_3 \times \mathbb{R}_+$ with gauge group $G^\vee$. This assertion reflects statement (ii) in Section 5.5.1, except that, since we started with $G^\diamond$ instead of G^*, the roles of G and $G^\vee$ are exchanged. In determining the gauge symmetry in this description, it suffices to consider the situation at large y, and we do not need to know what is happening at $y = 0$. However, because of the supersymmetry of the problem, we expect the boundary condition at $y = 0$ to be the usual one with the regular Nahm pole (for the five-dimensional bulk gauge group $G^\vee$, of course).

We write $\mathcal{K}$ for the space of physical ground states in construction (2). It can be obtained by studying the supersymmetric equations (5.36) in $G^\vee$ gauge theory. So in particular the spaces $\mathcal{K}^*$ and $\mathcal{K}$ that arise in our two constructions coincide with the ones that were denoted the same way in Section 4.3.

Now we compactify the time direction, possibly with a global symmetry twist:

(1′) In case (1), we replace M_6 by $S^1 \times W_3 \times D$, but making a twist by ζ around the S^1 direction. The resulting path integral on $S^1 \times W_3 \times D$ can be interpreted as a trace in $\mathcal{K}^*$. In the absence of the twist, the path integral would compute $\mathrm{Tr}_{\mathcal{K}^*} q^{\mathrm{P}}(-1)^{\mathrm{F}}$, but as we have included the twist, we get instead $\mathrm{Tr}_{\mathcal{K}^*} q^{\mathrm{P}}(-1)^{\mathrm{F}}\zeta$. This is the right hand side of one of the two formulas in (4.15).

(2′) In case (2), we again replace M_6 by $S^1 \times W_3 \times D$, but now without any twist in the S^1 direction. The path integral around S^1 now computes $\mathrm{Tr}_{\mathcal{K}} q^{\mathrm{P}}(-1)^{\mathrm{F}}$. This is the right hand side of the other formula in (4.15).

As for why these two six-dimensional constructions agree with the left hand-side of (4.15) – that is, with the path integral of Chern–Simons theory with gauge group G – we simply observe the following. In either of the two constructions, at distances

large compared to the size of S^1, we get a description by $G^\vee$ gauge theory on $W_3 \times \mathbb{R}_+$ with D3–D5 boundary conditions. Given this, we can retrace our way through the steps of Sections 3 and 2, first making an S-duality to a description by G gauge theory with D3–NS5 boundary conditions, and finally relating this to Chern–Simons theory on W_3 with gauge group G.

6. Another path to six dimensions

6.1. Overview

6.1.1. Some background. In this section, we will repeat the analysis of the present paper along a different route.

For a first orientation, let us recall some of the defect operators in gauge theories. A basic defect operator in dimension 1 is the Wilson line operator. In codimension 3, there are 't Hooft operators. These are the two types of defect operator that we have considered so far.

More obvious than the 't Hooft operator is another type of defect operator that appears in codimension 2. This is an operator associated with a prescribed monodromy. In gauge theory with gauge group G on any manifold X, let U be a submanifold of codimension 2. Let $\mathcal{C}$ be a conjugacy class in G. Then one considers gauge theory on $X \backslash U$ with the condition that the gauge fields have a monodromy around U that is in the conjugacy class $\mathcal{C}$. A surface operator supported on U is defined by asking in addition that the fields should have the mildest type of singularity consistent with this monodromy or (depending on the context) by imposing additional conditions on the singular behavior along U. We will call codimension two operators of this sort monodromy defects. We introduce this terminology because, in comparing related theories in different dimension, we want a way to emphasize the codimension rather than the dimension on which the defect is supported.

Chern–Simons theory is a theory in dimension 3, and since $3 - 2 = 1$, in this case the defect operators defined by monodromy are also line operators, just like the Wilson operators.[41] Moreover, in Chern–Simons theory, the two types of line operator are equivalent. This statement is a slight reformulation of matters explained in [102] and [105] and in much more detail in [8]. The basic reason for a relation between the two types of line operator can be seen for $G = \mathrm{U}(1)$. Consider U(1) Chern–Simons theory on a three-manifold W_3 at level k, coupled to a knot K that is

[41] Similarly, since $3 - 3 = 0$, an 't Hooft operator in a three-dimensional theory is simply a local operator. However, the Chern–Simons function $\mathrm{CS}(A)$ is not gauge-invariant in the presence of the singularity corresponding to an 't Hooft operator, and hence there are no 't Hooft operators in pure Chern–Simons theory. 't Hooft operators – which in this context are often called monopole operators – do exist in Chern–Simons theories with matter fields; see [66] and [14].

labeled by the charge n representation of U(1). The action is

$$I = -\frac{k}{4\pi}\int_{W_3} A\wedge \mathrm{d}A - n\oint_K A. \tag{6.1}$$

The equation of motion is

$$F = -\frac{2\pi n}{k}\delta_K, \tag{6.2}$$

where δ_K is a delta function that is Poincaré dual to K. This means that the gauge field A has a singularity along K, the monodromy around K being $\mathbb{M} = \exp(-2\pi i n/k)$. It is equivalent to consider Chern–Simons theory for ordinary U(1) gauge fields on W_3 with a Wilson operator of charge n on the knot K or Chern–Simons theory on W_3 for U(1) gauge fields that are required to have a singularity along K of the form (6.2).

This construction is particularly simple for $G = \mathrm{U}(1)$ because a representation is one-dimensional and a Wilson operator $\exp(in\oint_K A)$ is constructed by exponentiating a local expression that can be included in the action. The analog for a nonabelian gauge group G with a Wilson line associated to an irreducible representation R is to include in the microscopic description a matter system, supported on K, whose quantization gives the representation R. In view of the Borel–Weil–Bott theorem, such a system is the theory of maps $K\to G/T$, where the "flag manifold" G/T is endowed with a homogeneous line bundle whose first Chern class is the highest weight λ_R of the representation R. Thus, one considers a quantum theory of pairs (A,Φ), where A is a connection on a G-bundle $E\to W_3$ and Φ is a section of the G/T bundle $\mathcal{E}\to K$ that is associated to E (if E is understood as a principal G-bundle, one can set $\mathcal{E} = G/T\times_G E$).

After introducing Φ, one can gauge Φ away, since G/T is a homogeneous space, and then the equation of motion for A takes the form of (6.2) with the integer n replaced by the Lie algebra element λ_R. The monodromy around K, if computed classically, turns out to be $\mathbb{M} = \exp(-2\pi\lambda_R^*/k)$. ($\lambda_R$ is naturally an element of $\mathfrak{t}^\vee$; we have used the usual metric in which short roots have length squared two to map λ_R to an element of $\mathfrak{t}$ that we denote as λ_R^*.) It is known, however, that many formulas take their simplest form if k is replaced by $\Psi = k + h\,\mathrm{sign}\,k$ and λ_R^* by $\lambda_R^* + \varrho^*$, where ϱ is one-half the sum of the positive roots. The shift from k to Ψ has an interpretation that was explained in Section 2.4, and this interpretation indicates that all formulas of $\mathcal{N}=4$ super Yang–Mills theory should be expressed in terms of Ψ. Unfortunately, we do not have an equally clear picture of what the shift $\lambda_R\to\lambda_R+\varrho$ means in the context of $\mathcal{N}=4$ super Yang–Mills theory and hence we do not know whether this shift should be included in the microscopic formulas in this description. When we introduce the description by $\mathcal{N}=4$ super Yang–Mills theory, we will not incorporate this shift, and thus we will take the monodromy to be $\mathbb{M} = \exp(-2\pi\lambda_R^*/\Psi)$. But this is only a provisional choice and is one of many points in the present section that merit a more careful reconsideration.

6.1.2. Contents of this section. Although Wilson operators and monodromy defects are equivalent in Chern–Simons theory, they lead to two quite different pictures when we lift to four dimensions. A one-dimensional defect in three dimensions can be lifted to four dimensions as a one-dimensional defect. This is what we have done in the present paper, beginning in Section 2, in relating Wilson operators in three dimensions to Wilson or 't Hooft operators in four-dimensional gauge theory. Alternatively, a codimension two defect in three dimensions can be lifted to four dimensions as a codimension two defect. That will be our approach in the present section. The use of codimension two defects in four dimensions to describe Wilson operators in three dimensions is not essentially new; this actually was done in [108]. The motivation there was to study a semiclassical limit of Chern–Simons theory in which k and λ_R are both large, with a fixed ratio so that the monodromy $\mathbb{M}$ remains fixed. This semiclassical limit is related to the volume conjecture for Chern–Simons theory (see for instance [80] and [50]), and related developments. In the present paper, we started with Wilson operators rather than monodromy defects because this seemed to give the most direct route to Khovanov homology. However, in the present section we will describe at least the beginnings of an analogous story based on monodromy defects.

Monodromy defects in four dimensions are supported on a surface of dimension two and are often called *surface operators*. The appropriate ones were described in [55] and will be reviewed in Section 6.2, where we will also describe the basic four-dimensional construction that is related to Chern–Simons theory in this perspective. In Section 6.3, we describe the S-dual construction in four dimensions, and the resulting formulas for knot invariants, in terms of counting of solutions of elliptic differential equations. In Section 6.4, we lift the story to five dimensions, giving a description of Chern–Simons theory in terms of dimensions of vector spaces rather than counting of solutions, and in Section 6.4.3, we make the further lift to an ultraviolet-complete description in six dimensions. Finally, in Section 6.5, we attempt to use this form of the duality to actually say something about Chern–Simons knot invariants. What we are able to say is quite limited.

Thus, in brief, in the rest of this paper, we aim to recapitulate what we have done so far with Wilson operators of Chern–Simons theory replaced by the equivalent monodromy defects. But we make only the barest beginnings in this direction.

6.2. From three dimensions to four

6.2.1. Review of monodromy defects. Our first step is to relate Chern–Simons theory on a three-manifold W_3 to $\mathcal{N} = 4$ super Yang–Mills theory on $V_4 = W_3 \times \mathbb{R}_+$, but now in the presence of a monodromy defect. Just as in Section 2.4, in doing this, it is convenient to take the twisting parameter t to be real, so as to get a localization on the solutions of the elliptic differential equations $\mathcal{V}^+ = \mathcal{V}^- = \mathcal{V}^0 = 0$. And it is convenient to take the Q-invariant complex connection on the boundary of V_4 to be simply $\mathcal{A} = A + i\varphi$.

A monodromy defect supported on a knot $K \subset W_3$ will be extended to a monodromy defect in V_4. The monodromy defect is defined by specifying the singularity that fields are supposed to have along a two-dimensional surface $C \subset V_4$. For our analysis, we will take $C = K \times \mathbb{R}_+$, but more generally one may take C to be any surface in V_4 whose boundary is the original knot $K \times \{0\}$.

The singularity along C must be compatible with the localization equations $\mathcal{V}^+ = \mathcal{V}^- = \mathcal{V}^0 = 0$. In fact, the relevant monodromy defects, which have been described in [55], are half-BPS and are compatible with the localization equations for any value of the twisting parameter t.

