Available online at www.sciencedirect.com
SCIENCE DIRECT®
Annals of Physics 310 (2004) 428–492
ANNALS of PHYSICS
www.elsevier.com/locate/aop

A class of P, T-invariant topological phases of interacting electrons

Michael Freedman,[a] Chetan Nayak,[a,b,*] Kirill Shtengel,[a] Kevin Walker,[a] and Zhenghan Wang[a,c]

[a] *Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA*
[b] *Department of Physics and Astronomy, University of California, Los Angeles, CA 90095-1547, USA*
[c] *Department of Mathematics, Indiana University, Bloomington, IN 47405, USA*

Received 4 December 2003

Abstract

We describe a class of parity- and time-reversal-invariant topological states of matter which can arise in correlated electron systems in 2+1-dimensions. These states are characterized by particle-like excitations exhibiting exotic braiding statistics. P and T invariance are maintained by a 'doubling' of the low-energy degrees of freedom which occurs naturally without doubling the underlying microscopic degrees of freedom. The simplest examples have been the subject of considerable interest as proposed mechanisms for high-T_c superconductivity. One is the 'doubled' version of the chiral spin liquid. The chiral spin liquid gives rise to anyon superconductivity at finite doping and the corresponding field theory is $U(1)$ Chern–Simons theory at coupling constant $m = 2$. The 'doubled' theory is two copies of this theory, one with $m = 2$ the other with $m = -2$. The second example corresponds to Z_2 gauge theory, which describes a scenario for spin-charge separation. Our main concern, with an eye towards applications to quantum computation, are richer models which support non-Abelian statistics. All of these models, richer or poorer, lie in a tightly organized discrete family indexed by the Baraha numbers, $2\cos(\pi/(k+2))$, for positive integer k. The physical inference is that a material manifesting the Z_2 gauge theory or a doubled chiral spin liquid might be easily altered to one capable of universal quantum computation. These phases of matter have a field-theoretic description in terms of gauge theories which, in their infrared limits, are topological field theories. We motivate these gauge theories using a parton model or slave-fermion construction and show how they can be solved exactly. The structure of the resulting Hilbert spaces can be understood in purely combinatorial terms. The highly constrained nature of this combinatorial construction,

* Corresponding author. Fax: 1-310-825-5734.
E-mail address: nayak@physics.ucla.edu (C. Nayak).

0003-4916/$ - see front matter
doi:10.1016/j.aop.2004.01.006

phrased in the language of the topology of curves on surfaces, lays the groundwork for a strategy for constructing microscopic lattice models which give rise to these phases.

PACS: 71.10.Hf; 71.10.Pm

1. Introduction

In quantum mechanics, trajectories in which identical particles are permuted must be considered on an equal footing with those in which they are not. Consequently, particles can be classified according to the irreducible representations of the permutation group. This leaves bosons and fermions as the only allowed possibilities [1]. However, in two spatial dimensions, trajectories can be further classified according to the braid group, essentially because the combination of two counter-clockwise exchanges in succession can not be adiabatically deformed to no exchange at all. As a result, non-trivial *braiding* statistics is possible in two dimensions [2].

The simplest examples of particles with exotic braiding statistics are called *anyons*. They realize one-dimensional representations of the braid group: the quantum-mechanical wavefunction acquires a non-trivial phase as a result of a counter-clockwise exchange of one anyon with another (the complex conjugate phase is associated with a clockwise rotation) or the 2π rotation of an anyon. It is actually not necessary for the anyons to be identical, unlike with the permutation group. For instance, a phase $e^{i\theta_{ii}}$ could result when a particle of type i is exchanged with another of type i while a phase $e^{2i\theta_{ij}}$ results when a particle of type i winds around a particle of type j and returns to its original position (when the particles are distinguishable, a trajectory in which they are exchanged leads to a different configuration). Thus, the term 'statistics' is somewhat misleading in the two-dimensional case because the classification is not according to the permutation group. It is more correct to say, instead, that there is a topological interaction at work: an interaction between particles which only depends on how they are braided. This is reflected in the field-theoretic implementation of anyonic statistics, in which a $U(1)$ gauge field mediates the interaction between the particles. If the gauge field is governed by the $U(1)$ Chern–Simons action, then the interaction is purely topological in nature. The coefficient of the Chern–Simons term determines the phase assigned to a particle exchange. The quasiparticle excitations of most observed fractional quantum Hall states are believed to be anyons [3–6].

An even more exotic variety of braiding statistics is associated with multi-dimensional representations of the braid group. If there are p states of the system when there are particles at $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, then the effect of an exchange of identical particles may be represented by a $p \times p$ matrix acting on the p states of the system. The different matrices corresponding to different possible exchanges need not commute; hence, this type of statistics is called *non-Abelian* (braiding) statistics. Chern–Simons gauge theories with non-Abelian groups generically give rise to particles with such braiding properties. The leading candidate to describe the $\nu = 5/2$ fractional

quantum Hall state [7,8] is a state with *non-Abelian* (braiding) statistics [9–13]. Other quantum Hall states observed in the first excited Landau level [8,14] might also be non-Abelian, possibly described by some of the states proposed in [15–17].

The different n particle trajectories which begin and end at the same positions, up to exchanges, are classified by the elements of the *braid group*. If we do not allow exchanges, as in the case of distinguishable particles, then we have the 'pure' braid group. Different varieties of particle braiding statistics correspond to different representations of the braid group. These representations are realized in the Hilbert spaces of many-particle systems which support excitations with non-trivial braiding statistics and also in the corresponding Chern–Simons field theories.

The Chern–Simons field theories which describe exotic braiding statistics are topological quantum field theories (TQFTs). The gauge fields in these theories do not probe local geometry—in particular, they disregard distance—so they only respond to topological properties, of the particle trajectories and also of the manifold on which they play out. Consider the amplitude associated with a process in which two pairs of statistics $e^{\pi i/m}$ anyons and their anti-particles are created out of the vacuum. If the anyons braid around each other before the two pairs annihilate, the amplitude acquires a phase $e^{2\pi i\ell/m}$, where ℓ is the linking number of the trajectories. The non-Abelian Chern–Simons theories are related to more interesting link invariants such as the Jones polynomial. Even though the topology of the physical realizations which are probed in experiments is usually trivial, it is useful to consider more complicated topologies as a *gedanken* experiment probe of the underlying structure of the quasiparticle braiding properties. For instance, a system of statistics $e^{\pi i/m}$ anyons has a ground state degeneracy of m^g on a surface of genus g. The physical interpretation of these states is that quasiparticles pick up a phase $e^{2\pi ni/m}$, $n = 0, \ldots, m-1$ upon encircling the longitude of the torus. The underlying electrons have periodic boundary conditions in all of these states. By taking linear combinations of these states (or, in the language of conformal field theory, by applying the modular S-matrix), we can construct the states in which the quasiparticles acquire these phases upon encircling the meridian.

Chern–Simons theories give a local description, but a redundant one. The basic idea is familiar in the context of electromagnetism: one could eschew the gauge field $A_\mu = (\varphi, \mathbf{A})$ in favor of the physically measurable gauge-invariant electric and magnetic fields $\mathbf{E}, \mathbf{B}$ but only at the cost of introducing non-local interactions between fields and charged particles in order to implement the Aharonov–Bohm effect.

Indeed, any gauge theory will give rise to non-trivial braiding, by the Aharonov–Bohm effect. However, gauge theories with a continuous gauge group which are governed by a Maxwell term alone are strongly fluctuating. In 2+1 dimensions, they fluctuate so strongly that they always confine charged (under the gauge group) particles. Thus, the fractional excitations, which would have exhibited non-trivial braiding statistics if set free, are not, in the final analysis, part of the particle spectrum of the theory. A Chern–Simons term precludes such wild fluctuations by enslaving fluxes to charges. The other exception occurs when the gauge group is discrete, as can occur on a lattice or in the continuum when a continuous group is spontaneously broken to a discrete subgroup. Discrete gauge fields have a deconfined phase, in

which charges and fluxes (which are restricted to a discrete allowed spectrum) interact through the Aharonov–Bohm effect.

This elegant field-theoretic description of exotic braiding statistics begs the question: are any of these possibilities actually realized in the real world? The answer is almost surely in the affirmative, at least for the case of anyons: the quasiparticle and quasihole excitations of fractional quantum Hall ground states are, according to theory, anyons. However, there has not, to date, been a direct observation of the phases associated with braiding. Beyond this, relatively little is known. In order to bridge this gap, it is important to understand in what other contexts one can observe *topological phases* [18,19] (or, equivalently, *fractionalized phases*) described by TQFTs. In particular, it is still an open question what types of microscopic Hamiltonians have, in their infrared limits, *topological phases*.

We can derive some insight into this question by understanding the structure of the Hilbert spaces of TQFTs. The Hilbert space of a TQFT in a topologically trivial geometry with no quasiparticles present is completely trivial: there is just a single state in the theory. However, when quasiparticles are introduced or the theory is put on a higher-genus surface, the Hilbert space will consist of a set of degenerate ground states. (In the former case, these are ground states for a fixed quasiparticle number, assuming that the quasiparticles are not allowed to move.) The functional integral formulation of the theory encompasses all of these possibilities: one must simply perform the functional integral over fields defined on different manifolds and with different boundary conditions. From a canonical point of view, however, the different possible topologies seem to give rise to completely different Hilbert spaces; they aren't even the same size, unlike in most systems, where changing the boundary conditions has little effect in the thermodynamic limit. Thus, it would seem that the canonical formalism is ill-suited for TQFTs.

This is not the case. In fact, all of the Hilbert spaces associated with different topologies can be brought under the aegis of a mathematical structure called a *modular functor* [20]. The modular functor can be enhanced by including edge states and an 'annulus category' which transforms these states. As we will discuss in this paper, this complex can be understood in more conventional physics terms as the representation theory of the commutator algebra of the fundamental gauge-invariant variables of the theory. The structure which is thus revealed gives, we believe, important clues about the necessary structure in any microscopic model which could give rise to a topological phase.

An important step in this direction is taken by the combinatorial construction of TQFTs. In this construction, one builds the Hilbert space of a TQFT in the following abstract fashion. Rather than start with the space of gauge fields A_μ, or some gauge-fixed version thereof, and complex-valued wavefunctionals on this configuration space, one begins with a set of 'pictures,' collections of non-intersecting curves on a given surface (more precisely, 1-manifolds). Two pictures are considered to be equivalent if they can be continuously deformed into each other. One then looks at the space of complex-valued functions on this set; it is an infinite-dimensional vector space. However, one can imagine imposing constraints on this vector space to reduce it to a finite-dimensional one. It turns out there is a limited number of ways of doing

this, in fact just a single infinite sequence. If the wrong constraints are chosen, the resulting vector space will be zero-dimensional. Those favored conditions which lead to finite-dimensional vector spaces can be solved combinatorially. The construction can be generalized to surfaces with boundaries and 'punctures,' at which curves can terminate. The latter are quasiparticles, and their statistics can be calculated by taking one puncture around another and using the constraints to simplify the resulting picture. Such a representation of the states of a system in terms of loops will be familiar to some readers from analyses of dimer models, whose 'transition graphs' are configurations of loops [21]. The TQFTs which we discuss are natural generalizations of the Z_2 gauge theory which emerges in the quantum dimer model on non-bipartite lattices [22,23].

From the dimensions of their Hilbert spaces, we can guess that these TQFTs are closely related to but are not quite the same as known Chern–Simons theories. Unlike the latter, they are P, T-invariant. This comes about, in most cases, in an almost trivial way: the combinatorial construction leads to two decoupled Chern–Simons theories which are identical except that their chiralities are opposite. This theory is called the 'doubled' theory. The Hilbert space of the doubled theory is the tensor product of the Hilbert space of one copy of the Chern–Simons theory with its conjugate. As a result, the dimensions of the Hilbert spaces are the squares of the dimensions of the P, T-violating Chern–Simons theories. (In some 'pathological' cases, the doubled theory is not simply the tensor square of the chiral theory but actually includes extra structure automatically repairing 'flaws' in the chiral theory.) The combinatorial approach automatically leads to the doubled theory: no artificial doubling of the degrees of freedom of the system was introduced at the outset; doubling emerged in the topological phase.

These Chern–Simons theories are associated with gauge group $SU(2)$. This is a consequence of the wonderful identification between $SU(2)$ representations and diagrams of non-intersecting curves (unoriented embedded 1-manifolds) in a disk [24] established in the Rumer–Teller–Weyl theorem. In short, $SU(2)$ emerges from something very commonplace: sets of loops, which could represent domain walls, dimers, etc. Z_2 or $U(1)$ gauge groups occur in special cases in which they happen to coincide with a level $k = 1$ $SU(2)$ theory, as we discuss further in Section 5.3. (The $SU(2)$ gauge group is completely unrelated to the $SU(2)$ spin-rotational symmetry which may or may not be preserved by a correlated electron or spin system which is in such a topological phase.)

The combinatorial construction seems so far-removed from the ordinary methods of quantum field theory and many-body physics that it is initially somewhat surprising that it leads to familiar TQFTs. However, a connection can be made by viewing these Hilbert spaces as representations of the commutator algebra of gauge-invariant Wilson loop operators rather than the less tangible gauge fields themselves. A simple representation of the form described above is only available for the doubled theory. The elucidation of this path to the combinatorial construction of the doubled theories is one of the main goals of this paper.

The wavefunctions of these topological field theories are related to correlation functions in *two-dimensional* field theories, as is familiar from Laughlin's plasma

analogy [4] for quantum Hall wavefunctions and from the relation between Chern–Simons theory and conformal field theory [9,25,26].

In this paper, we will explore the physics of topological phases observed with successively finer microscopes. First, we will discuss the long-wavelength effective field theories which encapsulate the phenomena of exotic braiding statistics: Chern–Simons gauge theories and discrete gauge theories. The former have appeared in the context of both the quantum Hall effect and also anyon superconductivity [27,28] The latter have recently been studied [29,30] in a revival of earlier ideas about RVB states [22] and spin-charge separation as a mechanism for superconductivity [21,31–34]. These two theories are the first in a sequence which forms the subject of this paper. The other member of this sequence are our main interest because they have excitations exhibiting non-Abelian braiding statistics. One of the principal motivations for this work is the construction of models for topological quantum computation [35–39]; non-Abelian statistics (in fact, sufficiently rich non-Abelian statistics) is necessary in order to effect universal computation purely through braiding operations because Abelian statistics results in the mere accrual of phases. Our initial description is at the level of effective field theories, where one is at scales much larger than the characteristic size of quasiparticles, which are treated as point-like. By canonically quantizing these effective field theories, we can rephrase them in combinatorial terms. The combinatorial construction of TQFTs has a natural interpretation as an *intermediate scale* description. At this scale, the system can be described in terms of fluctuating curves. These curves must arise from shorter-distance physics and their dynamics must effectively impose certain constraints which lead to to the desired effective field theories at longer wavelengths. The final step in this program is constructing microscopic models of interacting electrons, which will be the subject of a later paper. There, we will explore models which describe physics at the lattice scale. In order to produce the physics in which we are interested, these models must give rise, through their local interactions, to domain-wall-like structures which are the input for the combinatorial construction.

In essence, the current state of understanding is that topological phases are related to systems which, at low-energies can be understood as composed of fluctuating *loops*. In mathematical terms, the modular functor is *localized* to *curves*. This is, for the most part, the phenomenon which we describe in this paper. The next step is to find *local* Hamiltonians from which fluctuating loops emerge as low-energy degrees of freedom. This would be the localization of these modular functors to *points*.

2. Doubled Chern–Simons theories

Chern–Simons theories break parity and time-reversal symmetries: a clockwise braid is not the same as a counter-clockwise one; it is its conjugate. There is a simple way to start with such a theory and make a parity and time-reversal invariant theory: take two decoupled copies with opposite chirality. The resulting theory is called the 'doubled' theory or the 'squared' theory. The particle spectrum is now squared in size since it is the tensor product of the spectra of the two theories (which are mirror

images of each other). This is a completely trivial procedure for a Chern–Simons theory, which hardly warrants a special name. However, there are certain pathological topological field theories which can be fixed by 'doubling' them and then specifying non-trivial braiding statistics between particles in the two tensor factors. As Drinfeld showed, this can be done in such a way as to cure the pathologies [40].