The singular solution that defines the monodromy defect operator is a solution on $\mathbb{R}^2$ with an isolated singularity at the origin $0 \in \mathbb{R}^2$. One can think of this $\mathbb{R}^2$ as the normal plane to C. The relevant solution on $\mathbb{R}^2$ is a solution of Hitchin's equations

$$\begin{aligned} F - \varphi \wedge \varphi &= 0, \\ \mathrm{d}_A \varphi &= 0, \\ \mathrm{d}_A \star \varphi &= 0 \end{aligned} \tag{6.3}$$

for the pair (A, φ). Any solution of these equations on $\mathbb{R}^2$, when pulled back to $\mathbb{R}^4 = \mathbb{R}^2 \times \mathbb{R}^2$, obeys the four-dimensional equations $\mathcal{V}^+ = \mathcal{V}^- = \mathcal{V}^0 = 0$ for every value of t. This is related to the fact that Hitchin's equations are actually half-BPS, that is, they preserve one-half the supersymmetry of $\mathcal{N} = 4$ super Yang–Mills theory.

We will consider only the most basic monodromy defect operator considered in [55] (as opposed to refinements that depend on the choice of a non-minimal Levi subgroup of G). The defect operator has parameters $(\alpha, \beta, \gamma, \eta)$. Here α, β, and γ are elements of the Lie algebra $\mathfrak{t}$ of a maximal torus $T \subset G$ (as described later, α is more precisely an element of $\mathfrak{t}/\Lambda_{\mathrm{cochar}} = T$). Introducing polar coordinates r, θ on $\mathbb{R}^2$, the singular solution of Hitchin's equations corresponding to $\alpha, \beta, \gamma \in \mathfrak{t}$ is

$$A = \alpha \, \mathrm{d}\theta, \tag{6.4}$$

$$\varphi = \beta \frac{\mathrm{d}r}{r} - \gamma \, \mathrm{d}\theta. \tag{6.5}$$

The defect operator is defined by saying that one studies $\mathcal{N} = 4$ super Yang–Mills fields in a space of fields that coincide with this singular solution modulo less singular terms, that is, modulo terms with a singularity milder than $1/r$. As an important example of the subtlety of this definition, let us consider the case that $\alpha, \beta, \gamma \to 0$, or more generally, the case that the triple (α, β, γ) becomes nonregular. (We call this triple *regular* if the subgroup of G that leaves fixed the solution (6.4) is only the maximal torus; more generally, we say that a collection of elements of $\mathfrak{t}$, T, and/or $T^\vee$ is *regular* if the collection is not left fixed by any nontrivial element of the Weyl group.) Naively, for $\alpha, \beta, \gamma \to 0$, it seems that the singularity associated to the defect operator disappears, but the correct statement is that the limit as $\alpha, \beta, \gamma \to 0$ is

a surface operator characterized by the fact that the singularity in the fields is milder than $1/r$. The generic behavior of Hitchin's equations for $\alpha, \beta, \gamma \to 0$ is given, as found in [91], by a solution that is slightly less singular than $1/r$. (We have seen a similar behavior in Section 3.6.5; for $\lambda \to 0$, the solution (3.101) does not become regular at $z = 0$, but reduces to the solution (3.88) that has a singularity that is slightly milder than $1/|z|$.) The gauge theory surface operator with nonregular parameters must be defined to allow the same behavior, as explained in detail in [55].

The parameters α and γ in (6.4) have the following simple interpretation. By virtue of Hitchin's equations, the complex connection $\mathcal{A} = A + i\varphi$ is flat on the complement of the point $r = 0$. Its monodromy around that singular point is

$$\mathbb{M} = \exp(-2\pi(\alpha - i\gamma)). \tag{6.6}$$

The combination $\beta + i\gamma$ also has a simple interpretation. Write φ for the $(1,0)$ part of the one-form φ; then away from the singularity, φ is holomorphic by virtue of Hitchin's equations. It has a pole at $z = 0$ with polar residue $(\beta + i\gamma)/2$:

$$\varphi = \frac{1}{2}(\beta + i\gamma)\frac{\mathrm{d}z}{z}\,. \tag{6.7}$$

Because of the subtlety noted in the last paragraph, we have to be careful in interpreting these formulas if the pairs (α, γ) or (β, γ) are nonregular. For example, for $\alpha = \gamma = 0$, although the model solution has monodromy $\mathbb{M} = 1$, a generic solution that coincides with the model solution modulo terms less singular than $1/r$, and therefore is allowed in the presence of the monodromy defect, has nontrivial but unipotent monodromy (that is, $\mathbb{M} - 1$ is nilpotent but otherwise unconstrained). This is relevant in the $G^\vee$ description introduced in Section 6.3, because there the vanishing of the parameters analogous to α and γ will be natural.

The fourth parameter η has a more quantum mechanical nature. As long as (α, β, γ) is a regular triple, the presence along a surface $C \subset V_4$ of a singularity of the form (6.4) means that, along C, the structure group of the G-bundle $E \to V_4$ is reduced to T. For $G = \mathrm{SU}(2)$, this means that the structure group of $E|_C$ reduces to $T = \mathrm{U}(1)$. A U(1) bundle over a two-manifold C has a $\mathbb{Z}$-valued first Chern class c_1. We can introduce a theta-angle η and include in the path integral a factor $\exp(2\pi i \eta c_1)$. If G is of rank greater than one, then, as explained in [55], a T-bundle over C has a natural characteristic class $\mathbb{m}$ that takes values in a lattice in $\mathfrak{t}$ that is known as the *cocharacter lattice* $\Lambda_{\mathrm{cochar}}$. The generalization of a theta-angle is a homomorphism from $\Lambda_{\mathrm{cochar}}$ to U(1); we write this homomorphism as $\mathbb{m} \to \exp(2\pi i(\eta, \mathbb{m}))$, where η takes values in $\mathfrak{t}^\vee/\Lambda_{\mathrm{char}}$. Here $\mathfrak{t}^\vee$ is the dual of $\mathfrak{t}$ and $\Lambda_{\mathrm{char}} \subset \mathfrak{t}^\vee$ is the character lattice. (More informally, η is simply a collection of theta-angles, one for each U(1) subgroup of T.) Moreover, $\mathfrak{t}^\vee/\Lambda_{\mathrm{char}}$ is naturally isomorphic to the maximal torus $T^\vee$ of the GNO or Langlands dual group $G^\vee$.

Reciprocally, a gauge transformation with a singularity at $r = 0$ can shift α by an element of $\Lambda_{\mathrm{cochar}}$, so α is naturally an element of $\mathfrak{t}/\Lambda_{\mathrm{cochar}}$, which is the maximal torus $T \subset G$. The element of T corresponding to α is simply $\exp(-2\pi\alpha)$.

The quadruple of parameters $(\alpha, \beta, \gamma, \eta)$ thus take values in $T \times \mathfrak{t} \times \mathfrak{t} \times T^\vee$ or more precisely in the quotient of this space by the Weyl group of G. Under electric-magnetic duality, T and $T^\vee$ are exchanged, and $\mathfrak{t}$ is mapped to $\mathfrak{t}^\vee$. A metric on $\mathfrak{t}$ gives a map from $\mathfrak{t}$ to $\mathfrak{t}^\vee$; we use the usual metric in which short roots have length squared 2, and write β^* and γ^* for the images of β and γ in $\mathfrak{t}^\vee$. The electric-magnetic duality transformation $\tau \to -1/\mathfrak{n}_\mathfrak{g}\tau$ then maps the quadruple $(\alpha, \beta, \gamma, \eta)$ to the quadruple $(\alpha^\vee, \beta^\vee, \gamma^\vee, \eta^\vee)$ defined in [55]:

$$(\alpha^\vee, \beta^\vee, \gamma^\vee, \eta^\vee) = (\eta, |\tau|\beta^*, |\tau|\gamma^*, -\alpha). \tag{6.8}$$

If the triple (α, β, γ) is nonregular, then our definition of η does not make sense. For example, if $G = \mathrm{SO}(3)$, the only nonregular triple is $\alpha = \beta = \gamma = 0$; this leaves SO(3) unbroken and so the reduction of the structure group of $E|_C$ to T, which we assumed in the definition of η, does not hold. Nevertheless, there is a well-behaved surface operator as long as the quadruple $(\alpha, \beta, \gamma, \eta)$ is regular. For example, a surface operator with parameters $(0, 0, 0, \eta)$ is hard to define directly in terms of G gauge theory, but in the S-dual description by $G^\vee$ gauge theory, the parameters are $(\eta, 0, 0, 0)$, and now it is obvious that there is no problem as long as η is regular. An alternative description of the surface operator which makes it clear that it behaves well as long as the quadruple $(\alpha, \beta, \gamma, \eta)$ is regular is presented in Section 3 of [56]. In this approach, the surface operator is defined by coupling gauge fields on the four-manifold V_4 to a supersymmetric sigma-model that is supported on the two-manifold $C \subset V_4$. In this description, α, β, γ, and η are parameters of the sigma-model. The sigma-model becomes singular (Coulomb and Higgs branches intersect) precisely when the quadruple $(\alpha, \beta, \gamma, \eta)$ is nonregular.

In Section 6.3, we will use $G^\vee$ gauge theory to develop a semiclassical method to calculate in the presence of a monodromy defect. Even though the monodromy defect makes sense as long as the quadruple $(\alpha^\vee, \beta^\vee, \gamma^\vee, \eta^\vee)$ is regular, a semiclassical picture based on $G^\vee$ gauge theory is possible only under the stronger condition that $(\alpha^\vee, \beta^\vee, \gamma^\vee)$ is regular. So we will usually make this assumption.

6.2.2. Specialization to $V_4 = W_3 \times \mathbb{R}_+$. So far, we have considered a monodromy defect supported on an arbitrary surface C in a general four-manifold V_4. Now let us specialize to the case that $V_4 = W_3 \times \mathbb{R}_+$ with $C = K \times \mathbb{R}_+$, K being a knot in W_3. Moreover, since our interest is in Chern–Simons theory, we assume that the boundary conditions at $y = 0$ are the D3–NS5 boundary conditions discussed in Section 2, or their generalization discussed from a more purely topological field theory point of view in [109].

The starting point in relating Chern–Simons theory on W_3 to $\mathcal{N} = 4$ super Yang–Mills theory on $W_3 \times \mathbb{R}_+$ is supposed to be that, given a critical point of the Chern–Simons function on W_3, one uses this critical point to define boundary conditions at $y = \infty$ for the $\mathcal{N} = 4$ path integral. To be more precise, in the absence of a knot, a critical point is a flat bundle $E \to W_3$, and, invoking a theorem of Corlette [22], such

a flat bundle (given a mild condition of semi-stability) can be promoted to a solution of the supersymmetric equations, which in three dimensions read $F-\varphi\wedge\varphi = \mathrm{d}_A\varphi = \mathrm{d}_A\star\varphi = 0$. In the presence of a knot K labeled by parameters α, β, γ, these equations acquire delta function sources:

$$\begin{aligned} F - \varphi\wedge\varphi &= 2\pi\alpha\,\delta_K\,, \\ \mathrm{d}_A\star\varphi &= 2\pi\beta\,\mathrm{d}s\wedge\delta_K\,, \\ \mathrm{d}_A\varphi &= 2\pi\gamma\,\delta_K\,. \end{aligned} \tag{6.9}$$

In these equations, δ_K is a delta function two-form Poincaré dual to K, and $\mathrm{d}s$ is a one-form defined along K that measures the length element of K defined using the Riemannian metric on W_3. (Multiplying it by δ_K, we promote it to a closed three-form $\mathrm{d}s\wedge\delta_K$ on W_3.) A generalization of Corlette's theorem to include such singularities is apparently not known in the context of Riemannian geometry, though there are such results in the context of Kähler manifolds, the most basic case being a Riemann surface [91]. Given a solution of these equations, we use it to define initial conditions for the Morse theory flow equations at $y = \infty$. The space of solutions of the flow equations gives an integration cycle Γ for Chern–Simons theory on the boundary at $y = 0$, in the presence of a monodromy defect. This procedure has been described in [108], though without the physical interpretation by $\mathcal{N} = 4$ super Yang–Mills theory.