The Chern–Simons theories of the Halperin (m,m,m) quantum Hall states[41] are examples of such 'pathological' theories. These are $U(1)\times U(1)$ Chern–Simons theories which can be written in a form in which one of the Chern–Simons terms has vanishing coefficient. Neutral excitations with non-vanishing spin quantum numbers (or isospin if the second component is a second layer rather than the opposite spin) have trivial braiding properties, unlike in the (m,m,n), $m\neq n$ states. As a result, the spin sector can condense and be gapless [42], which is a 'pathology' from the point of view of topological field theory. (From a more conventional physical standpoint, it's not a bug, it's a feature—and a remarkable one.) It can be 'fixed' by Drinfeld's procedure if one introduces a second, oppositely directed, (m,m,m) state (which one might wish to call a $(-m,-m,-m)$ state) and assigning relative statistics between the two spin sectors. It is not obvious that this can be done consistently in all such 'pathological' cases, but it can, as Drinfeld showed.

'Doubling' initially appears very unnatural from a physical point of view. Why should two opposite chirality copies of a theory arise? One might imagine semions living on one sublattice of a bipartite lattice and anti-semions living on the other or, perhaps, up-spin semions and down-spin anti-semions. However, these sound more like clutching at straws in an effort to preserve time-reversal invariance; in either case, the combination of time-reversal with translation or spin-rotation is still broken. However, 'doubling' can occur in a more organic way. In fact, it is not just a completely natural occurrence in our models, it is absolutely unavoidable. For instance, the simple model given in Section 8 is a 'doubled chiral spin-liquid' while, thus far, no model with short-ranged interactions has been shown to have an undoubled chiral spin-liquid ground state. In order to eliminate doubling, additional structure—possibly unnatural—must be added to the models.

From an algebraic point of view, it is useful to observe that in a system in a magnetic field, translations in the x and y directions do not commute. This is expressed in Chern–Simons theory by the statement that in the canonical formalism, the x and y components of the gauge field do not commute. Thus, they cannot be simultaneously diagonalized. In this sense, a Chern–Simons gauge field is 'half' of a gauge field: only half of its degrees of freedom can be specified in a given state. By doubling the theory, or taking two copies, we have, in a sense, doubled 'half' of a gauge field, thereby yielding one gauge field. Thus, it is possible, in the doubled theory to have a basis of states which are eigenstates of the Wilson loop operators—which are the natural gauge-invariant operators—associated with a gauge field. In the Abelian case, this can be done completely explicitly. We can construct a gauge field from the two opposite chirality gauge fields, and in its diagonal basis, the Wilson loop operators are analogous to number operators N, and operators which do not commute with them, such as the Wilson loop operators for either of the chiral gauge fields, are roughly analogous to creation and annihilation operators $a^{\dagger},a$. In the non-Abelian case, this

basis cannot be related in such a simple way to the original gauge fields, but roughly the same structure is present.

We will see that systems which have configurations which can be mapped onto those of a loop gas have a Hilbert space which naturally admits such a basis. Thus, if they enter a topological phase, we expect it to automatically be a doubled one.

3. Parton model construction for correlated electron systems

A useful heuristic for understanding the physics of a correlated electron Hamiltonian (in zero or non-zero magnetic field) involves rewriting the electron operator in terms of auxiliary 'parton' operators or slave fermion/boson operators. This strategy can also be applied to interacting bosons models, which could be realized in materials in which Cooper pairing leads to the emergence of effective bosonic degrees of freedom corresponding to the pairs. In this section, we will use the parton formalism to describe the phases of interest.

Consider a model of spins on a lattice. Suppose that there is spin $S = N/2$ at each site. The Hamiltonian may be of the form

$$H = J \sum_{\langle i,j \rangle} \mathbf{S}_i \cdot \mathbf{S}_j + \cdots, \tag{1}$$

where the ellipses denotes other terms including, perhaps, ring exchange terms or next-neighbor interactions. We now introduce an $SU(N)$ multiplet of spin-1/2 fermions $f_{a\alpha}$, with $a = 1, 2, \ldots, N$ and $\alpha = \uparrow, \downarrow$ so that

$$\mathbf{S}_i = f^\dagger_{a\alpha i} \sigma_{\alpha\beta} f_{a\beta i}. \tag{2}$$

The fermions must satisfy the constraints that there be N fermions per site:

$$f^\dagger_{a\alpha i} f_{a\alpha i} = N \tag{3}$$

and that there be a color singlet at each site

$$f^\dagger_{a\alpha i} T^{\underline{k}}_{ab} f_{a\beta i} = 0, \tag{4}$$

where $T^{\underline{k}}_{ab}$, $\underline{k} = 1, 2, \ldots, N^2 - 1$ are the generators of $SU(N)$ in the fundamental representation. These constraints guarantee that there is spin $N/2$ at each site.

Using this representation, we can rewrite the Hamiltonian in the form

$$H = -J \sum_{\langle i,j \rangle} f^\dagger_{a\alpha i} f_{b\alpha j} f^\dagger_{b\beta j} f_{a\beta i} + \cdots \tag{5}$$

In mean-field approximation, this Hamiltonian can be written as:

$$H = -J \sum_{\langle i,j \rangle} U_{ab,ij} f^\dagger_{a\alpha i} f_{b\alpha j} + \cdots, \tag{6}$$

where

$$U^{ij}_{ab} = \left\langle f^\dagger_{b\beta j} f_{a\beta i} \right\rangle. \tag{7}$$

Consider the following mean-field solution:

$$U_{ab}^{ij} = \left(\frac{t_{\text{eff}}}{J}\right) e^{i\phi_{ij}} \delta_{ab}, \tag{8}$$

where t_{eff} is a parameter to be determined from the saddle-point condition. Assuming that the system is on a square lattice of side a, we suppose that

$$\phi_{ij} = \frac{\pi}{k}(x_j - x_i)(y_j + y_i)/a^2 \tag{9}$$

so that

$$\prod_{\text{plaq.}} e^{i\phi_{ij}} = e^{2\pi i/k}. \tag{10}$$

Here, k is assumed to be an even integer. In this solution, the system generates an effective magnetic field for the $f_{a\alpha}$s such that there is half of a flux quantum per plaquette. The commensurability of the field ensures that precisely $k/2$ Landau levels are filled by each of the $f_{a\alpha}$s.

Does this solution actually occur? Is it stable? The answers to these questions depend, for the most part, on the particulars of the given Hamiltonian. Later in this paper, we will take some steps towards addressing them. While we cannot determine the absolute stability of this solutions using the approach of this section, we can test its local stability against small fluctuations. There are certain fluctuations which, a priori, are likely to be important. Consider the gauge transformations $f_{a\alpha i} \to W_{ab}^i f_{b\alpha i}, U_{ac}^{ij} \to W_{ab}^i U_{bd}^{ij} W_{dc}^{j\dagger}$, where W_{ab}^i are $U(N)$ matrices assigned to the points i of the lattice. These transformations leave the Hamiltonian invariant. Hence, the class of mean-field configurations

$$U_{ac}^{ij} = \left(\frac{t_{\text{eff}}}{t}\right) e^{i\phi_{ij}} \left(e^{ig_{ij}}\right)_{ac} \tag{11}$$

have vanishing energy cost if $g_{ij} = \Lambda_i - \Lambda_j$ for some $u(N)$ Lie algebra-valued matrices Λ_i, i.e., if g_{ij} is pure gauge. The energy cost should be a gauge-invariant function of g_{ij}; since gauge field fluctuations can be large we must consider them carefully. Let us break the $U(N)$ gauge field g_{ij} into $U(1)$ and $SU(N)$ parts, c_{ij} and a_{ij}, respectively. The latter is a traceless, Hermitian, $N \times N$ matrix, i.e., an $SU(N)$ Lie algebra-valued field. The time components of these gauge fields, c_0 and a_0, can be introduced as Lagrange multipliers which enforce the constraints (3) and (4).

To see that the gauge field fluctuations do not destabilize our mean-field solution, we integrate out the $f_{a\alpha}$s. The crucial point is that the $f_{a\alpha}$s fill $k/2$ Landau levels in the mean-field solution. By standard arguments, we then have the following effective action for c and a

$$\begin{aligned} S[a] &= NkS_{\text{CS}}[c] + kS_{\text{CS}}[a] \\ &= \frac{Nk}{4\pi}\int \epsilon^{\mu\nu\rho} c_\mu \partial_\nu c_\rho + \frac{k}{4\pi}\int \epsilon^{\mu\nu\rho}\left(a_\mu^{\underline{i}}\partial_\nu a_\rho^{\underline{i}} + \frac{2}{3} f_{\underline{ijk}} a_\mu^{\underline{i}} a_\nu^{\underline{j}} a_\rho^{\underline{k}}\right) \\ &= \frac{Nk}{4\pi}\int c \wedge \mathrm{d}c + \frac{k}{4\pi}\int \mathrm{tr}\left(a \wedge da + \frac{2}{3} a \wedge a \wedge a\right). \end{aligned} \tag{12}$$

The effective actions for the $U(1)$ gauge field c_μ and the $SU(N)$ gauge field $a_\mu^{ab} = a_\mu^i T_i^{ab}$ are the corresponding Chern–Simons actions at level k, i.e., with coupling constant k. Since each $f_{a\alpha}$ fills $k/2$ Landau levels, their response to an external gauge field must break P, T and be proportional to k; the requirement of gauge invariance then dictates (12), up to multiplication by an arbitrary integer. A direct calculation shows that this integer is 1. For the $U(1)$ gauge field, it is just the Hall conductivity of $k/2$ filled Landau levels of both spins and N colors. We have suppressed subleading terms, such as the Maxwell terms for c_μ and a_μ^{ab}, which should appear in Eq. (12) since they are irrelevant in the low-energy limit.

The evenness of the level k results from the presence of equal densities of up- and down-spin fermions. If spin-rotational symmetry were broken by the presence of different densities of $f_{aA\uparrow}$s and $f_{aA\downarrow}$s, then odd level k could also result.

The effective field theory (12) is a gapped theory, so gauge field fluctuations about this mean-field solution are *not* large [43]. In other words, if the configuration (8) is a saddle-point for some Hamiltonian, then this saddle-point is stable against fluctuations of the form (11). Said differently, the breaking of P, T permits the appearance of a 'mass term' for the gauge field—the Chern–Simons term—which renders the phase stable.

The appearance of a Chern–Simons term has another important effect: the excitations of the theory have exotic statistics. As a result of their interaction with the gauge field, the $f_{a\alpha}$s are anyons and, (except for the special case $k = 1$ which cannot occur anyway in this construction), their braiding statistics is non-Abelian. (It is not quite consistent to talk about the $f_{a\alpha}$s interacting with a Chern–Simons gauge field a_μ^i since we had to integrate out the $f_{a\alpha}$s in order to generate the Chern–Simons term for a_μ^i. However, we can introduce external source fields for the $f_{a\alpha}$s. When the $f_{a\alpha}$s are integrated out, a coupling between the source fields and a_μ^i is generated, so that when these sources are braided, the result is non-trivial.)

The phases which we have just discussed share the attractive features that they are stable and that they support quasiparticle excitations with exotic braiding statistics. They also spontaneously break parity and time-reversal invariance. The latter would seem, at first glance, to be a necessary condition for exotic statistics. However, the closely related 'doubled' theories, which we discussed in the previous section, also support exotic statistics, but they do not break P, T. This is a useful feature in, for instance, a theory of a material in which P, T violation has been experimentally ruled out. We will later discuss why such phases are likely to occur in certain types of models, but, for now, let us simply consider 'doubled' theories as another logical possibility.

Let us consider a model in which there is an integer spin N at each site of the lattice. Suppose we now introduce $2N$ spin-1/2 fermions $f_{aA\alpha}$, with $a = 1, 2, \ldots, N$, $A = \pm$, and $\alpha = \uparrow, \downarrow$ so that

$$\mathbf{S}_i = f_{aA\alpha i}^\dagger \sigma_{\alpha\beta} f_{aA\beta i}. \tag{13}$$

The fermions must satisfy the constraints that there be $2N$ fermions per site:

$$f_{a\alpha i}^\dagger f_{a\alpha i} = 2N \tag{14}$$

and that there be a color singlet at each site

$$f^{\dagger}_{aA\alpha i} T^{\underline{k}}_{abAB} f_{aB\beta i} = 0, \tag{15}$$

where $T^{\underline{k}}_{abAB}$, $\underline{k} = 1, 2, \ldots, 4N^2 - 1$ are the generators of $SU(2N)$ in the fundamental representation. These constraints guarantee that there is spin N at each site.

Following the steps which we made earlier, we make a mean-field approximation in which the Hamiltonian takes the form:

$$H = -J \sum_{\langle i,j \rangle} U_{abAB,ij} f^{\dagger}_{aA\alpha i} f_{bB\alpha j} + \cdots, \tag{16}$$

where

$$U^{ij}_{abAB} = \left\langle f^{\dagger}_{bB\beta j} f_{aA\beta i} \right\rangle \tag{17}$$

and consider the following mean-field solution:

$$\begin{aligned} U^{ij}_{ab++} &= \left(\frac{t_{\text{eff}}}{J}\right) \mathrm{e}^{\mathrm{i}\phi_{ij}} \delta_{ab}, \\ U^{ij}_{ab--} &= \left(\frac{t_{\text{eff}}}{J}\right) \mathrm{e}^{-\mathrm{i}\phi_{ij}} \delta_{ab}, \\ U^{ij}_{ab+-} &= U^{ij}_{ab-+} = 0 \end{aligned} \tag{18}$$

with

$$\phi_{ij} = \frac{\pi}{k} (x_j - x_i)(y_j + y_i)/a^2. \tag{19}$$

Again, k is assumed to be an even integer. The system generates an effective magnetic field for the $f_{a\alpha}$s in which $k/2$ Landau levels are filled by each of the $f_{a\alpha}$s.

Turning now to fluctuations about this saddle-point, we see that there are two $SU(N)$ gauge fields to go with the $U(1)$ gauge field. The low-energy fluctuations are of the form

$$\begin{aligned} U^{ij}_{ab++} &= \left(\frac{t_{\text{eff}}}{J}\right) \mathrm{e}^{\mathrm{i}\phi_{ij}} \mathrm{e}^{\mathrm{i}c_{ij}} \left(\mathrm{e}^{\mathrm{i}a^{+}_{ij}}\right)_{ab}, \\ U^{ij}_{ab--} &= \left(\frac{t_{\text{eff}}}{J}\right) \mathrm{e}^{-\mathrm{i}\phi_{ij}} \mathrm{e}^{\mathrm{i}c_{ij}} \left(\mathrm{e}^{\mathrm{i}a^{-}_{ij}}\right)_{ab}, \\ U^{ij}_{ab+-} &= U^{ij}_{ab-+} = 0. \end{aligned} \tag{20}$$

The theory actually has a $U(1) \times SU(2N)$ gauge symmetry. However, the saddle-point is only invariant under global $U(1) \times SU(N) \times SU(N)$ transformations. The other global $SU(2N)$ transformations are broken at the saddle-point level and, hence, the corresponding gauge fields are massive by the Anderson–Higgs mechanism.

Equivalently [34], symmetry allows a term in the effective action of the form

$$\mathrm{tr}\left(F_{ijkl} \mathrm{e}^{\mathrm{i}a_{ji'}} F_{i'j'k'l'} \mathrm{e}^{-\mathrm{i}a_{ji'}}\right), \tag{21}$$

where $ijkl$ and $i'j'k'l'$ are two plaquettes which are one lattice spacing apart so that they are connected by the link ji'. At the saddle-point, $F_{ijkl} = F_{i'j'k'l'} =$

$F = \mathrm{diag}(e^{2\pi i/k}, e^{-2\pi i/k})$ which which commutes with the elements of $SU(N) \times SU(N)$ but not the remaining elements of $SU(2N)$. Expanding (21) to second order in $a_{j'i}$, we find a term

$$\mathrm{tr}\Big(F\Big[[F, a_{ji'}], a_{ji'}\Big]\Big), \tag{22}$$

which is a mass term for the gauge fields associated with $SU(2N)$ generators which do not commute with $\mathrm{diag}(e^{2\pi i/k}, e^{-2\pi i/k})$, i.e., those not in $SU(N) \times SU(N)$.

The effective action for the $U(1) \times SU(N) \times SU(N)$ gauge fields can be derived by integrating out the fermions. We find:

$$S[a] = kS_{\mathrm{CS}}[a^+] - kS_{\mathrm{CS}}[a^-]. \tag{23}$$

Since the $f_{a-\alpha}$ fermions move in an effective field of flux $-2\pi/k$, there is a negative sign in front of the second term in (23). Furthermore, there is a cancellation between the contributions of the $f_{a+\alpha}$ and $f_{a-\alpha}$ fermions to the coefficient of the Chern–Simons term for the $U(1)$ gauge field c_μ. Hence, we have two opposite chirality $SU(N)$ Chern–Simons gauge fields. They are gapped and lead to exotic braiding statistics, but they preserve P and T since we can exchange a^+ and a^- when we perform time-reversal or a parity transformation. We will call these theories $SU(2)_k \times \overline{SU(2)_k}$ Chern–Simons theory, with the overline signifying that the second Chern–Simons term has opposite chirality.

The $U(1)$ gauge field does not have a Chern–Simons term. Thus, it is in one of two massive phases. (1) A confining phase in which gauge field fluctuations cause $U(1)$ charge to be confined and c_μ to be gapped. (2) A Z_2 phase which results when $\epsilon_{ab}\epsilon_{AB}\epsilon_{\alpha\beta}f_{aA\alpha}f_{bB\beta}$ condenses, breaking $U(1)$ to Z_2. The remaining Z_2 gauge field has a phase in which Z_2 charges are deconfined. The former theory can be considered as a subset of the latter: both of these phases have quasiparticles with exotic braiding statistics, but the former only has those which are created by fermion bilinears such as $f^\dagger_{aA\alpha}f_{bB\beta}$ while the latter has the full set of fractionalized quasiparticles. In the $SU(2)$ case, the former theory has only integer-'spin' (i.e., under the internal $SU(2) \times SU(2)$ gauge symmetry) quasiparticles while the latter has half-integer as well. We will call the former the 'even part' of $SU(2)_k \times \overline{SU(2)_k}$ Chern–Simons theory.

We have introduced these phases in the context of spin systems, but they can arise in a variety of contexts. The above analysis can be extended to finite doping in the usual way. We can also expect such phases in, for instance, correlated boson models. Consider an example of the latter,

$$H = -t\sum_{\langle i,j\rangle}\left(B_i^\dagger B_j + \mathrm{h.c.}\right) + \sum_{\langle i,j\rangle} V_{ij}N_iN_j, \tag{24}$$

where i, j label sites on a lattice, B_i creates a Cooper pair on site i, and $N_i = B_i^\dagger B_i$. We will assume that $V_{ii} = \infty$ so that we have hard-core bosons. Let us assume that there are precisely half as many bosons as lattice sites. We assume that $V_{ij} > 0$. For V_{ij} small, the system will be superconducting. For V_{ij} sufficiently large, the system will be insulating. We explore possible insulating phases. In order to do this, we write B_i in the following representation:

$$B_i = \frac{1}{\sqrt{N!}} \epsilon_{a_1 a_2 \ldots a_N} f_{a_1 i} f_{a_2 i} \cdots f_{a_N i}, \tag{25}$$

where $a_k = 1, 2, \ldots, N$. We have introduced an even number N of auxiliary fermions f_{ai} such that the boson B_i is an $SU(N)$ symmetric bound state of them.

Focusing on the hopping term, we see that the Hamiltonian can now be written as:

$$H = -\frac{t}{N!} \sum_{\langle i,j \rangle} \left(\epsilon_{a_1 a_2 \cdots a_N} f^\dagger_{a_1 i} f^\dagger_{a_2 i} \cdots f^\dagger_{a_N i} \times \epsilon_{b_1 b_2 \cdots b_N} f_{b_1 j} f_{b_2 j} \cdots f_{b_N j} + \text{h.c.} \right) + \cdots \tag{26}$$

The ellipses represent next-nearest-neighbor and longer-ranged interactions. In mean-field approximation, this Hamiltonian can be written as:

$$H = -t \sum_{\langle i,j \rangle} \left(U^{ij}_{\underline{ab}} f^\dagger_{ai} f_{bj} + \text{h.c.} \right) + \cdots, \tag{27}$$

where

$$U^{ij}_{\underline{ab}} = \frac{t_{\text{eff}}}{tN!} \epsilon_{a a_2 \ldots a_N} \epsilon_{b b_2 \ldots b_N} \times \left\langle f^\dagger_{a_1 i} \cdots f^\dagger_{a_N i} f_{b_2 j} \ldots f_{b_N j} \right\rangle. \tag{28}$$

Consider the following mean-field solution:

$$U^{ij}_{ab} = \left(\frac{t_{\text{eff}}}{t} \right) e^{i\phi_{ij}} \delta_{ab}, \tag{29}$$

where

$$\phi_{ij} = \frac{\pi m}{N} (x_j - x_i)(y_j + y_i)/a^2 \tag{30}$$

for some integer m. Only such fluxes are allowed since the bosons B must see vanishing flux. If the f_as see flux $2\pi m/N$ then the Bs see flux $2\pi m$ which is gauge equivalent to zero. Suppose that $2m$ is a divisor of N so that $k = N/2m$ is an integer. Then the f_as fill k Landau levels and, integrating out the fermions, we find that the effective field theory for fluctuations about this saddle-point is $SU(2mk)_k$ Chern–Simons theory.

The doubled version of this theory can be obtained by using a representation with two sets of N fermions f_{+a} and f_{-a}:

$$B_i = \frac{1}{\sqrt{2N!}} \epsilon_{a_1 a_2 \ldots a_N} f_{+a_1 i} f_{+a_2 i} \cdots f_{+a_N i} + \frac{1}{\sqrt{2N!}} \epsilon_{a_1 a_2 \ldots a_N} f_{-a_1 i} f_{-a_2 i} \cdots f_{-a_N i} \tag{31}$$

with a mean-field

$$\begin{aligned} U^{ij}_{ab++} &= \left(\frac{t_{\text{eff}}}{t} \right) e^{i\phi_{ij}} \delta_{ab}, \\ U^{ij}_{ab--} &= \left(\frac{t_{\text{eff}}}{t} \right) e^{-i\phi_{ij}} \delta_{ab}, \\ U^{ij}_{ab+-} &= U^{ij}_{ab-+} = 0. \end{aligned} \tag{32}$$

The effective field theory for fluctuations about this state is $SU(2mk)_k \times \overline{SU(2mk)_k}$.

This construction can be adapted to a system of itinerant electrons, for which we introduce an odd number of partons whose bound state is an electron. In this way and also by using other straightforward generalizations of the above constructions, we can obtain other phases of correlated electron systems which are described by the $SU(N)_k \times \overline{SU(N)_k}$ family of effective field theories.

In the remainder of this paper, we will focus on the case $N = 2$, i.e., the $SU(2)_k \times \overline{SU(2)_k}$ Chern–Simons theories. These can actually arise in two different ways from the above constructions. One is via the obvious path: a spin-2 magnet, for instance, can have an $SU(2)_k \times \overline{SU(2)_k}$ phase. The alternative route follows from *rank-level duality*: a spin-k magnet can have an $SU(k)_2 \times \overline{SU(k)_2}$ phase, which is equivalent to $SU(2)_k \times \overline{SU(2)_k}$. The latter construction suggests that the higher-level phases should be sought in higher-spin magnets, an observation on which we will comment further at the end of this paper.

The parton or slave-particle representations which we have discussed in this section have a Z_2 gauge symmetry under which f_+ and f_- are exchanged independently at each point. (In the spin systems which we discussed earlier, this Z_2 is a subgroup of $SU(2N)$.) This symmetry is broken at the saddle-point. P, T are also broken. However, the combination of a Z_2 gauge transformation and T or P leaves the saddle-point invariant. Thus, we have the symmetry-breaking pattern $Z_2^{\text{gauge}} \times Z_2^T \to Z_2^{\text{diagonal}}$. In this way, P, T are preserved since they can be identified with the diagonal Z_2 which remains. This is similar to the preservation of rotational invariance in monopole solutions of gauge theories: the solutions are not rotationally invariant, but the combination of a rotation and a gauge transformation leaves the solution invariant.

To summarize, we have seen in this section how the parton or slave-particle formalisms often used to discuss strongly correlated electron systems can lead to stable phases corresponding to doubled Chern–Simons theories. Thus far, we have not discussed why the doubled theories might arise rather than the undoubled ones or even some entirely different phases. In other words, the question which we have not addressed is why should one of these mean-field solutions have the minimum energy? In order to begin to answer this question, we will take a somewhat circuitous route which will involve solving them first. Special properties of the doubled theories will emerge once we have discussed them more fully. In order to understand the physics of these theories, we will need a set of non-perturbative methods, discussed in the following sections.

4. Solution of doubled Abelian theories

The Chern–Simons theories (both undoubled and doubled) which we have just encountered are topological field theories. Thus, they have only a finite number of degrees of freedom and are 'trivial' in the sense that they are a problem in quantum mechanics, rather than quantum field theory. On the other hand, the physics which they describe is non-trivial precisely because it is topologically invariant and measurable physical properties probe the topology of the manifold on which they live.

Thus, these theories are soluble, but not with the methods ordinarily used to solve field theories. In particular, we will need to use the canonical formalism almost entirely, although ideas from the functional integral approach will prove useful. In this section, we present the solution of these theories, which takes the form of a construction of their Hilbert spaces, together with physical observables acting on these Hilbert spaces. Their algebraic structure reveals connections with theories of exactly soluble statistical mechanical models and, perhaps more importantly, offers some valuable clues about which microscopic models might give rise to these topological phases. In this paper, we will make a few comments about microscopic models and defer a more serious discussion to a following paper.

Topological field theories have vanishing Hamiltonian, so they do not describe dynamics in the ordinary sense in which, say, a non-linear sigma model would. They only describe the braiding properties of quasiparticles and the sensitivity of the system to the topology of the surface on which it is realized. Consequently, the entire problem of solving these theories amounts to constructing their Hilbert spaces; there is no energy spectrum to compute because all states have zero energy. In other words, the problem is an algebraic one of constructing these Hilbert spaces as the representation spaces of the commutator algebra of the physical observables of the theories. In this section, we solve this problem for the 'warm-up' case of Abelian theories.

4.1. Abelian Chern–Simons theory

Let us begin our discussion of the effective field theories for topological phases of interacting electron systems with one of the simplest such theories, Abelian Chern–Simons theory:

$$S = \frac{m}{4\pi} \int \epsilon^{\mu\nu\rho} a_\mu \partial_\nu a_\rho. \tag{33}$$

The coefficient is chosen to be an integer divided by 4π in order to ensure invariance under large gauge transformations. A more refined effective field theory could include further terms such as a Maxwell term. However, these terms are irrelevant in the infrared, so we drop them.

Let us assume that we are working at energies much lower than that required to excite any of the particles to which the gauge field is coupled, so that we can consider (33) in isolation. This theory would appear to be completely trivial. Suppose we take Coulomb gauge, $a_0 = 0$. In taking this gauge, we must remember to impose the constraint which follows from varying a_0 in (33)

$$\nabla \times \mathbf{a} = 0. \tag{34}$$

In this gauge, the Lagrangian takes the form

$$\mathscr{L} = \frac{m}{2\pi} a_2 \partial_0 a_1 = \Pi_1 \partial_0 a_1 - 0, \tag{35}$$

where the canonical momentum conjugate to a_1 is

$$\Pi_1 \equiv \frac{\partial \mathscr{L}}{\partial(\partial_0 a_1)} = \frac{m}{2\pi} a_2. \tag{36}$$

From the second equality of (35), we see that the Hamiltonian vanishes. Thus, the effective action only describes the ground state—or states. The interesting structure of the theory follows from the canonical equal-time commutation relation

$$\left[a_1(\mathbf{x},t),\Pi_1(\mathbf{x}',t)\right]=\mathrm{i}\delta(\mathbf{x}-\mathbf{x}') \tag{37}$$

or

$$\left[a_1(\mathbf{x},t),a_2(\mathbf{x}',t)\right]=\frac{2\pi\mathrm{i}}{m}\delta(\mathbf{x}-\mathbf{x}'). \tag{38}$$

Note that the all-important integer m appears here in the commutation relation because it appears in front of the Chern–Simons action.

On the infinite plane or the sphere, the ground state is a unique, non-degenerate state. Pure Chern–Simons theory (i.e., without any other fields coupled to it) has no other states. However, suppose that the theory is defined on the torus. Then a gauge field satisfying $\nabla\times\mathbf{a}=0$ can still give rise to a non-trivial holonomy $W[\gamma]$ around the closed curve γ if γ winds around one of the non-trivial cycles of the torus.

$$W[\gamma]=\mathrm{e}^{\oint_\gamma \mathbf{a}\cdot\mathrm{d}\mathbf{l}}. \tag{39}$$

According to the constraint, $W[\gamma]$ does not depend on the precise curve γ but only on how many times it winds around the generators of the torus, i.e., on its homotopy class. Furthermore, it is clear that $W[\gamma]$ is multiplicative in the sense that its value for a curve γ which winds twice around one of the generators of the torus is the square of its value for a curve γ which winds once. Hence, we have only two independent variables. This is revealed by solving the constraint. In a coordinate system $\theta_1,\theta_2\in[0,2\pi]$ on the torus, we have $\mathbf{a}=(\alpha_1/2\pi,\alpha_2/2\pi)+\nabla\varphi$. If we take φ to be single-valued, then invariance under large gauge transformations requires that we make the identification $\alpha_i\equiv\alpha_i+1$. W_i and α_i are related by

$$W_i\equiv\mathrm{e}^{\oint_{\gamma_i}\mathbf{a}\cdot\mathrm{d}\mathbf{l}}=\mathrm{e}^{\mathrm{i}\alpha_i}. \tag{40}$$

From (38), we have the following equal-time commutation relations:

$$[\alpha_1,\alpha_2]=\mathrm{i}\frac{2\pi}{m}. \tag{41}$$

Since α_1,α_2 are not themselves gauge-invariant, we cannot simply use the analogy between their commutation relations and those of p, x for a single particle. We must work with the gauge invariant quantities W_i. Since

$$\mathrm{e}^{\mathrm{i}\alpha_1}\mathrm{e}^{\mathrm{i}\alpha_2}=\mathrm{e}^{[\alpha_1,\alpha_2]/2}\mathrm{e}^{\mathrm{i}\alpha_1+\mathrm{i}\alpha_2} \tag{42}$$

we have the commutation relation

$$W_1W_2=\mathrm{e}^{2\pi\mathrm{i}/m}W_2W_1. \tag{43}$$

This algebra can be implemented on a vector space in the following way:

$$\begin{aligned}&W_1|n\rangle=\mathrm{e}^{2\pi n\mathrm{i}/m}|n\rangle,\\&W_2|n\rangle=|n+1\rangle\end{aligned} \tag{44}$$

with $n \in Z$. This is a representation of this algebra on an infinite-dimensional Hilbert space. However, there are finite-dimensional truncations which also allow a representation of this algebra. Suppose that we simply restrict $n \in \{j, j+1, j+2, \ldots, j+qm-1\}$ for any integer j and any positive integer q. and define $W_2|j+qm-1\rangle = |j\rangle$. Such a vector space has dimension qm and it gives a perfectly good representation of the algebra. Which value of q gives the Hilbert space of $U(1)_m$ Chern–Simons theory?

To answer this question, first note that the phase space of the classical theory has finite volume. It is parameterized by $\alpha_{1,2} \in [0, 2\pi]$ defined in (40). This is analogous to the case for a single spin, where phase space is the surface of a sphere S^2, but it is unlike the case of a particle constrained to lie on a circle, where the coordinate takes values on a circle but the momentum is arbitrary, so that *phase space* is $S^1 \times R$. From the commutation relation for α_1, α_2, we see that α_1 and $m\alpha_2/2\pi$ are canonically conjugate coordinates on phase space; thus phase space has volume $2\pi m$. We expect the dimension of Hilbert space to be roughly equal to the volume of phase space, measured in units of $h = 2\pi\hbar = 2\pi$ (since we have set $\hbar = 1$). In the classical – or large m—limit, this should be an exact equality, so the only choice is the minimum possible one allowed by the algebra (43), $q = 1$, i.e., dimension m.

We can restate this by saying that the Hilbert space of the theory is obtained from the infinite-dimensional one of (44) by requiring that all physical states be annihilated by the projection operator

$$\sum_n (|n\rangle - |n+m\rangle)(\langle n| - \langle n+m|). \tag{45}$$

To summarize, the Hilbert space of $U(1)_m$ Chern–Simons theory is spanned by the basis vectors $|n\rangle$ with $n = 0, 1, \ldots, m-1$, i.e., the ground state is m-fold degenerate. On a genus g manifold, this generalizes to m^g. The inner product $\langle n|n'\rangle = \delta_{nn'}$ is fixed by the requirement that W_1 and W_2 be unitary.