The $\mathcal{N} = 4$ path integral on $W_3\times\mathbb{R}_+$ with the given boundary conditions at $y = \infty$ reproduces the path integral of Chern–Simons theory on the integration cycle Γ. However, we do face the fact that, at least generically, Γ is not equivalent to any standard integration cycle of Chern–Simons theory. In our earlier analysis in which knots were associated to Wilson operators rather than monodromy defects, to partly avoid this problem, we relied on the fact that there is an important case in which there is only one possible integration cycle. This was the case $W_3 = \mathbb{R}^3$: as $\mathbb{R}^3$ is simply-connected, the Chern–Simons functional for gauge fields on $\mathbb{R}^3$ has only one critical point up to a gauge transformation, and any possible integration cycle is equivalent to the standard one. Hence results obtained by the procedure of the present paper can be compared to results of ordinary Chern–Simons theory for expectation values of knots in $\mathbb{R}^3$. As soon as we allow a monodromy defect operator supported on some $K\subset\mathbb{R}^3$, the critical point and the integration cycle are no longer unique. (This is because there typically are inequivalent flat connections over $\mathbb{R}^3\backslash K$ with prescribed monodromy around K.) We will try to find something almost as convenient as we had from the Wilson loop point of view, but this will involve some assumptions and to some extent has been included in the present paper only to orient the reader about what one might hope for.

From the point of view of Chern–Simons theory, the natural problem involving a monodromy defect was described in Section 6.1.1: it is a path integral in the space of gauge fields on W_3 that have a singularity along K with prescribed monodromy. For simplicity, we assume that the monodromy is given by a semisimple (diago-

nalizable) element $\mathbb{M} \in G_{\mathbb{C}}$ (the more general case is discussed in [108]). Then $\mathbb{M}$ can be conjugated to the complex maximal torus $T_{\mathbb{C}} \subset G_{\mathbb{C}}$ and has the form $\mathbb{M} = \exp(-2\pi(\alpha - i\gamma))$, with $\alpha, \gamma \in \mathfrak{t}$. To describe a Chern–Simons path integral for gauge fields with monodromy conjugate to $\mathbb{M}$, we must use a monodromy defect operator with α and γ as two of its parameters.

What about the other parameters β and η? We must set the parameter η to zero for the following reason. What η multiplies is supposed to be a topological invariant, which for $G = \mathrm{SU}(2)$ would be the first Chern class of a U(1) bundle over $C = K \times \mathbb{R}_+$. To define the first Chern class as a topological invariant on the non-compact Riemann surface C, one needs trivializations of the U(1) bundle at both $y = \infty$ and $y = 0$. Although our boundary condition does allow a trivialization at $y = \infty$, it does not allow a trivialization at $y = 0$, where arbitrary fluctuations in $\mathcal{A} = A + i\varphi$ are allowed. More fundamentally, the integration cycle in Chern–Simons theory defined by Morse theory flow from a critical point (or even a connected family of critical points) is connected, so there is no hope of decomposing it in components according to the values of a generalized first Chern class.

As regards the parameter β, it has no natural meaning in Chern–Simons theory. This makes one wonder if one should set β to zero, but that does not seem to be the case in general. Given a flat bundle $E \to V_4 \backslash C$, for any value of β for which we can find a solution of (6.9), we can use this to give a boundary condition on $\mathcal{N} = 4$ super Yang–Mills theory at $y = \infty$. Since β has no role in the Chern–Simons interpretation of the theory, one would expect the resulting path integral on $W_3 \times \mathbb{R}_+$ to be independent of the choice of β. A smooth deformation of the integration cycle Γ, such as one gets by varying β, should not change its homology class.

There is, however, one important situation in which β must definitely be set to zero. Suppose that $G = \mathrm{U}(1)$. Then the second equation in (6.9) reduces to $\mathrm{d}\star\varphi = 2\pi\beta\, \mathrm{d}s \wedge \delta_K$, and this equation has no solution except for $\beta = 0$. The reason for this last statement is that the closed three-form $\mathrm{d}s \wedge \delta_K$ represents a nonzero element of de Rham cohomology (its integral is the circumference of the knot K), so unless $\beta = 0$, the closed form $2\pi\beta\, \mathrm{d}s \wedge \delta_K$ cannot be written as $\mathrm{d}\star\varphi$ for any φ.

More generally, for any G, in the case of a flat bundle $E \to W_3 \backslash K$ whose monodromy reduces to an abelian subgroup of G, the same argument shows that we must take $\beta = 0$.

It seems likely that what has just been described is essentially the only obstruction to varying β away from zero, and that for example in the case of an irreducible flat $G_{\mathbb{C}}$-bundle $E \to W_3 \backslash K$, one may take arbitrary β. However, as already noted, the appropriate generalization of Corlette's theorem does not appear to be available in the literature.

Comparing the formula $\mathbb{M} = \exp(-2\pi(\alpha - i\gamma))$ to the discussion at the end of Section 6.1.1, we see that if we want to use $\mathcal{N} = 4$ super Yang–Mills theory with a monodromy defect to generate a Chern–Simons path integral (albeit on an unusual integration cycle) with a Wilson loop in the representation R, we must relate the

parameters by

$$\frac{\lambda_R^*}{\Psi} = \alpha - i\gamma. \tag{6.10}$$

Here as usual $\Psi = k + h\,\mathrm{sign}(k)$, and the formula is provisional in the sense that possibly we should replace λ_R by $\lambda_R + \varrho$. A notable fact is that, since λ_R^*, α, and γ are all elements of the real Lie algebra $\mathfrak{t}$, in order to have $\gamma \neq 0$ we must take Ψ off the real axis. In this case, $q = \exp(2\pi i/\mathfrak{n}_{\mathfrak{g}}\Psi)$ does not have modulus 1, and a description by ordinary Chern–Simons theory (in which k and Ψ are integers) is not possible. In any event, from the point of view of $\mathcal{N} = 4$ super Yang–Mills theory, we are certainly not limited to values of α, γ, and Ψ that obey a relation such as (6.10) for some representation R.

6.2.3. An important detail. In the standard perturbative expansion of Chern–Simons theory on a three-manifold W_3 around a flat connection $\mathcal{A}_\rho$ associated to a representation ρ of the fundamental group, the leading contribution in the semiclassical limit is simply the exponential of the classical action $\exp(-ik\mathrm{CS}(\mathcal{A}_\rho))$. A one-loop correction converts this to

$$Z_{\mathrm{CS}} \sim \exp(-i\Psi\mathrm{CS}(\mathcal{A}_\rho)), \tag{6.11}$$

and this is the leading behavior of the Chern–Simons partition function for large Ψ.

In the analogous calculation in $\mathcal{N} = 4$ super Yang–Mills on $V_4 = W_3 \times \mathbb{R}_+$, we use $\mathcal{A}_\rho$ to define a boundary condition at $y = \infty$. To emphasize this, in the $\mathcal{N} = 4$ context, we write $\mathcal{A}_\infty$ instead of $\mathcal{A}_\rho$. Apart from an inessential Q-exact term, the $\mathcal{N} = 4$ description differs from the Chern–Simons description by an important constant in the action – the constant $-i\,\Psi\,\mathrm{CS}(\mathcal{A}_\infty)$, which can be found in (2.69). This means that while the leading behavior of the Chern–Simons path integral expanded around a flat connection $\mathcal{A}_\rho = \mathcal{A}_\infty$ is the exponential factor (6.11), this factor is completely absent in the corresponding $\mathcal{N} = 4$ path integral: it cancels between $y = 0$ and $y = \infty$. The relation between them is

$$Z_{\mathrm{CS}} = \mathfrak{N}_0 \exp(-i\Psi\mathrm{CS}(\mathcal{A}_\infty))Z_{\mathcal{N}=4}, \tag{6.12}$$

where we allow for the possibility of a constant factor $\mathfrak{N}_0$ as in (2.66).

This is not important in studying knots in $\mathbb{R}^3$ via Wilson loops, because in that context $\mathcal{A}_\infty$ is trivial. However, when we study knots via monodromy defects, $\mathcal{A}_\infty$ has a prescribed monodromy around K and is not trivial.

In the present paper, we will consider one question for which this is important. This is the framing anomaly for knots. Under a change in framing of a knot K, Z_{CS} transforms by a power of q – the framing anomaly. But in fact, the exponential of the classical action $\exp(-i\Psi\mathrm{CS}(\mathcal{A}_\infty))$ itself has a framing anomaly. As we will now explain, in a sense most of the framing anomaly is contained in the classical action and only a quantum correction to the framing anomaly is contained in $Z_{\mathcal{N}=4}$.

Consider first the case $G = \mathrm{U}(1)$. Inserting a Wilson operator $\exp(in\oint_K A)$

in effect adds a linear term to the action, namely the second term in (6.1). Since the action is quadratic in A, once we shift to a classical solution in the presence of the knot, the linear term in the action disappears. At this point, except for an additive constant – the value of the action at the classical solution – the action coincides with what it would be in the absence of the knot, and the rest of the quantum computation proceeds as if the knot were absent. Hence, for U(1) gauge theory, the framing anomaly for knots arises entirely from the evaluation of the classical action. For a discussion of the U(1) framing anomaly in this vein, see [78], Section 2.4.

The result of the computation is that for U(1) Chern–Simons theory, the partition function transforms under a unit change in framing of a knot by

$$Z_{\rm CS} \longrightarrow Z_{\rm CS}\, q^{n^2/2} = Z_{\rm CS} \exp(\pi i \Psi m^2), \tag{6.13}$$

where we use the fact that $\Psi = k$ for U(1), and $m = n/k = n/\Psi$ is essentially the logarithm of the monodromy around the knot (that monodromy is $\mathbb{M} = \exp(-2\pi i m)$, as we explained in relation to (6.2)). We stress that this formula is purely classical in the sense that it comes entirely from evaluating the classical action.