This has an interpretation in terms of the (quasi)particle spectrum of the theory—about which we might have thought that we would lose all information at low energies. Imagine creating a quasihole–quasiparticle pair, taking them around the meridian of the torus and annihilating them; call the corresponding operator T_1. Let T_2 be the operator for such a process around the longitude. If the quasiparticles have statistics π/m, then

$$T_1 T_2 = \mathrm{e}^{2\pi\mathrm{i}/m} T_2 T_1 \tag{46}$$

because the particles wind around each other during such a process, as depicted on the right of Fig. 1. This is precisely the same algebra (43) which we found above, with representations of minimal dimension m.

Hence, if we know that the ground state degeneracy of a system on a genus-g manifold is m^g, then one explanation of this degeneracy is that it has non-trivial quasiparticles of statistics $0, \pi/m, \ldots, (m-1)\pi/m$.

One awkward feature of the basis which we have just constructed for the Hilbert space of Abelian Chern–Simons theory is that we had to choose a particular direction on the torus. If θ_1 is the coordinate along the meridian of the torus, then the

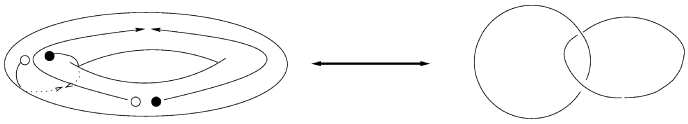

Fig. 1. The operation $T_1T_2T_1^{-1}T_2^{-1}$ results in a phase $e^{2\pi i/m}$ because it is equivalent to the braiding operation on the right.

Wilson loop operator W_1 corresponding to the meridian is diagonal in this basis, while the Wilson loop operator taken along the longitude is not. This appears to be unavoidable because the two operators do not commute. However, by considering the 'doubled' theory, a theory with two Chern–Simons fields with equal but opposite coupling constants, m and $-m$, we can have a more democratic-looking Hilbert space. This is more than just an aesthetic requirement, since the simplest microscopic models give rise to such Hilbert spaces, as we will see later.

4.2. Doubled Abelian Chern–Simons theory

The Hilbert space of the theory with action

$$S = \frac{m}{4\pi}\int \epsilon^{\mu\nu\rho}a_\mu\partial_\nu a_\rho - \frac{m}{4\pi}\int \epsilon^{\mu\nu\rho}c_\mu\partial_\nu c_\rho \tag{47}$$

is clearly just the tensor product of the Hilbert spaces associated with the two terms in the action or, in other words, the tensor product of the Hilbert space (44) with its complex conjugate (the conjugation results from the minus sign in front of the second term in Eq. (47). We will usually call this theory $U(1)_m \times \overline{U(1)}_m$ Chern–Simons theory, but we will sometimes call the theory the 'double' or 'Drinfeld double' of $U(1)_m$ Chern–Simons theory. This terminology is unnecessary in this simple case, but there are instances of pathological chiral topological field theories whose pathologies can be cured by doubling them and adding some extra structure, as Drinfeld found.

For simplicity, let us consider the case $m = 2$. Now, following (39), we can define the operators $W_+[\gamma]$ associated with the gauge field a_μ and the analogous operators $W_-[\gamma]$ associated with c_μ. We now have the tensor product of two operator algebras:

$$\begin{aligned} W_+[\gamma]W_+[\gamma'] &= (-1)^{I(\gamma,\gamma')}W_+[\gamma']W_+[\gamma],\\ W_-[\gamma]W_-[\gamma'] &= (-1)^{I(\gamma,\gamma')}W_-[\gamma']W_-[\gamma],\\ W_+[\gamma]W_-[\gamma'] &= W_-[\gamma']W_+[\gamma], \end{aligned} \tag{48}$$

$I(\alpha,\gamma)$ is the intersection number of α and γ. It is useful to define the operators $L[\gamma] = W_+[\gamma]W_-[\gamma]$, which commute with each other:

$$L[\gamma]L[\gamma'] = L[\gamma']L[\gamma]. \tag{49}$$

We introduced these operators to emphasize the point that (48) is not the tensor product of just any two algebras pulled off the street, but of two identical ones. As a result, there is a commuting set of operators $L[\gamma]$ associated with curves γ. For future reference, we display the relations which the $L[\gamma]$s obey with $W_+[\gamma]$, $W_-[\gamma]$:

$$\begin{aligned} L[\gamma]W_+[\gamma'] &= (-1)^{I(\gamma,\gamma')}W_+[\gamma']L[\gamma], \\ L[\gamma]W_-[\gamma'] &= (-1)^{I(\gamma,\gamma')}W_-[\gamma']L[\gamma]. \end{aligned} \tag{50}$$

One way of representing this algebra is on the Hilbert space $|n_+, n_-\rangle$ with $n_+, n_- = 0, 1$

$$\begin{aligned} L_1|n_+, n_-\rangle &= (-1)^{n_+ - n_-}|n_+, n_-\rangle, \\ L_2|n_+, n_-\rangle &= |n_+ + 1, n_- + 1\rangle, \end{aligned} \tag{51}$$

where $L_{1,2} = L[\gamma]$ with γ a meridian or longitude, respectively. $W_+[\gamma]$ and $W_-[\gamma]$ act as one would anticipate from (44). As expected, the ground state degeneracy is $m^2 = 4$.

Since $L_{1,2}$ commute, they can be simultaneously diagonalized. They are diagonal in the following basis:

$$|\ell_1, \ell_2\rangle = \sum_{n=0}^{1}(-1)^{n\ell_2}|n + \ell_1, n\rangle \tag{52}$$

(The generalization to arbitrary m is clear: the sum ranges from 0 to $m - 1$, and (-1) is replaced by $e^{2\pi i/m}$.) Then

$$L_{1,2}|\ell_1, \ell_2\rangle = (-1)^{\ell_{1,2}}|\ell_1, \ell_2\rangle. \tag{53}$$

Meanwhile

$$\begin{aligned} W_{+1}|\ell_1, \ell_2\rangle &= (-1)^{\ell_1}|\ell_1, \ell_2 + 1\rangle. \\ W_{+2}|\ell_1, \ell_2\rangle &= |\ell_1 + 1, \ell_2\rangle \end{aligned} \tag{54}$$

with similar relations for W_-. Thus, we can think of $L_{1,2}$ as being analogous to number operators N_k while $W_{+1,2}$ are analogous to raising operators $a_k^\dagger$. The analogy is not quite right because $W_{+1,2}$ do not commute with each other. However, this is a useful analogy nevertheless. (On the torus, we actually can take a commuting set of 'raising/lowering' operators, W_{+1} and W_{-2}). We have one such 'number' operator for each generator of the torus.

Note that states $|\Psi\rangle$ can be written as wavefunctions in this basis:

$$\Psi[\ell_1, \ell_2] = \langle \ell_1, \ell_2|\Psi\rangle \tag{55}$$

$\Psi[\ell_1, \ell_2]$ maps two integers modulo 2 into the complex numbers.

A slightly more abstract representation of the same Hilbert space will prove useful when we generalize this construction to more complicated theories. It will also suggest connections with microscopic models. Again, we will work on the torus, but the extension to other surfaces is straightforward. The basic idea is to define wavefunctions on the space of curve configurations on a given surface. The discussion of the previous paragraph can be framed in these terms if we think of ℓ_1, ℓ_2 as defining a topological class of curves.

In order to do this in more general terms, we need a few definitions. We define the following notation: let $\{\alpha\}$ be an *isotopy class* of one-dimensional submanifolds of the torus. A one-dimensional submanifold, α, of the torus is simply a set of non-intersecting smooth curves. If the submanifold is connected, then it is a curve;

however, we want to allow the multi-component case. We will use the term *multi-curve* to denote such a one-dimensional manifold. α and α' are in the same *isotopy class* if they can be smoothly deformed into each other [44]. We define a 'pre-Hilbert space,' $\tilde{\mathscr{H}}$ by associating an abstract vector $|\{\alpha\}\rangle$ to every isotopy class, $\{\alpha\}$, of one-dimensional submanifolds of the torus. By forming all linear combinations with complex coefficients, we arrive at the vector space which serves as our pre-Hilbert space.

$$\tilde{\mathscr{H}} = \left\{ \sum_{\{\alpha\}} c_{\{\alpha\}} |\{\alpha\}\rangle \middle| c_{\{\alpha\}} \in C \right\}. \tag{56}$$

The vectors in our pre-Hilbert space are complex-valued functionals of the isotopy classes of one-dimensional submanifolds of the torus [45]:

$$\psi[\{\alpha\}] = \langle\{\alpha\}|\psi\rangle. \tag{57}$$

We called this the 'pre-Hilbert space' of our theory because the actual Hilbert space of the theory, $\mathscr{H}$, is a subspace of $\tilde{\mathscr{H}}$. We define $\mathscr{H}$ as the subspace consisting of $\Psi[\{\alpha\}]$s satisfying the following constraints:

$$\begin{aligned} \Psi[\{\alpha\}] &= -\Psi[\{\alpha \cup \bigcirc\}], \\ \Psi[\{\alpha\}] &= -\Psi[\{\tilde{\alpha}\}], \end{aligned} \tag{58}$$

$\alpha \cup \bigcirc$ is the one-dimensional submanifold of the torus which is obtained from the union of α with a contractible loop; $\tilde{\alpha}$ is obtained from α by performing the cutting and rejoining operation $)(\rightarrow \asymp$ on any part of α (see Fig. 2). Since this operation, which is called 'surgery,' can be performed on any part of α, the relation (58) must hold for all possible $\{\tilde{\alpha}\}$.

Said differently, the Hilbert space of our theory is the subspace of pre-Hilbert space which is annihilated by the two projection operators:

$$\begin{aligned} K_{-1} &= (|\{\alpha\}\rangle + |\{\alpha \cup \bigcirc\}\rangle)(\langle\{\alpha\}| + \langle\{\alpha \cup \bigcirc\}|), \\ P_{2,-1} &= (|\{\alpha\}\rangle + |\{\tilde{\alpha}\}\rangle)(\langle\{\alpha\}| + \langle\{\tilde{\alpha}\}|). \end{aligned} \tag{59}$$

As before, the first projector must be applied for all possible contractible loops and the second projector must be applied for all possible $\{\tilde{\alpha}\}$.

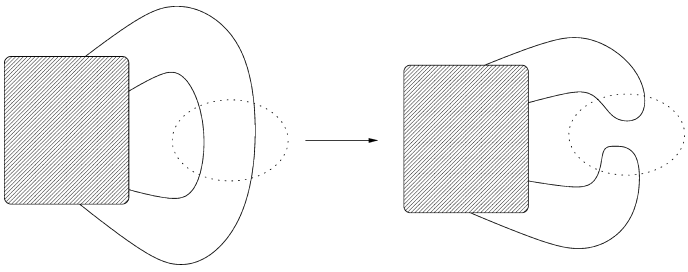

Fig. 2. A surgery operation of this type changes the wavefunction by a factor of −1. The shaded area represents an arbitrary multi-curve which completes the 1-manifold shown.

This Hilbert space is four-dimensional on the torus. To see this, observe that the operations $\alpha \to \alpha \cup \bigcirc$ and $)(\to \overset{\smile}{\frown}$ preserve modulo two the winding number of a multi-curve about the torus but do not preserve the winding number itself. There are four possible winding numbers about the meridian and longitude of the torus: (0,0), (1,0), (0,1), (1,1). We can define a function $\Psi_{(i,j)}$ which vanishes on all isotopy classes of one-dimensional submanifolds of the torus which do not have winding number (i,j) modulo two. We can assign $\Psi_{(i,j)}$ the value one for some isotopy class which has winding number (i,j) modulo two; it takes value ± 1 on all other isotopy classes which have winding number (i,j) modulo two, according to the relation (58). These four wavefunctions $\Psi_{(i,j)}$ form a basis of Hilbert space. A further point must be checked: that no 1-manifold α can be related back to itself by an odd number of surgeries. This would introduce the relation $\alpha = (-1)^{\text{odd}}\alpha$ or $\alpha = 0$. This fact is intuitively obvious, but a proof would take us on a topological digression, so we omit it here.

Thus, our Hilbert space is four-dimensional, as it must be if it is to be the same as the Hilbert space of $U(1)_2 \times \overline{U(1)_2}$ Chern–Simons theory, which we constructed earlier. To see that it, indeed, fulfills its raison d'être, namely to furnish a representation of the algebra (49) and (50) we define the operators $W^+[\gamma]$, $W^-[\gamma]$:

$$\begin{aligned} W^+[\gamma]\Psi[\{\alpha\}] &= \mathrm{i}^{n(\gamma,\alpha)}\Psi[\{\alpha \cup_R \gamma\}], \\ W^-[\gamma]\Psi[\{\alpha\}] &= \mathrm{i}^{n(\gamma,\alpha)}\Psi[\{\alpha \cup_L \gamma\}], \\ L[\gamma]\Psi[\{\alpha\}] &= (-1)^{I(\gamma,\alpha)}\Psi[\{\alpha\}], \end{aligned} \tag{60}$$

$n(\gamma,\alpha)$ is the number of intersections between γ and α (without regard to sign, unlike $I(\gamma,\alpha)$). The notation $\alpha \cup_R \gamma$ simply means the union of α and γ if they do not intersect; if they do, then the crossing is resolved by turning to the right as the intersection is approached along α, as shown in Fig. 3. $\alpha \cup_L \gamma$ is defined in the opposite way. It may not be obvious that $\alpha \cup_R \gamma$ depends only on the isotopy class of α. For instance, suppose that α and γ do not intersect. Then, α can be continuously deformed into isotopic α' such that α' intersects γ twice. Thus, if we act on $\Psi[\{\alpha\}]$ with $W^+[\{\gamma\}]$, we would seem to get a different factor, $i^{n(\gamma,\alpha)} = 1$ or $i^{n(\gamma,\alpha')} = -1$, depending on which representative of the isotopy class we choose. To make matters worse, $\alpha \cup_R \gamma$ and $\alpha' \cup_R \gamma$ are not in the same isotopy class. However, these two apparent problems cancel each other out, as a result of the second constraint in (58). (It's useful to draw a couple of pictures to see this.)

It is clear from the preceding considerations why the second constraint in (58) is necessary. The first constraint in (58) is necessitated by consistency with the second one, as may be seen from Fig. 4.

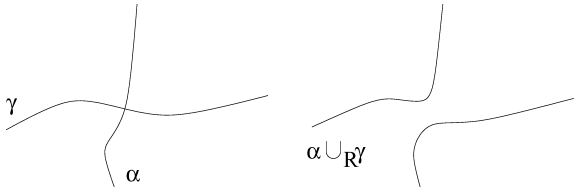

Fig. 3. When α and γ intersect, $\alpha \cup_R \gamma$ is defined as shown above.

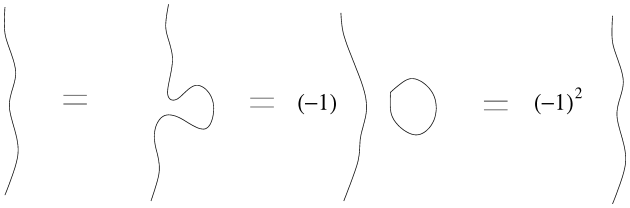

Fig. 4. The second constraint, which assigns a −1 to the operation of cutting and rejoining of two parallel strands requires that a contractible loop change the wavefunction by −1.

In defining the action of the Wilson loop operators, we ignored the fact that they are defined for parameterized curves $\gamma(s)$. This definition is manifestly invariant under orientation-preserving reparameterizations. Orientation-reversing reparameterizations simply conjugate the Wilson loop operator: the integral is done in the reverse direction. For the case $m = 2$, the Wilson loop operators have real eigenvalues 1, −1, so they are also invariant under orientation-reversal. Thus, we can be lazy and treat the γs as unoriented curves.

The inner product on Hilbert space is determined by the condition that $W^{\pm}[\gamma]$, $L[\gamma]$ be unitary. This can be accomplished by inheriting the obvious inner product from pre-Hilbert space or by simply taking $\langle \Psi_{(i,j)} | \Psi_{(i',j')} \rangle = \delta_{ii'}\delta_{jj'}$.

Thus, we have learned how to create a representation of the basic operator algebra (49) and (50) which defines $U(1)_2 \times \overline{U(1)}_2$ Chern–Simons theory in terms of simple operations on multi-curves, *so long as we impose a carefully chosen constraint structure on the Hilbert space.* Such a Hilbert space necessarily realizes the doubled theory $U(1)_2 \times \overline{U(1)}_2$ because it automatically admits a set of commuting operators $L[\gamma]$ as defined in the third line of (60). Such a set of operators is a feature of the doubled theory, according to (49), *but not of the undoubled theory* $U(1)_2$.

Note that the Chern–Simons constraint (34) is already implemented in the pre-Hilbert space by the condition that states depend only on the isotopy class of α and not on α itself. This structure will be common to all of the topological field theories which we will discuss. The additional structure (58) is special to $U(1)_2 \times \overline{U(1)}_2$ Chern–Simons theory, and it enables $\mathscr{H}$ to realize the canonical commutation relations of the theory. In more complex topological field theories, these relations must be generalized. It will turn out that this can only be done in a rather restricted set of ways. In Section 4, we will discuss this in detail.

First, let us carry a little further our analysis of the $U(1)_2 \times \overline{U(1)}_2$ Chern–Simons theory. Let us consider the theory on the plane, but with a quasiparticle located at the origin.

$$S = \frac{2}{4\pi}\int \mathrm{d}^2x\,\mathrm{d}t\epsilon^{\mu\nu\rho}a_\mu\partial_\nu a_\rho - \frac{2}{4\pi}\int \mathrm{d}^2x\,\mathrm{d}t\epsilon^{\mu\nu\rho}c_\mu\partial_\nu c_\rho + \int \mathrm{d}t(\rho_1 a_0(0,t) + \rho_2 c_0(0,t)). \tag{61}$$

We have coupled the gauge fields to a fixed, non-dynamical quasiparticle at the origin. The quasiparticle has charges $\rho_{1,2}$ under the two $U(1)$ gauge symmetries.

If $\rho_1 = \rho_2 = 0$, then there's no quasiparticle at the origin. If one of them is non-zero then there is a non-trivial quasiparticle at the origin. These four possibilities are the four (topologically) distinct types of quasiparticles allowed in the theory, including the trivial particle. The Chern–Simons constraints now read:

$$\begin{aligned} \nabla \times \mathbf{a} &= \pi\rho_1, \\ \nabla \times \mathbf{c} &= \pi\rho_2 \end{aligned} \tag{62}$$

or, equivalently,

$$\begin{aligned} W^+[\gamma] &= \mathrm{e}^{\mathrm{i}\pi\rho_1}, \\ W^-[\gamma] &= \mathrm{e}^{\mathrm{i}\pi\rho_2}. \end{aligned} \tag{63}$$

This can be incorporated into our representation of Hilbert space in terms of multi-curves if we enlarge our pre-Hilbert space to include multi-curves which terminate at the origin. In order to make this well-defined, we will also have to widen the origin into a finite-sized puncture in the plane. Let us also take the system to be on a finite disk, rather than the infinite plane. Thus, we must allow curves which are not closed but have two endpoints. The location of each such endpoint should be described by a quantum mechanical wavefunction $\psi(\theta)$ where θ is an angular coordinate on the inner or outer circle. Different ψs correspond to edge excitations, which we will discuss later. The endpoints of a curve may both be at the inner or outer boundary of the annulus or one may be at the inner boundary while the other is at the outer boundary. The former case corresponds to 'oscillator modes' of the edge (which are, in general, gapped); the latter, to different sectors (or different Verma modules) of the outer edge theory and different quasiparticle species at the origin. In the $U(1)_2 \times \overline{U(1)}_2$ Chern–Simons theory which we are now discussing, the surgery relation tells us that we can reduce any configuration to one in which there is no more than a single curve connecting the inner circle to the outer one. For simplicity of depiction, we will choose preferred points at the inner and outer boundaries and insist that curves which connect the two boundaries terminate at these points, as shown in Fig. 5. It does not matter which points we choose; different choices of preferred point are related to issues regarding boundary conditions and edge excitations, which we will discuss later. This is a particular choice of boundary condition. It is not natural physically, but it is convenient for now. Curves with two endpoints at the same boundary, or 'bigons,' will be neglected for now and discussed in the context of edge excitations.

From this construction, we see that the only allowed charges are $\rho_{1,2} = 0, 1$. A Hilbert space representation in terms of multi-curves on the annulus does not exist for any other value of $\rho_{1,2}$. This can be understood from a functional integral perspective by first noting from (63) that integer charges are only distinguished modulo 2. If ρ_1 is not an integer, then we can rescale a_μ to set $\rho_1 = 1$, thereby changing the coefficient of the Chern–Simons term. We really have a theory with a different (non-integer) coupling constant for one of the Chern–Simons gauge fields, and we shouldn't expect it to have a simple representation in terms of multi-curves.

Thus, we are essentially considering our system on an annulus. There are four basic pictures depicted in Fig. 6. The first pair have been combined into two linear

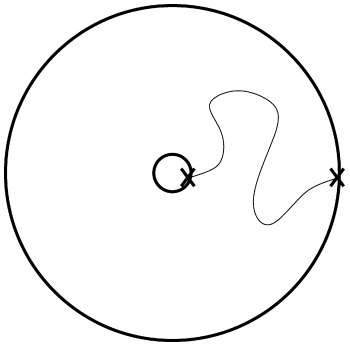

Fig. 5. Curves which terminate at the quasiparticle (i.e., the inner boundary) or the outer boundary of the annulus must do so at preferred points. The darker lines represent the boundary and the quasiparticle at the origin, while the lighter line represents a curve.

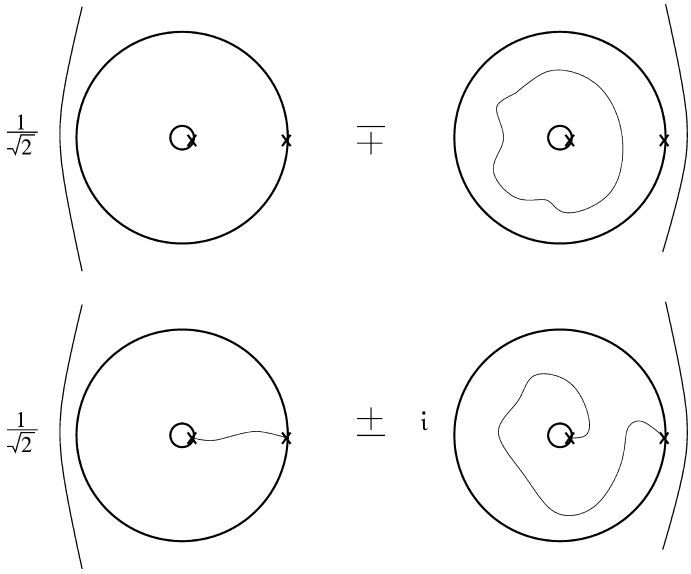

Fig. 6. The four species of quasiparticles (including the vacuum as the trivial quasiparticle) in $U(1)_2 \times \overline{U(1)_2}$ Chern–Simons theory.

combinations with relative coefficient ∓ 1; the second pair have been combined into two possible linear combinations with relative coefficient $\pm i$. All other one-dimensional submanifolds of the annulus (subject to the prescribed boundary conditions) can be obtained from these by applying the operations $\alpha \to \alpha \cup \bigcirc$ and $)(\to \smile$. Thus, we can, as we did in the case of the torus, take a basis of states which vanish on the isotopy classes of all but one of the pictures in Fig. 6. It necessarily also vanishes on all isotopy classes which can be obtained from these three by the operations $\alpha \to \alpha \cup \bigcirc$ and $)(\to \smile$. A given basis vector takes the value 1 on the isotopy class of one of the pictures and, therefore, takes the values ± 1 on the isotopy classes of those pictures and those which can be

obtained from it by the repeated use of the operations $\alpha \to \alpha \cup \bigcirc$ and $)(\to \asymp$. We will call the four states corresponding to the four pictures in Fig. 6, respectively, $|0,0\rangle$, $|1,1\rangle$, $|1,0\rangle$, $|0,1\rangle$. Using (60), we can show that $W^+[\gamma]$ has eigenvalues 1,−1,−1,1 on these four states, while $W^-[\gamma]$ has eigenvalues 1,−1,1,−1 if γ encircles the origin. $W^+[\gamma]$ and $W^-[\gamma]$ have eigenvalue 1 on all four states if γ does not encircle the origin. Hence, comparing with (63), we see that our labels can also be interpreted as $|\rho_1, \rho_2\rangle$.

These particular linear combinations have been chosen so that they are spin eigenstates (by 'spin,' we refer here to the eigenvalue under a rotation of the annulus, not to an internal quantum number, or, in more mathematical language, eigenstates under a Dehn twist). They also diagonalize all two-particle braids. According to the spin-statistics connection, this is automatic for braids of identical particles since we have taken spin eigenstates. We have furthermore chosen linear combinations such that a braid of two particles of two different particle species results in merely the acquisition of a phase, i.e., we have diagonalized two-particle braids. (The non-Abelian nature of the braid group can only be manifest when three or more particles are present.) As a result of these choices, our quasiparticles have the mathematical property of idempotence under 'stacking.' If we join annuli concentrically (or stack the topologically equivalent cylinders) then $(\rho_1, \rho_2) \circ (\rho_1', \rho_2') = (\rho_1, \rho_2)\delta_{\rho_1\rho_1'}\delta_{\rho_2\rho_2'}$, where $\circ$ denotes the stacking operation.

Imagine taking two of these quasiparticles, with charges (ρ_1, ρ_2) and (ρ_1', ρ_2'), very close together so that, viewed from a distance, they look like a single quasiparticle of charge (ρ_1^T, ρ_2^T). Following the graphical manipulations in Fig. 7, we find that $\rho_i^T = (\rho_i + \rho_i')\text{mod}2$. This process is called *fusion*.

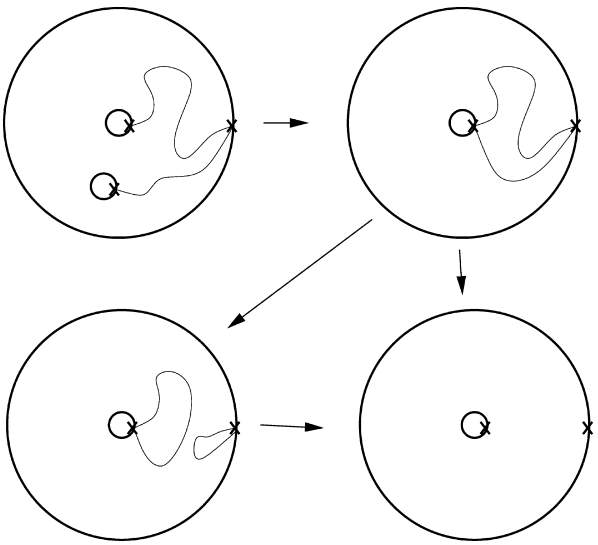

Fig. 7. The fusion of two particles. After the first step, the contractible loop can be shrunk directly or after applying the surgery operation. The same result is obtained either way (either 1 or 3 minus signs accrue). Strictly speaking, the particles are the superpositions of Fig. 6, so this picture must be superposed with three others; the result is the same.

We can also take one particle around another, as in Fig. 8. The effect of a counter-clockwise braid of two particles is the multiplication of the state by a phase $e^{2i\theta_{\rho_1\rho_1',\rho_2\rho_2'}} = e^{i\pi(\rho_1\rho_1' - \rho_2\rho_2')}$. Two identical particles can be exchanged counter-clockwise; the resulting phase is $e^{i\theta_{\rho_1\rho_1,\rho_2\rho_2}} = e^{(i\pi/2)(\rho_1^2 - \rho_2^2)} = e^{(i\pi/2)(\rho_1 - \rho_2)}$. An equivalent way of arriving at this result is via the spin-statistics theorem: we can imagine fusing the two particles first and then simply rotating the resulting particle by π. Under a rotation, a quasiparticle of charge (ρ_1^T, ρ_2^T) acquires a phase $e^{\pi i S} = e^{(i\pi/4)(\rho_1^T - \rho_2^T)}$ which is the same phase which we would obtain if we exchanged them first and then fused them.

To summarize, we can set up the Hilbert space not only of pure $U(1)_2 \times \overline{U(1)_2}$ Chern–Simons theory, but also of the theory with static quasiparticles in a pictorial representation.

The simplest generalizations of the preceeding are the other Abelian doubled Chern–Simons theories, $U(1)_m \times \overline{U(1)_m}$. These also have a pictorial representation, but it involves directed multi-curves and m-valent vertices at which they can terminate. We will not discuss them further here, but they are a straightforward generalization of the case $m = 2$.

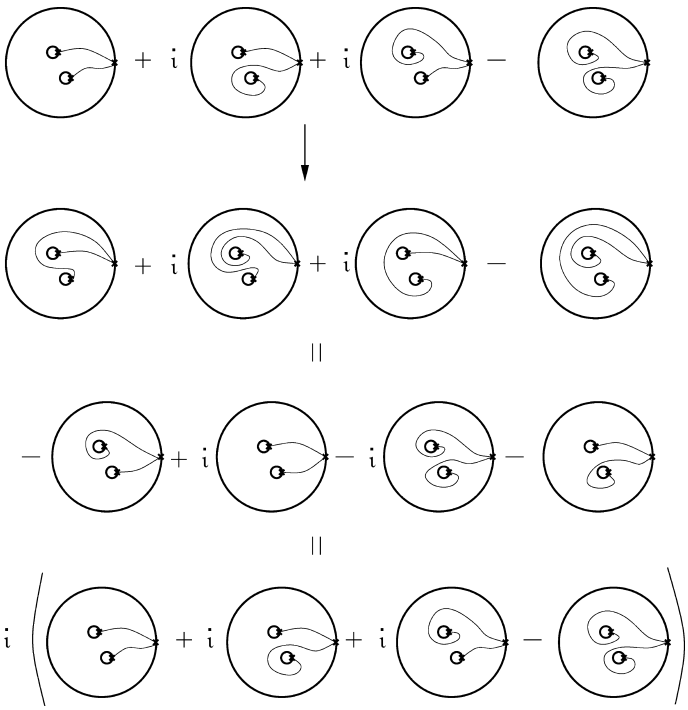

Fig. 8. The result of braiding two (1,1) quasiparticles. (1,1) quasiparticles are the type which are displayed in the lower part of Fig. 6, with the plus sign. They involve a superposition of two pictures, hence two quasiparticles involve a superposition of four pictures. A clockwise interchange (second panel) leads, after surgery (third panel), to a phase $\pi/2$.

4.3. Z_2 Gauge theory

As we mentioned in the introduction, discrete gauge theories also have topological phases. Consider the simplest, Z_2 gauge theory. There are two different ways of realizing such a theory. We could begin with a $U(1)$ gauge theory with Maxwell action which is coupled to a charge-2 matter field. When this matter field condenses, the $U(1)$ symmetry is broken to Z_2. This construction can be done directly in the continuum. Alternatively, one can work with Z_2 gauge fields from the beginning. However, one must, in such a case, work on a lattice. Let us follow the latter avenue. We consider a 2+1 dimensional space-time lattice on which there is an Ising gauge field degree of freedom $\sigma_z = \pm 1$ on each *link* of the lattice. We will label them by a lattice site, $\mathbf{x}$, and a direction $i = x, y, \tau$ so that there are three links associated with each site. The action is the sum over all plaquettes of the product of σ_zs around a plaquette:

$$S = -K \sum_{\text{plaq.}} \sigma_z \sigma_z \sigma_z \sigma_z. \tag{64}$$

To quantize this theory, it is useful to choose Coulomb gauge, $\sigma_z(\mathbf{x}, \tau) = 1$ for all $\mathbf{x}$. In this gauge, the Hamiltonian takes the form:

$$H = -\sum_{\mathbf{x},i} \sigma_x(\mathbf{x}, i) - K \sum_{\text{spatialplaq.}} \sigma_z \sigma_z \sigma_z \sigma_z. \tag{65}$$

In Coulomb gauge, there are residual global symmetries generated by the operators

$$G(\mathbf{x}) = \sigma_x(\mathbf{x}, x)\sigma_x(\mathbf{x}, y)\sigma_x(\mathbf{x} - \hat{\mathbf{x}}, x)\sigma_x(\mathbf{x} - \hat{\mathbf{y}}, y). \tag{66}$$

The extreme low-energy limit, in which this theory becomes topological, is the $K \to \infty$ limit. In this limit, $\sigma_z \sigma_z \sigma_z \sigma_z = 1$ for every spatial plaquette, recovering Kitaev's 'toric code' [35].

It is useful to define operators $W[\gamma]$ associated with closed curves γ on the lattice:

$$L[\gamma] = \prod_{\mathbf{x},i \in \gamma} \sigma_z(\mathbf{x}, i). \tag{67}$$

We also need operators $Y[\alpha]$ associated with closed curves on the dual lattice, i.e., closed curves which pass through the centers of a sequence of adjacent plaquettes.

$$Y[\alpha] = \prod_{\mathbf{x},i \perp \alpha} \sigma_x(\mathbf{x}, i). \tag{68}$$

The product is over all links which α intersects. $L[\gamma]$ is analogous to a Wilson loop operator while $Y[\gamma]$ creates a Dirac string.