For a general compact Lie group, the analog is

$$Z_{\rm CS} \longrightarrow Z_{\rm CS}\, q^{\mathfrak{n}_{\mathfrak{g}}(\lambda_R+2\varrho,\lambda_R)/2}, \tag{6.14}$$

where $(\ ,\ )$ is the usual inner product on $t^\vee$ in which short roots have length squared two, and $(\lambda_R, \lambda_R + 2\varrho)/2\Psi$, which reduces to $n^2/2k$ in the abelian case, is the dimension of a chiral primary field of highest weight λ_R in two-dimensional current algebra at level k. As usual, $q = \exp(2\pi i/\mathfrak{n}_{\mathfrak{g}}\Psi)$, so the factor of $\mathfrak{n}_{\mathfrak{g}}$ is absent if the formula is written in terms of Ψ.

In the same sense that the framing anomaly for knots is entirely classical in abelian gauge theory, it is mostly classical in the nonabelian case. If G is a nonabelian group, then the flat connection $\mathcal{A}_\infty$ over $W_3\backslash K$ may have nonabelian monodromy. But its restriction to a neighborhood of K in $W_3\backslash K$ is always abelian, since the fundamental group in such a neighborhood is the abelian group $\mathbb{Z}\times\mathbb{Z}$. The classical part of the framing anomaly comes only from the behavior of the classical solution near K, and can be obtained from the abelian formula (6.13) by replacing m^2 by (m,m), where m is the logarithm of the monodromy. For m we will take λ_R^*/Ψ, as explained at the end of Section 6.1.1. But this choice really needs more justification; it is not clear whether we should be making a shift $\lambda_R \to \lambda_R + \varrho$. At any rate, with our choice, we can express the factor by which Z transforms under a change in framing as

$$q^{\mathfrak{n}_{\mathfrak{g}}(\lambda_R+2\varrho,\lambda_R)/2} = \exp(\pi i \Psi(m,m)) q^{\mathfrak{n}_{\mathfrak{g}}(\lambda_R,\varrho)}. \tag{6.15}$$

On the right hand side, the first factor is classical and the second, which is subleading in the semiclassical limit (large Ψ with fixed m), is a quantum correction. However, there has been some guesswork in the way we have written the formula.

The reason that we have made this decomposition is the following. In view of (6.12), the classical part of the framing anomaly in $Z_{\rm CS}$ is contained in the factor

$\exp(-i\,\Psi\mathrm{CS}(\mathcal{A}_\infty))$. Only the quantum correction to the framing anomaly will appear in $Z_{\mathcal{N}=4}$. If therefore we accept the decomposition (6.15) at face value, then the transformation of $Z_{\mathcal{N}=4}$ under a unit change in the framing of a knot will be

$$Z_{\mathcal{N}=4} \longrightarrow Z_{\mathcal{N}=4}\, q^{\mathrm{n_g}(\lambda_R,\varrho)}. \tag{6.16}$$

6.3. The S-dual in the presence of a monodromy defect. The next step is S-duality. The gauge group is transformed from G to $G^\vee$, and the boundary condition at $y = 0$ becomes that of a D3–D5 system. The partition function can be evaluated by counting solutions of the supersymmetric equations $\mathcal{V}^+ = \mathcal{V}^- = \mathcal{V}^0 = 0$ with the appropriate elliptic boundary conditions.

In particular, the boundary condition at $y = 0$, away from the monodromy defect, is the familiar one associated with a regular Nahm pole. Near the monodromy defect, the boundary condition must be modified. As usual the corrected boundary condition is based on a model solution. The model solution should now be a solution on $\mathbb{R}^2\times\mathbb{R}_+$ of the three-dimensional reduction of our supersymmetric equations. We assume that a monodromy defect is present on the ray $\ell = p \times \mathbb{R}_+$, with p some point in $\mathbb{R}^2$. Near any point in $\mathbb{R}^2$ except p, the model solution should have a regular Nahm pole, and around any point of the ray ℓ except the endpoint at $y = 0$, it should have the singularity (6.4) of a monodromy defect. The interest in the model solution is its behavior at the exceptional point $p \times \{y = 0\}$ where ℓ meets the boundary; whatever this behavior is, we define a boundary condition by requiring this behavior where a monodromy defect meets the boundary. Happily, for $G^\vee = \mathrm{SO}(3)$, the requisite model solutions have been found, though not in complete generality, in Section 3.6.5. Equation (3.101) is the solution with $\alpha^\vee = 0$, $\beta^\vee, \gamma^\vee \neq 0$; equation (3.89) corresponds to $\alpha^\vee \neq 0$, $\beta^\vee = \gamma^\vee = 0$; and equation (3.88) exhibits the subtle behavior for $\alpha^\vee, \beta^\vee, \gamma^\vee \to 0$.

Now let us discuss what values we should take for the parameters $(\alpha^\vee, \beta^\vee, \gamma^\vee, \eta^\vee)$ in the context of topological field theory on $W_3 \times \mathbb{R}_+$. Since $\alpha^\vee$ corresponds to η in the description of Section 6.2, and in that context we had to set $\eta = 0$, we expect that we will have to set $\alpha^\vee = 0$. Indeed, there is a simple reason for this, which can be stated most briefly for $G^\vee = \mathrm{SO}(3)$. The solution (3.89) with $\alpha^\vee \neq 0$ makes perfect sense when W_3 is flat, but has a monodromy $\exp(-2\pi\alpha^\vee)$ around K. However, as we know from Section 3.4, away from a monodromy defect, the $G^\vee$ bundle $E \to W_3$ is the tangent bundle to W_3 with its Riemannian connection. For generic W_3, the Riemannian connection is irreducible and there is no way to "twist" it by a monodromy $\exp(-2\pi\alpha^\vee)$ around a knot $K \subset W_3$, while leaving it locally unchanged up to gauge transformation on the complement of K. Hence, the boundary condition of the D3–D5 system with generic W_3 and a monodromy defect only makes sense if $\alpha^\vee = 0$.

As concerns $\beta^\vee$, its status seems to be just parallel to that of β in the context of the Chern–Simons like description. At $y = \infty$, we pick a homomorphism $\rho^\vee\colon \pi_1(W_3\backslash K) \to G^\vee$, and then try to promote this to a solution of the supersymmetric equations (6.9) in the presence of the monodromy defect, now with parameters

$\alpha^\vee$, $\beta^\vee$, $\gamma^\vee$, of course. For a given $\rho^\vee$, we may use whatever $\beta^\vee$ is compatible with the equations.

To understand S-duality between the two descriptions, we need to know how the homomorphism $\rho\colon \pi_1(W_3\backslash K) \to G$ that is used to determine a boundary condition at $y = \infty$ on one side of the duality is related to the homomorphism $\rho^\vee\colon \pi_1(W_3\backslash K) \to G^\vee$ that is similarly used on the other side. We get a clue from the hypothesis that the only case in which β or $\beta^\vee$ must vanish is an abelian representation. The relation $\beta^\vee = |\tau|\beta^*$ shows that $\beta^\vee$ is constrained to vanish if and only if β is so constrained. So we are led to conjecture that $\pi_1(W_3\backslash K)$ is mapped by ρ to a commutative subgroup of G if and only if it is mapped by $\rho^\vee$ to a commutative subgroup of $G^\vee$. This conjecture is particularly powerful if $W_3 = S^3$, for then there is precisely one choice of ρ or $\rho^\vee$ with given monodromy around K and with abelian image. (This statement would not hold if we replace the knot K by a link with several components.) So in that case, the conjecture is that the abelian representation ρ is mapped to the abelian representation $\rho^\vee$.

More generally, the number of free parameters in the choice of β or $\beta^\vee$ is the rank of G minus the rank of the automorphism group of ρ or $\rho^\vee$. So a generalization of the above argument indicates that the map from ρ to $\rho^\vee$ preserves the rank of the automorphism group.

As for the other parameters, from (6.8) we have $\eta^\vee = -\alpha$, $\gamma^\vee = |\tau|\gamma^*$. In the Chern–Simons-like description, the model depends holomorphically on the logarithm of the monodromy $\alpha - i\gamma$, so in the dual description, it depends holomorphically on $\eta^\vee + i\gamma$.

An important detail is dual to the discussion of (6.7). In the $G^\vee$ description, for $\alpha^\vee = \gamma^\vee = 0$, the monodromy around K is unipotent, but not necessarily 1.

6.3.1. The partition function. In Section 3, solutions of the supersymmetric equations $\mathcal{V}^+ = \mathcal{V}^- = \mathcal{V}^0 = 0$ were labeled by the instanton number P (whose precise definition depended on a framing of both W_3 and K). The contribution of a given solution to the partition function was $(-1)^g q^{\mathrm{P}}$, where $(-1)^g$ is the sign of the fermion determinant in expanding around the given solution, P is its instanton number, and $q = \exp(2\pi i/\mathfrak{n}_{\mathfrak{g}}\Psi)$. In the present context, assuming $\beta^\vee$ and $\gamma^\vee$ are not both zero (we have set $\alpha^\vee = 0$), there is an additional topological invariant. When the $G^\vee$ bundle $E \to V_4$ is restricted to a two-manifold $C \subset V_4$, its structure group reduces to $T^\vee$, so roughly speaking it has a generalized first Chern class $\mathbb{m}^\vee$ valued in Λ_{char}. (We postpone to Section 6.3.2 some subtleties that arise if C is not compact, which is the case in our application to knots.)

How does the contribution of a given classical solution to the partition function depend on $\eta^\vee$ and $\gamma^\vee$? The dependence on $\eta^\vee$ is a simple factor of $\exp(2\pi i(\eta^\vee, \mathbb{m}^\vee)) = \exp(-2\pi i(\alpha, \mathbb{m}^\vee))$. Since the partition function is holomorphic in $\alpha - i\gamma$, the full dependence on α and γ must be a factor $\exp(-2\pi i(\alpha - i\gamma, \mathbb{m}^\vee))$. We will not show explicitly how to calculate the γ-dependence, but we expect that this will involve a computation somewhat analogous to (2.60) and (2.62): in the presence of a

monodromy defect, when one writes the action as a Q-exact term plus a topological invariant, the topological invariant includes a multiple of $(\gamma, \mathfrak{m}^\vee)$.

We can now write a formula for the partition function along the lines of (3.15). Let S be the set of solutions of the supersymmetric equations. For $s \in S$, let n_s, $\mathfrak{m}_s^\vee$, and $(-1)^{g_s}$ be the values of P, $\mathfrak{m}^\vee$, and the sign of the fermion determinant for the classical solution corresponding to s. The partition function is then

$$Z(q) = \sum_{s\in S} q^{n_s} \exp(-2\pi i(\alpha - i\gamma, \mathfrak{m}_s^\vee))(-1)^{g_s}. \tag{6.17}$$

Making use of (6.10) and the definition of q, we can write this as

$$Z(q) = \sum_{s\in S} q^{n_s - \mathfrak{n}_\mathfrak{g}(\lambda_R, \mathfrak{m}_s^\vee)}(-1)^{g_s}. \tag{6.18}$$

Alternatively, let $w_{r,\mathfrak{c}}$ be the "number" of solutions of $\mathrm{P} = r$ and $\mathfrak{m}^\vee = \mathfrak{c}$, where in computing this number we weight each solution with the sign of the fermion determinant. Then

$$Z(q) = \sum_{r,\mathfrak{c}} w_{r,\mathfrak{c}} q^{r - \mathfrak{n}_\mathfrak{g}(\lambda_R, \mathfrak{c})}. \tag{6.19}$$

These formulas have the usual proviso that λ_R should possibly be replaced by $\lambda_R + \varrho$.