Let us consider the space of states which are annihilated by the Hamiltonian; this is the Hilbert space of the $K \to \infty$ limit. When restricted to states within this Hilbert space, $L[\gamma]$ and $Y[\alpha]$ satisfy the operator algebra

$$\begin{aligned} &L[\gamma]Y[\alpha] = (-1)^{I(\gamma,\alpha)} Y[\alpha]L[\gamma], \\ &[L[\gamma], L[\alpha]] = [Y[\gamma], Y[\alpha]] = 0. \end{aligned} \tag{69}$$

Now, it is clear that such an operator algebra can be represented on a vector space which is very similar to the Hilbert space of $U(1)_2 \times \overline{U(1)}_2$ Chern–Simons theory:

$$\begin{aligned} L[\gamma]\Psi[\{\alpha\}] &= (-1)^{I(\gamma,\alpha)}\Psi[\{\alpha\}], \\ Y[\gamma]\Psi[\{\alpha\}] &= \Psi[\{\alpha \cup \gamma\}]. \end{aligned} \tag{70}$$

The notable difference is that the allowed states must now satisfy the constraints

$$\begin{aligned} \Psi[\{\alpha\}] &= \Psi[\{\alpha \cup \bigcirc\}], \\ \Psi[\{\alpha\}] &= \Psi[\{\tilde{\alpha}\}]. \end{aligned} \tag{71}$$

Again, $\tilde{\alpha}$ is obtained from α by performing the surgery operation $)(\rightarrow \smile$ on any part of α.

If α is contractible, then $Y[\alpha]$ commutes with all other operators in the theory, so its effect on any wavefunction should be multiplication by a scalar. If we take this scalar to be 1, then we have the first constraint above. The second constraint is necessary in order to realize the operator algebra (70) and is also required by consistency with the first. As a result of the second line of (71), $\alpha \cup \gamma$ in (70) can be either $\alpha \cup_R \gamma$ or $\alpha \cup_L \gamma$ since they are equivalent in the low-energy Hilbert space.

Again, we can characterize Hilbert space as the space of states annihilated by two projection operators,

$$\begin{aligned} K_1 &= (|\{\alpha\}\rangle - |\{\alpha \cup \bigcirc\}\rangle)(\langle\{\alpha\}| - \langle\{\alpha \cup \bigcirc\}|) \\ P_{2,1} &= (|\{\alpha\}\rangle - |\{\tilde{\alpha}\}\rangle)(\langle\{\alpha\}| - \langle\{\tilde{\alpha}\}|). \end{aligned} \tag{72}$$

As a shorthand, we will summarize such relations in the manner shown in Fig. 9

Again, the Hilbert space on the torus is four-dimensional and there are four quasiparticle species corresponding to them. The corresponding pictures are the same as in $U(1)_2 \times \overline{U(1)}_2$ Chern–Simons theory, namely the four pictures in Fig. 6. However, as a result of the $+$ sign in the second line of (71), the coefficients are different. $\mp$ in the top panel of Fig. 6 becomes $\pm$, while $\pm i$ in the second panel becomes simply $\pm$. The fusion rule for quasiparticles (ρ_1, ρ_2) and (ρ_1', ρ_2') is still $\rho_i^T = (\rho_i + \rho_i')\text{mod}2$.

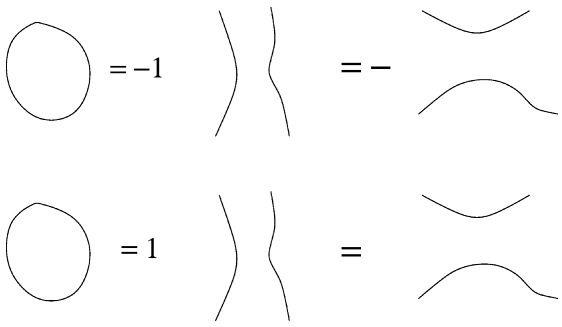

Fig. 9. As a shorthand, we will denote the constraints satisfied by the physical Hilbert space as shown. In the top panel, we have the relations of $U(1)_2 \times \overline{U(1)}_2$ Chern–Simons theory, while in the bottom panel, we have the relations of Z_2 gauge theory.

However, these quasiparticles have spins $e^{2\pi i S} = e^{i\pi\rho_1\rho_2}$. A counter-clockwise braid of two quasiparticles yields a phase $e^{i\pi(\rho_1\rho_2'+\rho_1'\rho_2)}$.

There is a natural generalization of this construction to discrete G gauge theories. In these theories, the variables are elements g_i of the discrete group G on each link i of the lattice. The action is then of the form

$$S = -K\sum_{\text{plaq.}} \mathrm{Tr}(g_1 g_2 g_3 g_4). \tag{73}$$

We will not discuss these theories further here, but they, too, have a representation similar to the ones we have constructed, but with some extra features (e.g., directed curves) of the type present in the $U(1)_m \times \overline{U(1)}_m$ Chern–Simons theories.

4.4. Some comments on doubled theories

The two Hilbert spaces which we have just constructed are representation spaces for the underlying Wilson loop algebras of the $U(1)_2 \times \overline{U(1)_2}$ Chern–Simons theory,

$$\begin{aligned} &L[\gamma]W^+[\gamma'] = (-1)^{I(\gamma,\gamma')}W^+[\gamma']L[\gamma],\\ &L[\gamma]L[\alpha] = L[\alpha]L[\gamma],\\ &W^+[\gamma]W^+[\alpha] = (-1)^{I(\gamma,\gamma')}W^+[\alpha]W^+[\gamma]. \end{aligned} \tag{74}$$

(since W^- can be recovered from L and W^+, they form a complete set) and the Z_2 gauge theory,

$$\begin{aligned} &L[\gamma]Y[\alpha] = (-1)^{I(\gamma,\alpha)}Y[\alpha]L[\gamma],\\ &L[\gamma]L[\alpha] = L[\alpha]L[\gamma],\\ &Y[\gamma]Y[\alpha] = Y[\alpha]Y[\gamma]. \end{aligned} \tag{75}$$

As we noted earlier, it is tempting to think of the $L[\gamma]$s as 'number' operators and the $W^+[\gamma]$s or $Y[\gamma]$s as being somewhat similar to creation/annihilation operators. Indeed, $L[\gamma]$ essentially counts how much flux is enclosed by the curve γ, while $W^+[\gamma]$ and $Y[\gamma]$ can increase or decrease this flux. The only difference between the two theories is in the third lines of (74) and (75): the $Y[\gamma]$s commute with each other while the $W^+[\gamma]$s do not.

This leads to rather significant differences between the two theories although both have degeneracy 4 on the torus. The $U(1)_2 \times \overline{U(1)_2}$ theory has in its spectrum of quasiparticles two species which are semions and anti-semions. Z_2 gauge theory has no quasiparticles with non-trivial self-statistics. The only non-trivial statistics are off-diagonal statistics between different particle types. Mathematically, these differences follow from the simple fact that the composition algebra for Wilson loops on T^2 in the Z_2 gauge theory is commutative and, therefore, is equal to $\oplus_{i=1}^4 C$; in the $U(1)_2 \times \overline{U(1)_2}$ theory, it is is isomorphic to M_2, the algebra of 2×2 matrices. (Any finite-dimensional C^* algebra must be a direct sum matrix algebras [46], so these are the only possibilities.)

In fact, both of these theories are 'doubled' theories. $U(1)_2 \times \overline{U(1)_2}$ Chern–Simons theory, in the naive way: it is the tensor square of $U(1)_2$ Chern–Simons theory. Z_2 gauge theory is an example of a non-trivial double. From a formal perspec-

tive, it is the double of a theory with a single gauge field with vanishing Chern–Simons coefficient. Such a theory is gapless, and might represent a superfluid.

We will proceed momentarily to a discussion of the $SU(2)_k \times \overline{SU(2)_k}$ Chern–Simons theories, but it is worth pausing for a moment to note some of the general features of the theories which we have just constructed since these features will reappear in a slightly more complicated guise. We constructed our theories by finding a representation for the algebra of Wilson loop operators. We found that this could be done on a subspace of the vector space of functionals of 1-manifolds, a subspace which was selected by requiring that allowed wavefunctionals assign the value $d = \pm 1$ to a contractible loop and be invariant under a surgery procedure on 1-manifolds. The doubled $SU(2)_k$ Chern–Simons theories theories have different values of d and surgery procedures.

5. Solution of doubled $SU(2)_k$ Chern–Simons theories

5.1. Wilson loop algebra

We now turn to our main concern in this paper, the $SU(2)_k \times \overline{SU(2)_k}$ Chern–Simons theories. Our aim is to construct the Hilbert spaces of these theories in a representation similar to the ones which we just used for $U(1)_2 \times \overline{U(1)_2}$ Chern–Simons theory and Z_2 gauge theory.

The action of $SU(2)_k$ Chern–Simons theory is

$$S = \frac{k}{4\pi} \int \epsilon^{\mu\nu\rho} \left(a^{\underline{a}}_{\mu} \partial_{\nu} a^{\underline{a}}_{\rho} + \frac{2}{3} f_{\underline{abc}} a^{\underline{a}}_{\mu} a^{\underline{b}}_{\nu} a^{\underline{c}}_{\rho} \right) = \frac{k}{4\pi} \int \mathrm{tr} \left(a \wedge \mathrm{d}a + \frac{2}{3} a \wedge a \wedge a \right), \quad (76)$$

$a^{\underline{a}}_{\mu}$ is a gauge field taking values in the Lie algebra $su(2)$, $a^{\underline{a}}_{\mu} T^a$, with index $a = 1, 2, 3$ running over the generators T^1, T^2, T^3 of $su(2)$. In the second line, we have rewritten the action in the more compact language of differential forms. The integer, k, is the coupling constant.

In constructing physical observables, we must exercise a little more care than in the Abelian case because the gauge fields at different points will not commute, $[a^a_{\mu}(x_1) T^a, a^b_{\mu}(x_2) T^b] \neq 0$. Thus, the exponential integral must be path-ordered

$$U[\gamma] \equiv \mathscr{P} \mathrm{e}^{\oint_{\gamma} \mathbf{a}^c T^c \cdot \mathrm{d}\mathbf{l}} = \sum_{n=0}^{\infty} \int_0^{2\pi} \mathrm{d}s_1 \int_0^{s_1} \mathrm{d}s_2 \ldots \times \int_0^{s_{n-1}} \mathrm{d}s_n \left[\dot{\gamma}(s_1) \cdot \mathbf{a}^{\underline{a}_1}(\gamma(s_1)) T^{a_1} \ldots \dot{\gamma}(s_n) \cdot \mathbf{a}^{\underline{a}_n}(\gamma(s_n)) T^{a_n} \right], \quad (77)$$

where $\gamma(s)$, $s \in [0, 2\pi]$ is an arbitrary parameterization of the curve γ. This quantity is an $SU(2)$ matrix, which transforms in the adjoint representation of $SU(2)$ at the starting point $\gamma(0)$. To get a gauge-invariant quantity, we must take the trace:

$$W[\gamma] = \mathrm{tr}(U[\gamma]). \quad (78)$$

As in the $U(1)_2 \times \overline{U(1)_2}$ case, the Wilson loop operators in this theory are, strictly speaking, defined for parameterized curves $\gamma(s)$. It is manifestly invariant under

orientation-preserving reparameterizations. However unlike in the $U(1)_2 \times \overline{U(1)_2}$ case, it is not quite true here that $W[\gamma]$ is real, so the orientation of γ cannot be treated so cavalierly. Fortunately, the spin-1/2 representation of $SU(2)$ is pseudo-real, meaning that it is equal to its conjugate representation upon multiplication on the left and right by the antisymmetric tensor $\epsilon_{\alpha\beta}$. These factors disappear upon taking the trace, so we can, again, ignore the orientation of γ, i.e., if we orient γ and denote by γ^{-1} the same curve followed in the opposite direction then

$$W[\gamma^{-1}] = \mathrm{tr}(U[\gamma]^\dagger) = \mathrm{tr}(\epsilon U[\gamma]\epsilon) = \mathrm{tr}(U[\gamma]) = W[\gamma]. \tag{79}$$

Proceeding in parallel with the Abelian case, we derive the equal-time commutations relations from the temporal gauge form of the action (76)

$$\left[a_1^a, a_2^b\right] = \mathrm{i}\frac{2\pi}{k}\delta^{\underline{ab}}. \tag{80}$$

In the spin-1/2 representation,

$$\left[a_1^a \tau_{AB}^a, a_2^b \tau_{CD}^b\right] = \mathrm{i}\frac{2\pi}{k}\delta^{\underline{ab}}\tau_{AB}^a\tau_{CD}^b = \mathrm{i}\frac{2\pi}{k}\cdot 3\cdot[\delta_{AD}\delta_{BC} - \delta_{AC}\delta_{BD}]. \tag{81}$$

We must now be a little more careful about issues of gauge-invariance. The above commutation relations hold in a field theory which contains 'too many' degrees of freedom, most of which are pure gauge. Continuing in parallel with our discussion of the Abelian case, we will work entirely with gauge-invariant Wilson loop operators. In so doing, we are eliminating the pure gauge degrees of freedom. In the Abelian case, this is trivial since the gauge transformation is simply $\mathbf{a} \to \mathbf{a} - \mathbf{d}f$. In the non-Abelian case, it is $\mathbf{a} \to g\mathbf{a}g^{-1} - \mathbf{d}gg^{-1}$, so there is a non-trivial Jacobian which results upon eliminating the gauge degrees of freedom. As a result of this Jacobian, the commutation relations of the reduced theory are modified by the shift $k \to k+2$, as shown in [25]. (The alternative possibility is to work with the full set of degrees of freedom of the theory and then require that the Hilbert space contain only gauge-invariant states. In such a case, the inner product will be non-trivial and will lead to this 'quantum correction.') We will see how this shift arises from a different perspective in Section 6, so we defer a discussion until then.

We work in this reduced theory, containing only gauge-invariant degrees of freedom. Hence, we make this shift when we compute the commutator of two Wilson loop operators, which is:

$$[W[\gamma], W[\gamma']] = 2\sin\left(\frac{\pi}{2(k+2)}\right)\sum_i \left(W[\gamma\circ_i\gamma'] - W[\gamma'\circ_i\gamma]\right). \tag{82}$$

The summation is over all intersections i of the curves γ and γ'. In this expression, $\gamma\circ_i\gamma'$ is the curve obtained by starting at intersection i, following γ' until it returns to i, and then taking γ. In so doing, we temporarily introduce orientations for γ and γ' so that their tangent vectors form a right-handed dyad at their intersection point. Alternatively, we could simply say that as the intersection is approached along γ, we turn to the left to join to γ'. This construction is closely related to the Goldman bracket [47]. Strictly speaking, we should have a factor of $(\gamma,\alpha)_i$ multiplying $\gamma\circ_i\gamma'$,

where $(\gamma,\alpha)_i$ is the sign of the ith intersection between γ and α. However, since the orientation is unimportant for closed $SU(2)$ Wilson loops, as we just noted, we can drop this factor. However, for loops terminating at boundaries or for other Lie groups, one must retain this factor.

In the doubled theory, we have two mutually commuting sets of Wilson loop operators, $W_{\pm}[\gamma]$, which satisfy the algebra

$$\left[W_+[\gamma], W_+[\gamma']\right] = (A - A^{-1}) \sum_i \left(W_+[\gamma \circ_i \gamma'] - W_+[\gamma' \circ_i \gamma]\right), \tag{83}$$

$$[W_-[\gamma], W_-[\gamma']] = -(A - A^{-1}) \sum_i \left(W_-[\gamma \circ_i \gamma'] - W_-[\gamma' \circ_i \gamma]\right), \tag{84}$$

$$[W_+[\gamma], W_-[\gamma']] = 0, \tag{85}$$

where we have introduced $A = \mathrm{i}\exp(\pi\mathrm{i}/2(k+2))$. Note that if γ and γ' have multiple intersections then $\gamma \circ_i \gamma'$ will be a self-intersecting curve.

To construct a Hilbert space which furnishes a representation of this algebra, we begin with a structure similar to that of the Hilbert spaces which we derived earlier: it is a subspace of the vector space of functionals of isotopy classes of loops which satisfy

$$\Psi[\{\alpha \cup \bigcirc\}] = \mathrm{d}\Psi[\{\alpha\}], \tag{86}$$

where d is a constant to be determined. The second condition in (58) is replaced by a more complicated one which we will construct later. For the following discussion, it is useful to introduce the notion of d-isotopy. Two multi-curves are related by d-isotopy if one can be deformed into the other by a combination of isotopy and the elimination of contractible loops. The latter operation results in a factor of d multiplying the wavefunctional.