To be more exact, though we have kept the notation minimal, all these formulas describe a partition function in $\mathcal{N} = 4$ supersymmetric $G^\vee$ gauge theory with a boundary condition at $y = \infty$ set by a suitable homomorphism $\rho^\vee : \pi_1(W_3 \backslash K) \to G^\vee$, and with a monodromy defect operator whose parameters are determined by the representation R of G. For some purposes, it may be best to write these formulas in terms of the logarithm of monodromy $\alpha - i\gamma$, but as they can be elegantly written in terms of λ_R, we have done so.

6.3.2. The framing anomaly revisited. For the case $V_4 = W_3 \times \mathbb{R}_+$, $C = K \times \mathbb{R}_+$, because C is not compact, the definition of $\mathfrak{m}^\vee$ depends on a trivialization of $E|_C$ at both ends of $\mathbb{R}_+$. The dependence on a choice of trivialization at $y = \infty$ means that the right topological data in fixing the boundary condition at infinity is a little more than the choice of $\rho^\vee$, but we will not say more about this.

The dependence on the trivialization at $y = 0$ leads to a framing anomaly for the $\mathcal{N} = 4$ partition function on $W_3 \times \mathbb{R}_+$ in the presence of a monodromy defect. We can see this as follows. The restriction of the $G^\vee$ bundle $E \to W_3 \times \mathbb{R}_+$ to the boundary $W_3 \times \{y = 0\}$ is the tangent bundle TW_3 of W_3, or more exactly it is the $G^\vee$ bundle associated to the SO(3) bundle TW_3 by a principal embedding $\xi : \mathfrak{su}(2) \to \mathfrak{g}^\vee$. A framing of the knot K trivializes the restriction of TW_3 to K, so it trivializes the restriction of $E|_C$ to $C \cap \{y = 0\}$. Thus a framing of K (together with whatever data was used at $y = \infty$) makes $\mathfrak{m}^\vee$ well-defined, so that we can write the formula (6.19) for the $\mathcal{N} = 4$ partition function. Under a unit change of framing of K, $\mathfrak{m}^\vee$ transforms to $\mathfrak{m}^\vee - \varrho$, and this gives the expected formula (6.16).

The statement about how $\mathbb{m}^\vee$ transforms under a change in framing amounts to the following. For $G^\vee = \mathrm{SO}(3)$, a unit change of framing shifts $\mathbb{m}^\vee$ by one unit, that is by $\varrho_{\mathrm{SU}(2)}$. (A weight of $G = \mathrm{SU}(2)$ is an element of $\mathfrak{t}^\vee = \mathfrak{t}_{\mathrm{SO}(3)}$, so in particular $\varrho_{\mathrm{SU}(2)} \in \mathfrak{t}_{\mathrm{SO}(3)}$.) A minus sign comes from comparing orientations. For general G, the homomorphism $\xi\colon \mathfrak{su}(2) \to \mathfrak{g}$ maps $\varrho_{\mathrm{SU}(2)} \in \mathfrak{t}_{\mathfrak{so}(3)}$ to $\varrho = \varrho_G \in \mathfrak{t}^\vee \subset \mathfrak{g}^\vee$ (this is a standard fact about principal $\mathfrak{su}(2)$ subalgebras), and this gives our result. But since we do not really know where the shift $\lambda_R \to \lambda_R + \varrho$ should enter in the present formalism, what we have described is more a scenario than a derivation of the framing anomaly.

6.4. Lifting to five or six dimensions

6.4.1. Five dimensions. The next step is to lift to five dimensions, following the same logic as in Section 4. We promote the solutions of the four-dimensional equations $\mathcal{V}^+ = \mathcal{V}^- = \mathcal{V}^0 = 0$ on $V_4 = W_3 \times \mathbb{R}_+$ to time-independent solutions of the five-dimensional supersymmetric equations (5.36) on $S^1 \times V_4$. Here S^1 is viewed as the time direction. We lift the monodromy defect supported on $K \times \mathbb{R}_+ \subset V_4$ to a monodromy defect supported on $S^1 \times K \times \mathbb{R}_+$.

The basic idea of a monodromy defect in five-dimensional super Yang–Mills theory on a five-manifold M_5 is similar to what it is in four dimensions, and can be described without specializing to the setting of the present paper. The support of a monodromy defect is now a three-manifold U which is of codimension two in M_5. As long as the triple of parameters $(\alpha^\vee, \beta^\vee, \gamma^\vee)$ is regular, a monodromy defect in five dimensions can be defined by postulating in the normal plane to U the same type of singularity as in (6.9). For φ in this formula, we take two of the scalar fields of five-dimensional super Yang–Mills theory. Which two depends on the context. In our application, $M_5 = M_4 \times \mathbb{R}_+$, $U = C \times \mathbb{R}_+$ for some $C \subset M_4$, and three of the scalar fields are twisted to a field $B \in \Omega^{2,+}(M_4) \otimes \mathrm{ad}(E)$. Along C, $\Omega^{2,+}(M_4)$ has the decomposition (5.59) with a two-dimensional real subbundle corresponding to $\mathcal{L}$, and the part of B valued in this subbundle is what appears in the five-dimensional analog of (6.9).

The most striking difference from four dimensions is possibly that the monodromy defect operator has no parameter corresponding to $\eta^\vee$, because the generalized first Chern class is now associated not to a spacetime history but to a physical state. In other words, if the triple $(\alpha^\vee, \beta^\vee, \gamma^\vee)$ is regular, then the bundle $E \to M_5$, when restricted to U, has abelian structure group $T^\vee$ and its curvature is a $\mathfrak{t}^\vee$-valued closed two-form f that is defined along U. Then $\star_U f$ (here $\star_U$ is the Hodge star operator for the three-manifold U) is a conserved current defined on U. Its integral on an initial value surface $C \subset U$ is a conserved quantity in the sense that it only depends on the homology class of C. We call this conserved quantity $\mathbb{m}^\vee$. (What $\mathbb{m}^\vee$ means when the triple $(\alpha^\vee, \beta^\vee, \gamma^\vee)$ is nonregular will be explained in Section 6.4.2. Technical issues in the definition of $\mathbb{m}^\vee$ involving the fact that in our application to knots, the relevant C is not compact were discussed in Section 6.3.2.)

For our application, we take $M_5 = \mathbb{R} \times W_3 \times \mathbb{R}_+$, $U = \mathbb{R} \times K \times \mathbb{R}_+$, where K is a knot in the three-manifold W_3. The space $\mathcal{K}$ of physical states defined on the initial value surface $K \times \mathbb{R}_+$ is then graded by the conserved charges P, F, and $\mathfrak{m}^\vee$.

The time-independent solutions on M_5 supply a basis for a space $\mathcal{K}_0$ of approximate supersymmetric ground states. A salient fact here – just as in the absence of the monodromy defect – is that from a four-dimensional perspective, a time-independent solution has a $\mathbb{Z}_2$-valued invariant, the sign of the fermion determinant. But from a five-dimensional perspective, this $\mathbb{Z}_2$-valued invariant is the mod 2 reduction of a $\mathbb{Z}$-valued invariant, the R-charge or fermion number F. This is a large part of the reason that the lift to five dimensions gives a richer theory than the four-dimensional one.

$\mathcal{K}_0$ is an approximation to the space $\mathcal{K}$ of exact supersymmetric ground states. To determine $\mathcal{K}$, one follows the standard recipe described in Section 4.2. One considers solutions that interpolate between different time-independent solutions in the far past and the far future. By counting such solutions in an appropriate way, one constructs the operator Q of (4.3) whose cohomology is $\mathcal{K}$.

By the same reasoning as in Section 4.2.1, we can restate (6.19) as a formula for the partition function via a trace in $\mathcal{K}$:

$$Z(q) = \mathrm{Tr}_{\mathcal{K}}\, q^{\mathrm{P}-\mathfrak{n}_{\mathfrak{g}}(\lambda_R,\mathfrak{m}^\vee)}(-1)^{\mathrm{F}}. \tag{6.20}$$

More generally, we can consider knot cobordisms interpolating between two knots K and K' by considering in $\mathbb{R} \times W_3 \times \mathbb{R}_+$ a monodromy defect supported on $C \times \mathbb{R}_+$, where $C \subset \mathbb{R} \times W_3$ is asymptotic to $\mathbb{R} \times K$ in the past and $\mathbb{R} \times K'$ in the future. Still more generally, we can replace $\mathbb{R} \times W_3$ with any oriented four-manifold M_4, and C by any oriented two-manifold in M_4.

6.4.2. The non-regular case and an action of G. The description of the monodromy defect in five dimensions via the singularity (6.9) is adequate when the triple $(\alpha^\vee, \beta^\vee, \gamma^\vee)$ is regular. For the general case, one needs a more powerful point of view.

The monodromy defect can be alternatively defined by coupling the five-dimensional $G^\vee$ gauge theory to a three-dimensional supersymmetric theory known as $T(G^\vee)$. ($T(G^\vee)$ was systematically discussed in [40] for all $G^\vee$; the prototype $T(\mathrm{SU}(2))$ is a basic example of three-dimensional mirror symmetry [60]. $T(G^\vee)$ is a rather subtle theory which, for example, can be interpreted as the universal kernel of geometric Langlands duality, as briefly explained in Section 3.5 of [107].) The theory $T(G^\vee)$ has OSp(4|4) superconformal symmetry; it has an action of $G^\vee$ on its Higgs branch and G on its Coulomb branch.[42] We couple $T(G^\vee)$ to $G^\vee$

[42] It is believed that the groups that act faithfully are the adjoint forms of $G^\vee$ and G, so the distinction between them is unimportant in the simply-laced case. The mirror of $T(G^\vee)$ is $T(G)$. In parallel with the Fayet–Iliopoulos parameters that are introduced momentarily, there is a mirror triple of mass parameters that violate the $G^\vee$-symmetry; these are not relevant in the present context as the $G^\vee$-symmetry is gauged.

gauge theory using the $G^\vee$ action on the Higgs branch. $T(G^\vee)$ can be deformed by Fayet–Iliopoulos parameters $(\alpha^\vee, \beta^\vee, \gamma^\vee)$; this breaks the G-symmetry to the maximal torus, eliminates the Coulomb branch, and makes the Higgs branch smooth. Once the Higgs branch is smooth, the theory is infrared free and one can aim for a classical description of the defect operator associated to coupling to $T(G^\vee)$. This classical description involves the singularity postulated in (6.9). The steps involved in reducing from a description involving a coupling to a field theory on the defect to a description involving the singularity are similar to what they are in one dimension less; see Section 3 of [56].

Describing the defect operator by coupling the bulk gauge theory to $T(G^\vee)$ has the advantage of making sense when the triple $(\alpha^\vee, \beta^\vee, \gamma^\vee)$ is nonregular. Let us consider the extreme case that these parameters vanish. Then the theory admits an action of G, acting only on fields supported along the defect. The conserved quantities $\mathbb{m}^\vee$ generate the action of the maximal torus of G, in the sense that the group element corresponding to $\eta^\vee \in T$ is $\exp(2\pi i(\eta^\vee, \mathbb{m}^\vee))$.