We will find it notationally convenient to define $\Psi[\{\alpha\}]$ for intersecting loops so long as intersections are resolved as over- or under-crossings. In other words, we allow the 1-manifold α to no longer be embeddable in a surface. We will relate in the following way $\Psi[\cdot]$ evaluated on α with crossings to its values on smooth 1-manifolds embedded in a surface. We define:

$$\Psi[\{\alpha\}] = A\Psi[\{\alpha'\}] + A^{-1}\Psi[\{\alpha''\}] \tag{87}$$

where α' and α'' are the two ways of resolving the intersection, as depicted in Fig. 10. This relation can be applied repeatedly to remove all crossings in α, We could have

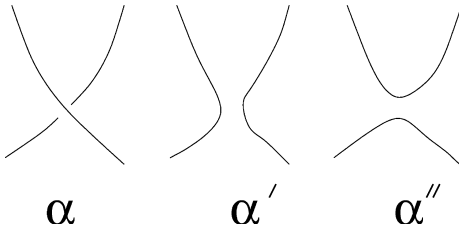

Fig. 10. α' and α'' are the two ways of resolving the crossing in α.

implemented this decomposition into resolutions in the definitions of $W_{\pm}[\gamma]$, but it is a little simpler this way.

The action of the Wilson loop operators is given by

$$\begin{aligned} W_{+}[\{\gamma\}]\Psi[\{\alpha\}] &= \Psi[\{\alpha\bigstar\gamma\}], \\ W_{-}[\{\gamma\}]\Psi[\{\alpha\}] &= \Psi[\{\gamma\bigstar\alpha\}]. \end{aligned} \tag{88}$$

The operation $\bigstar$ is defined as follows: if α and γ do not intersect then it is simply the union $\alpha \cup \gamma$. However, all intersections between α and γ are resolved by specifying that γ always crosses over α in $\alpha\bigstar\gamma$; in $\gamma\bigstar\alpha$, α always crosses over γ.

We also need to define $W_{+}[\{\gamma\}]$ for self-intersecting curves such as $\gamma \circ_i \gamma'$. A natural definition is to start at a given point on the curve, follow the curve in a given direction, and rule that 'later' sections of the curve always cross over 'earlier' sections. In other words, we parameterize the curve as $\gamma(t)$, with $t \in [0, 2\pi]$ and $\gamma(0) = \gamma(2\pi)$. We say that if $\gamma(t) = \gamma(t')$, then $\gamma(t')$ crosses over $\gamma(t)$ if $t' > t$. However, this depends on both the starting point and the direction. Hence, we average over all possible starting points and both directions. Thus, $W_{+}[\{\gamma\}]$ for self-intersecting γ is a normalized sum $\frac{1}{n}\sum_m W_{+}[\{\gamma_m\}]$ where the γ_m have intersections specified as over-crossings or under-crossings. Note that this is not a sum over all possible ways of choosing the intersections to be over- or under-crossings. If there are n intersections, then there are n different γ_m which can result in this way, not 2^n.

Before verifying that the operators defined above satisfy the desired commutation relations, let us first make sure that they are well-defined. Since $\Psi[\{\alpha\}]$ only depends on the isotopy class of α, $W_{\pm}[\{\gamma\}]$ should only depend on the isotopy class of γ. Thus, $\alpha\bigstar\gamma$ must be invariant under a continuous deformation of γ into γ' which has two new intersections with α, as shown in Fig. 11. Applying the definitions (88), we see that $\alpha\bigstar\gamma$ is invariant under isotopy moves of α and γ if:

$$d = -A^2 - A^{-2} = 2\cos\left(\frac{\pi}{k+2}\right). \tag{89}$$

The isotopy invariance of $W_{\pm}[\{\gamma\}]\Psi[\{\alpha\}]$ follows from its close relation to the *Kauffman bracket* [48] of the 1-manifold obtained by overlaying γ on α (or the reverse, in the case of W_{-}). The Kauffman bracket is defined for multi-curves in R^3 by projecting them to the plane, but keeping track of over-crossing and undercrossings. The crossings are resolved according to the rule (87) and all unknotted loops are accorded a factor of d. The only different feature in our Hilbert space is that the loops are

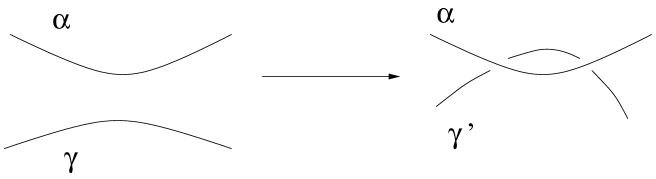

Fig. 11. $\alpha\bigstar\gamma$ must be defined so that $\alpha\bigstar\gamma = \alpha\bigstar\gamma'$.

actually embedded in some surface so that some unknotted loops are not contractible in that surface.

We now verify that $W_{\pm}[\{\gamma\}]$ as defined in (88) obey the commutation relations (83). First, note that $W_{+}[\{\gamma\}]$ and $W_{-}[\{\gamma'\}]$ commute trivially because

$$W_{+}[\{\gamma\}]W_{-}[\{\gamma'\}]\Psi[\{\alpha\}] = \Psi[\{\gamma'\bigstar\alpha\bigstar\gamma\}] = W_{-}[\{\gamma'\}]W_{+}[\{\gamma\}]\Psi[\{\alpha\}]. \tag{90}$$

Now consider the commutation relation between $W_{+}[\{\gamma\}]$ and $W_{+}[\{\gamma'\}]$ (the situation for $W_{-}[\{\gamma\}]$ and $W_{-}[\{\gamma'\}]$ is so similar that we need only discuss W_{+}). If there are no intersections between γ and γ', then both the left- and right-hand-sides of (83) vanish. Suppose γ and γ' have a single intersection k. The left-hand-side of (83) acting on a state $\Psi[\{\alpha\}]$ is

$$\begin{aligned}
\mathrm{LHS} &= \Psi[\{\alpha\bigstar\gamma'\bigstar\gamma\}] - \Psi[\{\alpha\bigstar\gamma\bigstar\gamma'\}] \\
&= A\Psi[\{\alpha\bigstar(\gamma'\circ_k\gamma)\}] + A^{-1}\Psi[\{\alpha\bigstar(\gamma\circ_k\gamma')\}] - A\Psi[\{\alpha\bigstar(\gamma\circ_k\gamma')\}] \\
&\quad - A^{-1}\Psi[\{\alpha\bigstar(\gamma'\circ_k\gamma)\}] \\
&= \left(A - A^{-1}\right)\left(\Psi[\{\alpha\bigstar(\gamma'\circ_k\gamma)\}] - \Psi[\{\alpha\bigstar(\gamma\circ_k\gamma')\}]\right) \\
&= \mathrm{RHS}.
\end{aligned} \tag{91}$$

Consider now the case in which γ and γ' have two intersections. In order to analyze this, it is useful to symbolically denote the two resolutions of the first crossing by x_1, y_1 and the two resolutions of the second crossing by x_2, y_2. Then the left-hand-side of (83) is:

$$\begin{aligned}
\mathrm{LHS} &= \left(Ax_1 + A^{-1}y_1\right)\left(Ax_2 + A^{-1}y_2\right) - \left(Ay_1 + A^{-1}x_1\right)\left(Ay_2 + A^{-1}x_2\right) \\
&= \left(A^2 - A^{-2}\right)(x_1x_2 - y_1y_2).
\end{aligned} \tag{92}$$

Meanwhile, the right-hand-side is:

$$\begin{aligned}
\mathrm{RHS} &= (A - A^{-1})\left(x_1\left(A + A^{-1}\right)(x_2 + y_2)/2 - y_1(A + A^{-1})(x_2 + y_2)/2\right) \\
&= (A^2 - A^{-2})(x_1x_2 - y_1y_2)
\end{aligned} \tag{93}$$

so (83) is satisfied.

If γ crosses over γ' at both intersections (or vice versa), then we can deform either one so that there is no intersection. Thus the equality which we have shown is trivial: both sides of the equation vanish because the corresponding Wilson loop operators commute, as we discussed earlier. If γ crosses over γ' at one intersection and under it at the other, then the commutator will be non-trivial, but still satisfies (83), as we have just seen. The general case of arbitrary γ, γ' is similar.

As in the Abelian case, we have found an infinite-dimensional vector space on which we can represent the commutator algebra of our theory. As in that case, we must truncate this vector space because the classical phase space has finite volume. Consider the torus. The holonomy $U_{+}[\gamma_1]$ of the a_{+} gauge field about the meridian, γ_1, of the torus will be some $SU(2)$ rotation. Since the homotopy group of the torus is Abelian, the holonomy about the longitude of the torus $U_{+}[\gamma_2]$ must commute with $U_{+}[\gamma_1]$, i.e., $[U_{+}[\gamma_1], U_{+}[\gamma_2]] = 0$. Hence, we must specify two $SU(2)$ rotations about the same axis. The direction of this axis is not invariant under an $SU(2)$ gauge

transformation, so we need only specify two angles. The same is true for the a_- gauge field. Hence, the phase space of the theory is the product of two tori, $T^2 \times T^2$.

Since the phase space of the classical theory has finite volume, its Hilbert space is finite-dimensional. In order to truncate our Hilbert space to a finite-dimensional one, we must specify surgery relations which must be obeyed by states in Hilbert space, analogous to the surgery relation of the $d = -1$ theory shown in the lower panel of Fig. 9. For the level k theory, there is only one possible relation which is consistent with the corresponding value of d, the amplitude associated with a contractible loop. The value of d so strictly constrains the Hilbert space of our theory that we have no freedom at all in our choice of a surgery relation—the analogue of $)(= -\breve{\frown}$. If we introduce no constraint, the Hilbert space on the torus (or any higher genus surface) will be infinite-dimensional. If we introduce a constraint which is too severe for the given d, such as $)(= \breve{\frown}$ for $d = \sqrt{2}$, then the Hilbert space (on any surface) will be zero-dimensional since there won't be any wavefunctions which satisfy both constraints. For a given d, there is a unique constraint which is just right—neither too trivial, nor too severe—so that there are finite-dimensional Hilbert spaces associated with different surfaces. In fact, for arbitrary d, there are no such non-trivial constraints, so the Hilbert space must be either infinite-dimensional or zero-dimensional. Only for – surprise, surprise—the very same sequence of ds which we have found in this section in 89, $d = 2\cos(\pi/(k+2))$, are there non-trivial constraints which lead to finite-dimensional Hilbert spaces. In Section 5.5, we will find these constraints.

As we will see in that section, such surgery relations depend strongly on d. We found above that d is determined by the condition that states and operators be invariant under isotopy which, in turn, follows from the Chern–Simons constraint which requires that the connection be flat. However, it is useful to consider another perspective on how d is determined in $SU(2)_k \times \overline{SU(2)_k}$ Chern–Simons theory. In order to do this, it will be useful to have an inner product on our Hilbert space. We will initially define this inner product on our infinite-dimensional Hilbert spaces, but we will later show that its restriction to the finite-dimensional Hilbert spaces constructed in Section 5.5 is the correct inner product of our theory.

5.2. Inner product

The infinite-dimensional pre-Hilbert spaces of the previous section have a natural inner product

$$\langle\{\alpha'\}|\{\alpha\}\rangle = \begin{cases} 1 & \text{if } \alpha' \cong \alpha, \\ 0 & \text{otherwise,} \end{cases} \tag{94}$$

where $\cong$ denotes equivalence under isotopy. The pleasant surprise about this inner product is that $W_\pm[\gamma]$ are Hermitian with respect to it

$$\begin{aligned} \langle\{\alpha'\}|W_+[\gamma]|\{\alpha\}\rangle &= \langle\{\alpha'\}|\{\alpha \bigstar \gamma\}\rangle \\ &= (\langle\{\alpha\}|\{\alpha' \bigstar \gamma\}\rangle)^* \\ &= (\langle\{\alpha\}|W_+[\gamma]|\{\alpha'\}\rangle)^*. \end{aligned} \tag{95}$$

The second equality can be understood by noting that the inner product vanishes unless $\alpha' \equiv \alpha \cup_R \gamma$ or $\alpha' \equiv \alpha \cup_L \gamma$. However, $\alpha' \equiv \alpha \cup_R \gamma$ implies that $\alpha \equiv \alpha' \cup_L \gamma$, and vice versa. The equality follows since one is accompanied by a factor of A; the other, a factor of A^{-1}.

Classically, $W_{\pm}[\gamma]$ are real because the spin-1/2 representation of $SU(2)$ is pseudo-real, as we noted in the previous section. Thus, it is natural to demand that they be Hermitian operators in the quantum theory and, indeed, we could take this requirement as the defining condition of our inner product. Fortunately, as we have just seen, this leads to precisely the same inner product as the natural one on pre-Hilbert space. Eventually, we will restrict this inner product to the finite-dimensional subspace which will form the Hilbert space of our theory.

5.3. Accidental symmetry and some coincidences

Before giving any detailed calculations, some low-level 'coincidences' should be brought to light so that they do not cause confusion later. These are essentially of the same form as the 'accidental' $SU(2)$ symmetry of a free boson at the self-dual compactification radius $R = 1/\sqrt{2}$ which underlies the Abelian bosonization of an $SU(2)$ doublet of fermions. A free chiral boson φ has the conserved current $\mathrm{i}\partial\varphi$. The existence of this current endows it with a $U(1)$ Kac–Moody algebra. The theory has central charge $c = 1$ (with respect to the Virasoro algebra) for any compactification radius R, $\varphi \equiv \varphi + 2\pi R$, so the description in terms of a $U(1)$ Kac–Moody algebra is perfectly acceptable. However, at radius $R = 1/\sqrt{2}$, there are additional dimension-1 fields $\mathrm{e}^{\pm \mathrm{i}\varphi\sqrt{2}}$—currents—which are allowed by the angular identification. These three currents form an $SU(2)_1$ Kac–Moody algebra. Thus, at this special radius, we can describe the theory equally well in terms of an $SU(2)_1$ Kac–Moody algebra, which also has central charge 1 with respect to its enveloping Virasoro algebra.

Let us now turn to the Abelian theories discussed in previous sections, (A) $U(1)_2 \times \overline{U(1)_2}$ and (B) Z_2 gauge theory. They are 'doubled' $SU(2)_1$ theories in a sense which we now describe. Consider the action of a Wilson loop operator in doubled $SU(2)$ Chern–Simons theory.

$$W_+[\{\gamma\}]\Psi[\{\alpha\}] = \Psi[\{\alpha \bigstar \gamma\}]. \tag{96}$$

Consider an intersection between α and γ. Using the prescription (87) for resolving overcrossings,

$$W_+[\{\gamma\}]\Psi[\{\alpha\}] = A\Psi[\{\alpha \cup_L \gamma\}] + A^{-1}\Psi[\{\alpha \cup_R \gamma\}]. \tag{97}$$

If we now take $A = \mathrm{e}^{\pi \mathrm{i}/6}$, which is almost *but not quite what we expect for* $k = 1$, and use the surgery relation $)(= -\smile$ (we will not show that this is the correct relation for the $k = 1$ theory until Section 5.5, but let us go ahead and use this relation nevertheless) then we find that this is simply

$$W_+[\{\gamma\}]\Psi[\{\alpha\}] = 2\mathrm{i}\sin(\pi/6)\Psi[\{\alpha \cup_L \gamma\}] = \mathrm{i}\Psi[\{\alpha \cup_L \gamma\}]. \tag{98}$$

Furthermore, $d = -A^2 - A^{-2} = -1$. Thus, this theory is equivalent to the $U(1)_2 \times \overline{U(1)_2}$ Chern–Simons theory. If, on the other hand, we take $A = \mathrm{i}\mathrm{e}^{\pi \mathrm{i}/6}$, as

expected for $k = 1$, then we find $d = 1$, which leads to Z_2 gauge theory. Thus, doubled $SU(2)_1$ is Z_2 gauge theory, while a slightly modified version is $U(1)_2 \times U(1)_2$. Chern–Simons theory.