Naively speaking, it appears that, upon setting $\alpha^\vee$, $\beta^\vee$, and $\gamma^\vee$ to zero, since the theory has a G action, the cohomology of Q would also admit such an action and the trace (6.20) would then be a trace in a G-module. This would have strong implications for the knot invariants – probably too strong. An instructive problem arises here. Precisely when the triple $(\alpha^\vee, \beta^\vee, \gamma^\vee)$ is nonregular, the theory $T(G^\vee)$ flows to a non-trivial CFT in the infrared. The noncompactness of the initial value surface $K \times \mathbb{R}_+$ then becomes essential and it is likely that the continuous spectrum cannot be ignored. Even in the nonregular case, it is possible to express the partition function $Z(q)$ as a trace analogous to (6.20) in a much bigger Hilbert space – the space of all physical states of the $(0, 2)$ model, without reducing to the cohomology of Q. But it may not be possible to reduce to a discrete spectrum of BPS states with G action. For example, trying to do so would entail setting $|q| = 1$ in the expansions made for the unknot in Section 6.5.

6.4.3. Lifting to six dimensions. The last step of this type is the lift to an ultraviolet-complete description in six dimensions, along the lines of Section 5. The six-dimensional geometry is now $M_4 \times D$, where D is a two-manifold with U(1)-symmetry.

The six-dimensional theory is classified by the choice of a simply-laced Dynkin diagram, and the distinction between G and $G^\vee$ arises from a subtle choice mentioned in footnote 27. (To relate the six-dimensional theory to gauge theory of a Lie group that is not simply-laced, one makes one of the two constructions described in Section 5.5.2.) Since the six-dimensional theory is not infrared-free, it is not clear that a system consisting of the six-dimensional theory with a codimension two defect can be obtained by coupling the six-dimensional theory to a four-dimensional theory that is defined independently. However, the combined system consisting of the six-dimensional theory with a four-dimensional defect does exist. In fact, there are a family of half-BPS codimension two defects; see [36], [37], and [9]. They parallel

the corresponding half-BPS monodromy defects described in gauge theory in [55] and associated to Levi subgroups of G. We will consider here only the "full" defect which in reduction to gauge theory corresponds to a monodromy defect operator with the full set of parameters $(\alpha, \beta, \gamma, \eta)$.

The six-dimensional theory does not have a Lie group or gauge group of symmetries, but in the presence of a codimension two defect, it does have a global symmetry group, which is a form of G. The full defect corresponds after reduction on a circle to the monodromy defect in five-dimensional gauge theory that we have derived from (6.4). In six dimensions, the full defect is characterized only by the parameters $\beta^\vee$ and $\gamma^\vee$. (One may as well call these parameters β and γ, as the six-dimensional description is symmetrical between G and $G^\vee$.) $\alpha^\vee$ arises if, in compactifying on a circle to get to five dimensions, one twists by the element $\exp(-2\pi\alpha^\vee)$ of the global symmetry group.[43] As we have already discussed, $\eta^\vee$ is not present as a parameter in five dimensions; instead the five-dimensional theory has a conserved current with $\mathbb{m}^\vee$ as the conserved charge.

It is clear what to do with a codimension two defect in the context of the present paper. We place such a defect on $C \times D \subset M_4 \times D$, where $C \subset M_4$ is an oriented two-manifold. Upon reducing on the U(1) orbits on D, we return to the five-dimensional construction that we have already analyzed. To study a knot, we make the usual specialization to $M_4 = \mathbb{R} \times W_3$, $C = \mathbb{R} \times K$.

6.5. Using the duality. In the part of this paper that was based on representing knots by Wilson operators, there were a few technical problems in actually using the duality to learn about Chern–Simons theory for knots in a three-manifold W_3. One problem is that if W_3 is compact, then gauge theory on $W_3 \times \mathbb{R}_+$ with a reducible flat connection at infinity leads to infrared divergences. Their role in the duality is not yet understood. Another problem is that in defining a boundary condition at $y = \infty$, we have to pick a homomorphism $\rho\colon \pi_1(W_3) \to G_{\mathbb{C}}$; we do not know how this is related to the homomorphism $\rho^\vee\colon \pi_1(W_3) \to G_{\mathbb{C}}^\vee$ that one introduces in the dual description. Happily, in an important situation – knots in $\mathbb{R}^3$ with only gauge transformations that are trivial at infinity allowed – these issues do not arise.

For the equivalent story with monodromy defects, we are not so fortunate. We can still avoid infrared divergences by taking $W_3 = \mathbb{R}^3$. But now to study a knot K, we have to consider homomorphisms from the fundamental group of $\mathbb{R}^3 \backslash K$ to $G_{\mathbb{C}}$ or $G_{\mathbb{C}}^\vee$, with a prescribed monodromy around K. Because of the prescribed monodromy, there is no longer a trivial flat connection, and once one only allows gauge transformations that are trivial at infinity, any non-trivial flat connection becomes non-isolated. So to proceed, we need to learn something about the relation between the Chern–Simons path integral and that of $\mathcal{N} = 4$ super Yang–Mills for the case that the flat connection at infinity is not isolated. Also, for generic K, there are multiple homomorphisms of

[43] The form of G that acts as a global symmetry group in six dimensions has not been fully analyzed and may depend on a choice as in footnote 27. It appears that after reducing on a circle, the global symmetry group coincides with the gauge group.

$\pi_1(\mathbb{R}^3\backslash K)$ to $G_{\mathbb{C}}$ or $G^\vee_{\mathbb{C}}$, even when the conjugacy class of the monodromy around K is prescribed. So we cannot avoid the question of the relation under duality of the homomorphisms ρ and $\rho^\vee$.

In short, to actually use the duality based on monodromy defects, we need to learn more. And so far we have only mentioned questions of principle. In practice, for either the duality based on Wilson operators or that based on monodromy defects, to learn a lot one will need to know more about actually solving the equations.

Rather than say nothing at all, we will make a few remarks about the unknot $K_0 \subset \mathbb{R}^3$. The fundamental group of $\mathbb{R}^3\backslash K_0$ is simply the abelian group $\mathbb{Z}$, so it has up to conjugacy only one homomorphism to G or $G^\vee$ with prescribed monodromy, and the image of this homomorphism is abelian. So there is essentially only one possible integration cycle in Chern–Simons theory, and the standard integration cycle must coincide with the one we get in the $G^\vee$ description using the unique possible flat connection at infinity. The Chern–Simons action of an abelian flat connection vanishes (with the canonical framing), so we do not need to worry about a factor in the duality involving the classical action. There might be a correction to the formula involving the fact that the abelian flat connection is not isolated (in the context of $\mathbb{R}^3\backslash K_0$), or a constant $\mathfrak{N}_0$, as in (2.66), but we will just proceed and see what happens.

For simplicity, we consider the case of $G = \mathrm{SU}(2)$. The path integral for a Wilson operator in the spin j representation placed on the unknot in $\mathbb{R}^3$ is

$$J(q; K_0, j) = \frac{q^{(2j+1)/2} - q^{-(2j+1)/2}}{q^{1/2} - q^{-1/2}}. \tag{6.21}$$

We would like to express this function in the form of (6.19), which for $G = \mathrm{SU}(2)$ should become

$$J(q; K_0, j) = \sum_{r,c} w_{r,c} q^{r-cj}. \tag{6.22}$$

What sort of expansion will this be? Actually, there are two expansions that we should make. In general, in the $G^\vee$ description, we have $\alpha^\vee = 0$, and in the present case, we are relying on an abelian homomorphism $\rho^\vee$, so also $\beta^\vee = 0$. Hence if $\gamma^\vee = 0$, then we are in the nonregular case described at the end of Section 6.4.2, where the space of BPS states may not be well defined. So we prefer to take $\gamma^\vee \neq 0$. In this case, as explained at the end of Section 6.2.2, q does not have modulus 1, so there are two cases, $|q| < 1$ or $|q| > 1$. In these two cases, we will interpret (6.22) as a Laurent series around $q = 0$ or $q = \infty$, respectively.

There are simple expansions of this type which moreover are consistent with the fact that in (6.22) the coefficients $w_{r,c}$ are supposed to be independent of j. We use either

$$\frac{1}{q^{1/2} - q^{-1/2}} = -q^{1/2} \sum_{t=0}^{\infty} q^t, \quad |q| < 1, \tag{6.23}$$

or

$$\frac{1}{q^{1/2}-q^{-1/2}} = q^{-1/2}\sum_{t=0}^{\infty} q^{-t}, \quad |q|>1. \tag{6.24}$$

For example, the first leads to the formula

$$J(q;K_0,j) = \left(-q^{j+1}+q^{-j}\right)\sum_{t=0}^{\infty} q^t, \tag{6.25}$$

in which the finite Laurent polynomial J is written as the difference of two infinite Laurent series. This expansion takes the form (6.22); the coefficients $w_{r,c}$ are nonzero if and only if $c = \pm 1$ and r is a positive integer, or $r = 0$ with $c = 1$. A similar formula can be written straightforwardly for $|q| > 1$. Of course, to be satisfied with the expansion (6.25) or its cousin for $|q| > 1$, one would like to know that solutions with the claimed topological invariants actually exist. In the present context, it is unclear why there are solutions leading to the geometric series in (6.25). Possibly a hint comes from recent approaches to related problems such as [27].

The fact that one has to make two different expansions may be special to a reducible flat connection. In the case of an irreducible flat connection, one is free to take $\beta^\vee \neq 0$, and this means that $\gamma^\vee$ can be varied in an arbitrary way while avoiding nonregular triples. This suggests that the contribution to the path integral of an irreducible flat $G^\vee$ connection with monodromy around K will be given by a Laurent polynomial (powers of q bounded above and below) rather than a Laurent series (powers of q bounded in only one direction). At any rate, there is plenty to understand.

References

[1] M. Aganagic, H. Ooguri, N. Saulina, and C. Vafa, Black holes, q-deformed $2d$ Yang–Mills, and nonperturbative topological strings. *Nucl. Phys. B* **715** (2005), 304–348. MR 2135642 Zbl 1207.81147

[2] M. Aganagic and M. Yamazaki, Open BPS wall crossing and M-theory. *Nucl. Phys. B* **834** (2010), 258–272. MR 2610993 Zbl 1204.81132

[3] P. C. Argyres, A. Kapustin and N. Seiberg, On S-duality for non-simply-laced gauge groups, *J. High Energy Phys.* **0606** (2006), 043. MR 2233812 iopscience.iop.org/1126-6708/2006/06/043/

[4] A. Ashtekar and R. Tate, *Lectures on nonperturbative canonical gravity.* World Scientific, River Edge, NJ, 1991. MR 1157631 Zbl 0948.83500

[5] P. S. Aspinwall and M. Gross, The SO(32) heterotic string on a K3 surface. *Phys. Lett. B* **387** (1996), 735–742. MR 1416960

[6] M. F. Atiyah, On framings Of 3-manifolds. *Topology* **29** (1990) 1–7. MR 1046621 Zbl 0716.57011

[7] D. Bar-Natan, On Khovanov's categorification of the Jones polynomial. *Alg. Geom. Topology* **2** (2002), 337–370. MR 1917056 Zbl 0998.57016 www.msp.warwick.ac.uk/agt/2002/02/p016.xhtml

[8] C. Beasley, Localization for Wilson loops in Chern–Simons theory. Preprint 2009. arXiv:0911.2687

[9] F. Benini, Y. Tachikawa, and D. Xie, Mirrors of $3d$ Sicilian theories. *J. High Energy Phys.* **9** (2010), 1–32. MR 2776954

[10] D. Berenstein, R. Corrado, W. Fischler, and J. M. Maldacena, Operator product expansion for Wilson loops and surfaces in the large N limit. *Phys. Rev. B* **59** (1999), 105023. MR 1709200

[11] J. Bernstein, I. Frenkel, and M. Khovanov, A categorification of the Temperley–Lieb algebra and Schur quotients of $U(\mathfrak{sl}_2)$ via projective and Zuckerman functors. *Selecta. Math. New. Ser.* **5** (1999), 199–241. MR 1714141 Zbl 0981.17001

[12] M. Bershadsky, V. Sadov, and C. Vafa, D-branes and topological field theories. *Nucl. Phys. B* **463** (1996), 420–434. MR 1393648 Zbl 1004.81560

[13] J. Birman and H. Wenzel, Braids, link polynomials and a new algebra. *Trans. Am. Math. Soc.* **313** (1989), 249–273. MR 992598 Zbl 0684.57004 www.ams.org/journals/tran/1989-313-01/S0002-9947-1989-0992598-X/home.html

[14] V. Borokhov, A. Kapustin, X. Wu, Topological disorder operators in three-dimensional conformal field theory. *J. High Energy Phys.* **0211** (2002), 049. MR 1955430

[15] R. L. Bryant and S. M. Salamon, On the construction of some complete metrics with exceptional holonomy. *Duke Math. J.* **58** (1989), 829–850. MR 1016448 Zbl 0681.53021

[16] C. G. Callan, Jr. and J. Maldacena, Brane Dynamics From the Born–Infeld Action. *Nucl. Phys. B* **513** (1998), 198–212. MR 1016448 Zbl 0958.81105

[17] J. M. Camino, A. Paredes, and A. V. Ramallo, Stable wrapped branes. *J. High Energy Phys.* **0105** (2001), 011. MR 81T30 iopscience.iop.org/1126-6708/2001/05/011/

[18] S. Cautis and J. Kamnitzer, Knot homology via derived categories of coherent sheaves I. The $\mathfrak{sl}(2)$ case. *Duke Math J.* **142** (2008), 511–588. MR 2411561 Zbl 1145.14016

[19] S. Cecotti, A. Neitzke, and C. Vafa, R-Twisting and $4d/2d$ correspondence. Preprint 2010. arXiv:1006.3435

[20] B. Chen, W. He, J.-B. Wu, and L. Zhang, M5-branes and Wilson surfaces. *J. High Energy Phys.* **0708** (2007), 067. MR 81T30 iopscience.iop.org/1126-6708/2007/08/067/

[21] N. R. Constable, R. C. Myers and O. Tafjord, The noncommutative bion core. *Phys. Rev. D* **61** (2000), 106009. MR 1790784

[22] K. Corlette, Flat G-bundles with canonical metrics. *J. Diff. Geom.* **28** (1988), 361–382. MR 965220 Zbl 0676.58007

[23] L. Crane and I. B. Frenkel, Four-dimensional topological quantum field theory, Hopf categories, and the canonical bases. *J. Math. Phys.* **35** (1994), 5136–5154. MR 1295461 Zbl 0892.57014

[24] S. Deser, R. Jackiw, and S. Templeton, Topologically massive gauge theory. *Ann. Phys.* **140** (1982), 372–411. Reprint *Ann. Phys.* **281** 2000), 409–449. MR 665601

[25] D. E. Diaconescu, D-Branes, monopoles and Nahm equations. *Nucl. Phys. B* **503** (1997) 220–238. MR 58D27 Zbl 0938.81034

[26] R. Dijkgraaf, C. Vafa, and E. Verlinde, M-Theory and a topological string duality. Preprint 2006. arXiv:hep-th/0602087

[27] T. Dimofte, S. Gukov, and L. Hollands, Vortex counting and Lagrangian 3-manifolds. Preprint 2010. arXiv:1006.0977

[28] E. D'Hoker, J. Estes, M. Gutperle, and D. Krym, Exact half-BPS flux solutions in M-theory II: Exact solutions asymptotic to $\mathrm{AdS}_7 \times S^4$. *J. High Energy Phys.* **0812** (2008), 044. MR 2469896 iopscience.iop.org/1126-6708/2008/12/044/

[29] N. Drukker, D. Gaiotto, and J. Gomis, The virtue of defects in 4D gauge theories and 2D CFT's. Preprint 2010. arXiv:1003.1112

[30] N. M. Dunfield, S. Gukov, and J. Rasmussen, The superpolynomial for knot homologies. *Exp. Math.* **15** (2006), 129-159. MR 2253002 Zbl 1118.57012

[31] A. Floer, Morse theory for Lagrangian intersections, *J. Differ. Geom.* **28** (1988), 513–547. MR 965228 Zbl 0674.57027

[32] D. S. Freed and R. E. Gompf, Computer calculation of Witten's 3-manifold invariant. *Commun. Math. Phys.* **141** (1991), 79–117. MR 1133261 Zbl 0739.53065

[33] E. Frenkel, Lectures on the Langlands program and conformal field theory. In P. Cartier et al. (eds), *Frontiers in number theory, physics, and geometry* II. *On conformal field theories, discrete groups and renormalization. Papers from the meeting, Les Houches, France, March 9–21, 2003.* Springer Verlag, Berlin, 2007, 389–533. Zbl 1196.11091 MR 2290768

[34] I. B. Frenkel and M. Khovanov, Canonical bases in tensor products and graphical calculus for $U_q(\mathfrak{sl}_2)$. *Duke Math. J.* **87** (1997), 409–480. MR 1446615 Zbl 0883.17013

[35] P. Freyd, D. Yetter, J. Hoste, W. B. R. Lickorish, K. Millett, and A. Ocneanu, A new polynomial invariant of knots and links. *Bull. Am. Math. Soc., New Ser.* **12** (1985), 239–246. MR 776477 Zbl 0572.57002 www.ams.org/journals/bull/1985-12-02/S0273-0979-1985-15361-3/home.html

[36] D. Gaiotto, $\mathcal{N} = 2$ dualities. Preprint 2009. arXiv:0904.2715

[37] D. Gaiotto and J. Maldacena, The gravity duals of $\mathcal{N} = 2$ superconformal field theories. Preprint 2009. arXiv:0904.4466

[38] D. Gaiotto, G. W. Moore, and A. Neitzke, Framed BPS states. Preprint 2010. arXiv:1006.0146

[39] D. Gaiotto and E. Witten, Supersymmetric boundary conditions in $\mathcal{N} = 4$ super Yang–Mills theory. *J. Stat. Phys.* **135** (2009), 789–855. MR 2548595 Zbl 1178.81180

[40] D. Gaiotto and E. Witten, Janus configurations, Chern–Simons couplings, and The theta-angle in $\mathcal{N} = 4$ Super Yang–Mills Theory. *J. High Energy Phys.* **1006** 2010, 097. MR 2680316

[41] D. Gaiotto and E. Witten, Knot invariants from four-dimensional gauge theory. Preprint 2011. arXiv:1106.4789

[42] D. Gaitsgory, Twisted Whittaker model and factorizable sheaves. *Sel. Math. New Ser.* **13** (2008), 617–659. MR 2403306 Zbl 1160.17009

[43] G. W. Gibbons, D. N. Page, and C. N. Pope, Einstein metrics on S^3, $\mathbb{R}^3$ and $\mathbb{R}^4$ bundles. *Commun. Math. Phys.* **127** (1990), 529–553. MR 1040893 Zbl 0699.53053

[44] J. Gomis, S. Matsuura, T. Okuda, and D. Trancanelli, Wilson loop correlators at strong coupling: from matrices to bubbling geometries. *J. High Energy Phys.* **0808** (2008), 068. MR 2434518 iopscience.iop.org/1126-6708/2008/08/068/

[45] J. Gomis and F. Passerini, Holographic Wilson loops. *J. High Energy Phys.* **0608** (2006), 074. MR 2249907 iopscience.iop.org/1126-6708/2006/08/074/

[46] R. Gopakumar and C. Vafa, On the gauge theory/geometry correspondence. *Adv. Theor. Math. Phys.* **3** (1999), 1415–1443. MR 1796682 Zbl 0972.81135

[47] R. Gopakumar and C. Vafa, M-Theory and topological strings I, II. Preprint 1998. arXiv:hep-th/9809187 arXiv:hep-th/9812127

[48] R. Graham and E. Witten, Conformal anomaly of submanifold observables in AdS/CFT correspondence. *Nucl. Phys. B* **546** (1999), 52–64. MR 1682674 Zbl 0944.81046

[49] R. Gregory, J. A. Harvey, and G. W. Moore, Unwinding strings and T-duality of Kaluza–Klein and H-Monopoles. *Adv. Theor. Math. Phys.* **1** (1997), 283–297. MR 1605632 Zbl 0901.53073

[50] S. Gukov, Three-dimensional quantum gravity, Chern–Simons theory, and the A-polynomial. *Commun. Math. Phys.* **255** (2005), 577–627. MR 2134725 Zbl 1115.57009

[51] S. Gukov, Surface operators and knot homologies. Preprint 2007. arXiv:0706.2369

[52] S. Gukov, A. Iqbal, C. Kozcaz, and C. Vafa, Link homologies and the refined topological vertex. *Comm. Math. Phys.* **298** (2010), 757–785. MR 2670927 Zbl 1207.81123

[53] S. Gukov, A. S. Schwarz, and C. Vafa, Khovanov–Rozansky homology and topological strings. *Lett. Math. Phys.* **74** (2005), 53–74. MR 2193547 Zbl 1105.57011

[54] S. Gukov and J. Walcher, Matrix factorizations and Kauffman homology. Preprint 2005. arXiv:hep-th/0512298

[55] S. Gukov and E. Witten, Gauge theory, ramification, and the geometric Langlands program. In J. David et al. (eds), *Current developments in mathematics, 2006*. International Press, Somerville (MA), 2008, 35–180. Zbl 05504307 MR 2459305

[56] S. Gukov and E. Witten, Rigid surface operators. *Adv. Theor. Math. Phys.* **14** (2010), 87–178. MR 2684979 Zbl 1203.81114

[57] A. Haydys, Fukaya–Seidel category and gauge theory. Preprint 2010. arXiv:1010.2353

[58] M. Henningson, Boundary conditions for GL-twisted $N = 4$ SYM. Preprint 2011. arXiv:1106.3845

[59] M. Henningson and N. Wyllard, Zero-energy states of $\mathcal{N} = 4$ SYM on T^3: S-duality and the mapping class group. *J. High Energy Phys.* **0804** (2008), 066. MR 2425237 iopscience.iop.org/1126-6708/2008/04/066/

[60] K. Intriligator and N. Seiberg, Mirror symmetry in three dimensional gauge theories. *Phys. Lett. B* **387** (1996), 513–519. MR 1413696

[61] A. Iqbal, C. Kozcaz, and C. Vafa, The refined topological vertex. *J. High Energy Phys.* **0910** (2009), 069. MR 2607441 iopscience.iop.org/1126-6708/2009/10/069/

[62] V. F. R. Jones, A polynomial invariant for links via von Neumann algebras. *Bull. Am. Math. Soc., New Ser.* **12** (1985), 103–111. MR 766964 Zbl 0564.57006 www.ams.org/journals/bull/1985-12-01/S0273-0979-1985-15304-2/home.html

[63] V. F. R. Jones, Hecke algebra representations of braid groups and link polynomials. *Ann. Math.* (2) **126** (1987), 335–388. MR 908150 Zbl 0631.57005

[64] J. Kamnitzer, The Beilinson–Drinfeld grassmannian and symplectic knot homology. In D. A. Ellwood (ed), *Grassmannians, moduli spaces and vector bundles.* Clay Math. Institute, Cambridge (MA), and Amer. Math. Soc., Providence (RI), 2011, 81–94. MR 2807850

[65] A. Kapustin, Wilson 't Hooft operators in four-dimensional gauge theories and S-duality. *Phys. Rev. D* **74** (2006), 025005. MR 2249977

[66] A. Kapustin and M. J. Strassler, On mirror symmetry in three-dimensional abelian gauge theories. *J. High Energy Phys.* **9904** (1999), 021. MR 1710589 Zbl 0953.81097 iopscience.iop.org/1126-6708/1999/04/021/

[67] A. Kapustin and E. Witten, Electric-magnetic duality and the geometric Langlands program. *Commun. Number Theory Phys.* **1** (2007), 1–236. MR 2306566 Zbl 1128.22013 www.intlpress.com/CNTP/p/2007/CNTP-1-1_1-236.pdf

[68] L. Kauffman, State models and the Jones polynomial. *Topology* **26** (1987), 395–407. MR 899057 Zbl 0622.57004

[69] M. Khovanov, A categorification of the Jones polynomial. *Duke. Math. J.* **101** (2000), 359–426. MR 1740682 Zbl 0960.57005

[70] M. Khovanov and L. Rozansky, Matrix factorizations and link homology. *Fundam. Math.* **199** (2008), 1–91. MR 2391017 Zbl 1145.57009 journals.impan.pl/cgi-bin/doi?fm199-1-1

[71] B. Kostant, The principal three-dimensional subgroup and the Betti numbers of a complex simple Lie group. *Am. J. Math.* **81** (1959), 973–1032. MR 0114875 Zbl 0099.25603

[72] D. Krefl, S. Pasquetti, and J. Walcher, The real topological vertex at work. *Nucl. Phys. B* **833** (2010), 153–198. MR 2611000 Zbl 1204.81145

[73] P. B. Kronheimer and T. S. Mrowka, Knot homology groups from instantons. Preprint 2008. arXiv:0806.1053

[74] P. B. Kronheimer and T. S. Mrowka, Khovanov homology is an unknot-detector. Preprint 2010. arXiv:1005.4346

[75] J. M. F. Labastida, M. Mariño, and C. Vafa, Knots, links and branes at large *N. J. High Energy Phys.* **0011** (2000), 007. MR 180659 Zbl 0990.81545 iopscience.iop.org/1126-6708/2000/11/007/

[76] J. M. F. Labastida and M. Mariño, Polynomial invariants for torus knots and topological strings. *Commun. Math. Phys.* **217** (2001), 423–449. MR 1821231 Zbl 1018.81049

[77] O. Lunin, 1/2 BPS States In M Theory And Defects In The Dual CFTs. *J. High Energy Phys.* **0710** (2007), 014. MR 2357964 iopscience.iop.org/1126-6708/2007/10/014/

[78] M. Mariño, *Chern–Simons theory, matrix models and topological strings.* Clarendon Press, Oxford, 2005. MR 2177747 Zbl 1093.81002

[79] M. Mariño, String theory and the Kauffman polynomial. *Commun. Math. Phys.* **298** (2010), 613–643. MR 2670922 Zbl 1207.81129

[80] H. Murakami, An introduction to the volume conjecture and its generalizations. *Acta Math. Vietnam* **33** (2008), 209–253. MR 2501844 Zbl 1179.57023

[81] R. C. Myers, Dielectric-branes. *J. High Energy Phys.* **9912** (1999), 022. MR 1743060 Zbl 0958.81091 iopscience.iop.org/1126-6708/1999/12/022/

[82] H. Ooguri and C. Vafa, Knot invariants and topological strings, *Nucl. Phys. B* **577** (2000), 419–438. MR 1765411 Zbl 1036.81515

[83] D. Panyushev, On the Dynkin index of a principal $\mathfrak{sl}_2$ subalgebra. *Adv. Math.* **221** (2009) 1115–1121. MR 2518633 Zbl 1163.17013

[84] A. M. Polyakov, Fermi–Bose transmutations induced by gauge fields. *Mod. Phys. Lett. A* **3** (1988), 325–328. MR 927055

[85] J. H. Przytycki and P. Traczyk, Invariants of links of Conway type. *Kobe J. Math.* **4** (1987), 115-139. MR 945888 Zbl 0655.57002

[86] X. Qi and S.-C. Zhang, Topological insulators and superconductors. Preprint 2010. arXiv:1008.2026

[87] P. Ramadevi and T. Sarkar, On link invariants and topological string amplitudes. *Nucl. Phys. B* **600** (2001), 487–511. MR 1833409 Zbl 1097.81742

[88] J. F. Schonfeld, A mass term for three-dimensional gauge fields. *Nucl. Phys. B* **185** (1981), 157–171.

[89] A. S. Schwarz, The partition function of degenerate quadratic functional and Ray-Singer invariants. *Lett. Math. Phys.* **2** (1978), 247–252. MR 0676337 Zbl 0383.70017

[90] P. Seidel and I. Smith, A link invariant from the symplectic geometry of nilpotent slices. *Duke math J.* **134** (2006), 453–514. MR 2254624 Zbl 1108.57011

[91] C. Simpson, Harmonic bundles on noncompact curves. *J. Am. Math. Soc.* **3** (1990), 713–770. MR 1040197 Zbl 0713.58012 www.ams.org/journals/jams/1990-03-03/S0894-0347-1990-1040197-8/home.html

[92] S. Sinha, C. Vafa, SO and Sp Chern–Simons at large N. Preprint 2000. arXiv:hep-th/0012136

[93] A. Strominger, Open p-branes. *Phys. Lett. B* **383** (1996), 44–47. MR 1402864 Zbl 0903.53053

[94] Y. Tachikawa, $\mathcal{N} = 2$ S-duality via outer-automorphism twists. *J. Phys. A, Math. Theor.* **44** (2011), 182001. MR 2788718 Zbl 1214.81169

[95] P. K. Townsend, The eleven-dimensional supermembrane revisited. *Phys. Lett. B* **350** (1995), 184–188. MR 1331064

[96] A. Tsuchiya and Y. Kanie, Vertex operators in the conformal field theory on $\mathbb{P}^1$ and monodromy representations of the braid group. *Lett. Math. Phys.* **13** (1987), 303–312. Extended version in M. Jimbo, T. Miwa, A. Tsuchiya (eds), *Conformal field theory and solvable lattice models* (*Kyoto, 1986*). Academic Press, Boston (MA), 1988, 297–372. MR 895293 Zbl 0631.17010

[97] V. G. Turaev, The Yang–Baxter equation and invariants of links. *Inv. Math.* **92** (1988), 527–553. MR 939474 Zbl 0648.57003

[98] C. Vafa, Geometric origin of Montonen–Olive duality. *Adv. Theor. Math. Phys.* **1** (1997), 158–166. MR 1489300 Zbl 0889.53054

[99] C. Vafa and E. Witten, A strong coupling test of S-duality. *Nucl. Phys. B* **431** (1994), 3–77. MR 1305096 Zbl 0964.81522

[100] E. Witten, Supersymmetry and Morse theory. *J. Differ. Geom.* **17** (1982), 661–692. MR 683171 Zbl 0499.53056

[101] E. Witten, Topological quantum field theory. *Commun. Math. Phys.* **117** (1988), 353–386. MR 953828 Zbl 0656.53078

[102] E. Witten, Quantum field theory and the Jones polynomial. *Commun. Math. Phys.* **121** (1989), 351-399. MR 990772 Zbl 0726.57010

[103] E. Witten, Chern–Simons gauge theory as a string theory. In H. Hofer et al. (eds.), *The Floer memorial volume.* Birkhäuser, Basel, 1995, 637-678. MR 1362846 Zbl 0844.58018

[104] E. Witten, Some comments on string dynamics. In Bars et al. (eds.), *Strings 95: future perspectives in string theory, USC, Los Angeles, March 13–18, 1995.* World Scientific, Singapore, 1996, 501–523. MR 1660736

[105] E. Witten, Dynamics of quantum field theory. In P. Deligne et. al. (eds.), *Quantum fields and strings: a course for mathematicians.* Amer. Math. Soc., Providence (RI), 1999, Lecture 7.7, 1119–1424. MR 1701615

[106] E. Witten, Conformal field theory in four and six dimensions. In U. Tillman (ed.), *Topology, geometry and quantum field theory. Proceedings of the 2002 Oxford symposium in honour of the 60th birthday of Graeme Segal, Oxford, U.K., June 24–29, 2002.* Cambridge University Press, Cambridge (U.K.), 2004, 405–419. MR 2079382 Zbl 1101.81096

[107] E. Witten, Geometric Langlands from six dimensions. In P. R. Kotiuga (ed.), *A celebration of the mathematical legacy of Raoul Bott. Based on the conference, CRM, Montreal, Canada, June 9–13, 2008.* Amer. Math. Soc., Providence (RI), 2010, 281–310. MR 2648898 Zbl 1216.81129

[108] E. Witten, Analytic continuation Of Chern–Simons theory. In J. E. Andersen et al. (eds), *Chern-Simons gauge theory: 20 years after.* Amer. Math. Soc., Providence (RI), and International Press, Somerville (MA), 2011, 347–446. MR 2809462

[109] E. Witten, A new look at the path integral of quantum mechanics. Preprint 2010 arXiv:1009.6032

[110] S. Yamaguchi, Bubbling geometries for half-BPS Wilson lines. *Int. J. Mod. Phys. A* **22** (2007), 1353–1374. MR 2309814 Zbl 1125.81039

[111] J. P. Yamron, Topological actions from twisted supersymmetric theories. *Phys. Lett. B* **213** (1988), 325–330. MR 965719

[112] K. Zarembo, Supersymmetric Wilson loops. *Nucl. Phys. B* **643** (2002), 157–171. MR 1936917

Received January 19, 2011

Edward Witten, Natural Sciences, Institute for Advanced Study, Einstein Drive, Princeton, NJ 08540, and Department of Physics, Stanford University, Palo Alto, CA 94305, U.S.A.
E-mail: witten@ias.edu