In Section 5.5, we will show that these theories are so tightly constrained by combinatorial relations, so that self-consistency essentially specifies the entire structure. In the combinatorial world [49,50], at level k the $SU(2)_k$ theory is fully specified once a primitive $4r$th root of unity A, with $r = k + 2$, is given. (If A is only a primitive $2r^{\text{th}}$ root, the modular S-matrix is singular, but the Drinfeld double—as opposed to the mere tensor square of the singular theory—does have a non-singular modular S-matrix.) Our theory A is the doubled $SU(2)_1$ theory for $A = e^{2\pi i/12}$ while theory B is the Drinfeld doubled $SU(2)_1$ theory for $A = ie^{2\pi i/12}$. Both of these have level 1 as $SU(2)$ theories and, according to the Kauffman relation $d = -A^2 - A^{-2}$ have $d = -1$ and $d = +1$. So, although we introduced these theories through their relation to the Abelian groups $U(1)$ and Z_2, they are also low-level $SU(2)$ theories.

For the combinatorially defined $SU(2)$ theories (and their doubles) to be unitary, there is a strong restriction on A. In fact, for $k > 1$, A *must* be chosen as $A = \pm ie^{2\pi i/4r}$, $r = k + 2$, in order that the intrinsic inner product, even on the closed surface Y of genus 2, be positive [51]. The unitarity of our first example (theory A of Section 5.3) is a bit of an exception as seen in the table of undoubled $SU(2)$ theories below.

	k even	$k = 1$	$k \geqslant 3$ odd
$A = e^{2\pi i/4(k+2)}$	Non-unitary S-matrix non-singular	Unitary (d = −1) S-matrix non-singular	Non-unitary S-matrix non-singular
$A = ie^{2\pi i/4(k+2)}$	Unitary S-matrix non-singular	Unitary S-matrix singular *but* non-singular S-matrix for doubled theory = Z_2 gauge theory	Unitary S-matrix singular *but* S non-singular if restricted to integer spins

5.4. Contractible Wilson loops: the value of d

In Section 5.1, we saw that isotopy invariance fixed $d = 2\cos(\frac{\pi}{k+2})$ in the level k theory. This is somewhat surprising since d is the eigenvalue of $W_{\pm}[\bigcirc]$ in any state in the theory. Since the gauge fields $a_{\pm}$ are flat, we would expect their holonomies to be trivial, $U_{\pm}[\gamma] = 1_2$ and, therefore, $W_{\pm}[\gamma] = 2$. Why is d reduced from its naively expected value?

In our discussion of the Abelian theories, we were able to get away with a rather fast and loose treatment of Wilson loop operators. In general, these operators need

to be regularized. To see why, consider the expectation value of a product of Wilson loops in the full 2+1-dimensional theory:

$$\langle W[\gamma_1] \dots W[\gamma_n] \rangle = \int \mathscr{D}a W[\gamma_1] \cdots W[\gamma_n] \mathrm{e}^{S_{\mathrm{CS}}}. \tag{99}$$

In an Abelian theory, this is equal to $\exp(\frac{2\pi}{m} \sum_{ij} L(\gamma_i, \gamma_j))$ where $L(\gamma_i, \gamma_j)$ is the linking number of γ_i and γ_j. The problem is caused by the $i = j$ terms. The self-linking number is not well-defined without some kind of regularization, e.g., point-splitting. In mathematical terms, this is called a *framing* of the curve γ_i. One thickens the curve γ_i into a ribbon and then computes the linking number of the curves at the two ends of the ribbon. Clearly, this is not unique, since the ribbon can twist an arbitrary number of times, but once a framing has been chosen, a well-defined calculation can be done and the result for different choices of framing can be related to each other. In more physical terms, the amplitude for such a process depends not only on how the different anyons wind around each other but also—since each has fractional spin—on how each particle rotates during the process. This extra information—which is equivalent to the framing—must be specified in order to have a well-defined process.

In an Abelian theory, we can always choose a framing so that an unknotted contractible loop has self-linking number zero. Thus, it is possible to ignore this subtlety. In our construction of $U(1)_2 \times \overline{U(1)_2}$ Chern–Simons theory above, we actually took an unknotted loop to have self-linking number 1 since $W_+[\gamma]\Psi[\{\alpha\}] = \Psi[\{\alpha \cup_R \gamma\}] = -\Psi[\{\alpha\}] = \mathrm{e}^{\pi \mathrm{i} \cdot 1}\Psi[\{\alpha\}]$. In a non-Abelian theory, it is not possible to choose a framing so that the value of a Wilson loop is always unity if the loop is an unknotted contractible loop. It will, however, always be some constant, which we have called d:

$$\Psi[\{\alpha \cup \bigcirc\}] = d\Psi[\{\alpha\}]. \tag{100}$$

Let us compute d in $SU(2)_k$ Chern–Simons theory, following the arguments of Witten [26]. We will find that it is real, so its value will clearly be the same in the tensor square of the theory, which is our ultimate interest. Consider Chern–Simons theory on $D^2 \times R$—space is a disk D^2 and R is the time direction—and a curve $\gamma \in D^2 \times R$.

It may seem that the functional integral (99), even when evaluated for a single Wilson loop $W[\gamma]$ is not quite the same thing as the operator $W[\gamma]$ which is the starting point for our canonical quantization procedure because a curve γ in the functional integral need not lie on a constant time spatial slice. However, the distinction is illusory if the curve is unknotted because the Chern–Simons action is independent of the spacetime metric. Thus we can foliate spacetime into 'spatial slices' in any way that we like, in particular so that our unknotted curve γ lies on such a slice. This can be said slightly differently by noting that the value of (99) for an unknotted contractible loop does not depend on whether or not it lies in a single spatial plane, it only depends on its topological class. Hence, a computation of $W[\gamma]$ using the functional integral should be the same as the result obtained from canonical quantization.

For convenience, let us assume that γ does not lie in a single spatial slice D^2 and that it intersects any spatial slice in either 2 or zero points (except for the two spatial

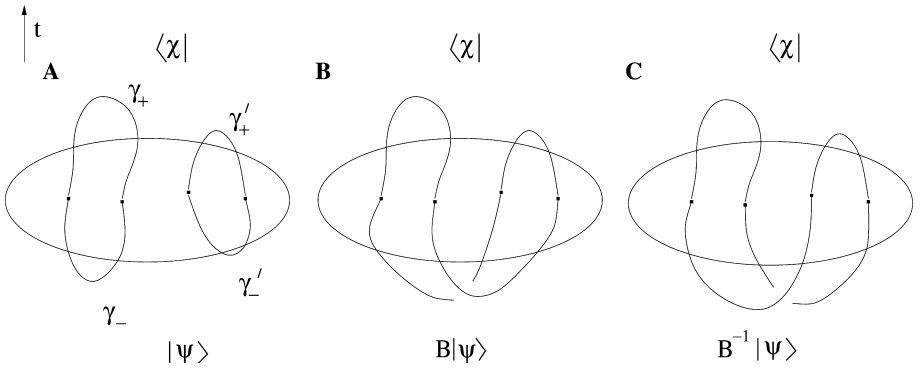

Fig. 12. The functional integrals which give (A) $\langle\chi|\psi\rangle$, (B) $\langle\chi|B|\psi\rangle$, (C) $\langle\chi|B^{-1}|\psi\rangle$.

slices which are tangent to γ). We define another topologically trivial curve γ' which is simply a copy of γ translated spatially. Since the functional integral only depends on the topological class of the curves γ, γ', we can take them far apart so that $\langle W[\gamma]W[\gamma']\rangle$ decouples into $\langle W[\gamma]\rangle\langle W[\gamma']\rangle$. Thus, $d^2 = \langle W[\gamma]W[\gamma']\rangle$.

Let us divide $D^2 \times R$ into two halves, $D^2 \times (-\infty, 0]$ and $D^2 \times [0, \infty)$ such that γ, γ' each intersect the spatial slice $t = 0$ at two points. We will call these points $\mathbf{x}_1, \mathbf{x}_2$ and $\mathbf{x}'_1, \mathbf{x}'_2$, as depicted in Fig. 12. Each is divided by this slice into two arcs, which we call γ_-, γ_+ and γ'_-, γ'_+. Then we can define a state in the $t = 0$ Hilbert space by performing the functional integral

$$\psi[A(\mathbf{x})] = \int_{a(\mathbf{x},0)=A(\mathbf{x})} \mathscr{D}a(\mathbf{x},t) W[\gamma_-]W[\gamma'_-] \times \mathrm{e}^{\int_{-\infty}^{0} \mathrm{d}t \int \mathrm{d}^2x \mathscr{L}_{\mathrm{CS}}}. \tag{101}$$

We can also define the state

$$\chi[A(\mathbf{x})] = \int_{a(\mathbf{x},0)=A(\mathbf{x})} \mathscr{D}a(\mathbf{x},t) W[\gamma_+]W[\gamma'_+] \times \mathrm{e}^{\int_{0}^{\infty} \mathrm{d}t \int \mathrm{d}^2x \mathscr{L}_{\mathrm{CS}}}. \tag{102}$$

The inner product of these two states is the functional integral which we to compute:

$$d^2 = \langle\chi|\psi\rangle = \int \mathscr{D}a(\mathbf{x},t) W[\gamma]W[\gamma'] \mathrm{e}^{\int_{-\infty}^{\infty} \mathrm{d}t \int \mathrm{d}^2x \mathscr{L}_{\mathrm{CS}}}. \tag{103}$$

Consider, now, the state obtained by deforming γ_-, γ'_- in order to perform a counterclockwise exchange of $\mathbf{x}_2$ and $\mathbf{x}'_1$. This state, which we will call $B|\psi\rangle$ is depicted in Fig. 12. Consider, as well, $B^{-1}|\psi\rangle$, obtained by deforming γ_-, γ'_- in order to perform a clockwise exchange of $\mathbf{x}_2$ and $\mathbf{x}'_1$. From the figure, we see that

$$\langle\chi|B|\psi\rangle = d, \tag{104}$$

$$\langle\chi|B^{-1}|\psi\rangle = d. \tag{105}$$

In Section 5.6, we will show that the four-quasiparticle Hilbert space is two-dimensional. Thus, B has two eigenvalues, λ_1, λ_2, so that

$$B - (\lambda_1 + \lambda_2) + \lambda_1\lambda_2 B^{-1} = 0, \tag{106}$$

which, in turn, implies that

$$d - (\lambda_1 + \lambda_2)d^2 + \lambda_1\lambda_2 d = 0 \tag{107}$$

so that

$$d = \frac{1 + \lambda_1\lambda_2}{\lambda_1 + \lambda_2}. \tag{108}$$

In Section 6, we will calculate these eigenvalues and show that

$$d = 2\cos\left(\frac{\pi}{k+2}\right). \tag{109}$$

5.5. Truncation of Hilbert space: Jones–Wenzl projectors

In this section, we take up the issue of truncating the pre-Hilbert spaces of Section 5.1 to finite-dimensional ones. We will do this by finding the analogues of the second equations of (58) and (71) which reduced the respective Hilbert spaces of $U(1)_2 \times \overline{U(1)_2}$ Chern–Simons theory and Z_2 gauge theory.

We begin by considering the structure of such constraints in general. Suppose that we wish to impose a relation involving n strands of a given 1-manifold (i.e., n segments of the 1-manifold which are 'close together'). We assume that there are no relations involving fewer than n strands; if there were, we could always use it to reduce a set of n strands to fewer than n strands, and the n-strand relation would be superfluous at best and incompatible at worst. Now consider our n-strand relation. If we were to connect the endpoints of two of the strands, then we would have an $n-1$ strand relation. By assumption, this is impossible. Hence, the putative $n-1$ relation must actually vanish identically. The same must be true for any other way of connecting two strands to yield an $n-1$ strand relation. Needless to say, any $n-2$, $n-3$, ... strand relations obtained in such a way must also vanish identically. This is a severe condition on our n strand relation.

In order to construct relations which satisfy this condition, it is useful to introduce the Temperley–Lieb algebra, TL_n. This algebra is most simply described in pictorial terms. The Temperley–Lieb algebra on n curves is made up of all planar diagrams without intersections in which n curves enter at the bottom and exit at the top. Some elements of the Temperley–Lieb algebra are depicted in the top panel of Fig. 13. The algebra is generated by $1, e_1, e_2, \ldots, e_{n-1}$, depicted in the bottom panel of Fig. 13. In the identity element of the Temperley–Lieb algebra, all curves go straight through. All other elements of the Temperley–Lieb algebra involve 'turnarounds' which join two incoming or two outgoing curves. The multiplication operation is simply stacking two pictures. Addition is simply formal linear superposition, i.e., pictures are multiplied by arbitrary complex numbers and superposed. The generators e_i satisfy the defining relations

$$e_i^2 = \mathrm{d}e_i, \tag{110}$$

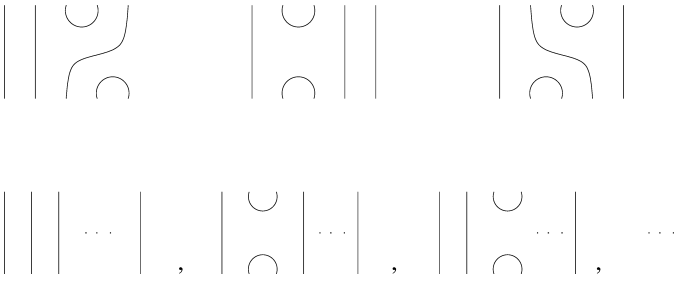

Fig. 13. In the top panel, three elements of TL_5 are depicted. In the bottom panel, the generators $1, e_1, e_2, \ldots, e_{n-1}$ are shown.

$$e_i e_j = e_j e_i \quad \text{for } |i-j| \geqslant 2, \tag{111}$$

$$e_i e_{i\pm 1} e_i = e_i. \tag{112}$$

Our desired n-curve relation is some element of the Temperley–Lieb algebra. Call it P_n. The condition that all $n-1, n-2, n-3, \ldots$ curve relations which can be derived from it must vanish identically is the statement that P_n annihilates every generator except the identity. The P_n are known as the Jones–Wenzl projectors.

Such projection operators are unique, up to an overall scalar factor. To see this, imagine that there were two such operators P_n and P'_n. Since P_n and P'_n are themselves elements of the Temperley–Lieb algebra, they can be written as $P_n = 1 + f$ and $P'_n = 1 + g$ (we are free to choose a scalar factor, so we set the coefficient of the identity to one in both expressions). Then $P_n P'_n = P_n(1+g) = P_n$. However, it is also equal to $P_n P'_n = (1+f)P'_n = P'_n$. Hence, $P_n = P'_n$.

We can construct the P_ns recursively. In order to do this, we first define the numbers Δ_n, which are the *traces* of the P_ns: we join every curve coming out of the top with its partner at the bottom, as shown in Fig. 14, and evaluate the resulting

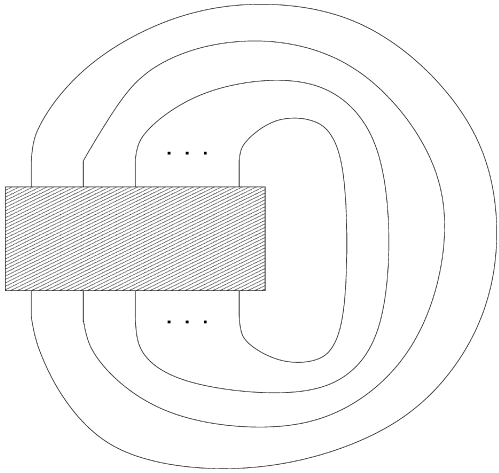

Fig. 14. The trace of an arbitrary element (the shaded region) of the Temperley–Lieb algebra.

diagram by assigning a factor of d for each closed loop. Now, P_2 can be found by inspection, $P_2 =)(-\frac{1}{d}\smile\frown$ (look familiar?). Let us suppose that we know $P_{n-1}, P_{n-2}, P_{n-3}, \ldots, P_2$. Then, connect one of the curves entering P_{n-1} with the corresponding curve leaving the top. The result annihilates all turnarounds on $n-2$ curves, so it is proportional to P_{n-2}. By comparing their traces, we see that the constant of proportionality is simply the ratios of the traces $\Delta_{n-2}/\Delta_{n-1}$. Now, consider the element of TL_n depicted in Fig. 15. Clearly, the $n-2$ turnarounds which are entirely on the first $n-1$ curves are annihilated. Consider a turnaround on the last two curves. As we see from the picture if we note that $P_{n-1}P_{n-2} = P_{n-1}$, it, too is annihilated. Thus, the operator shown in Fig. 15 is, indeed, the desired projection operator.

Now let us apply this result to the problem of finding an n-curve relation for a topological field theory with a pre-Hilbert space of states $\Psi[\{\alpha\}]$. As we noted above, if we connect any adjacent endpoints out of the set of $2n$ endpoints of n curves, then the

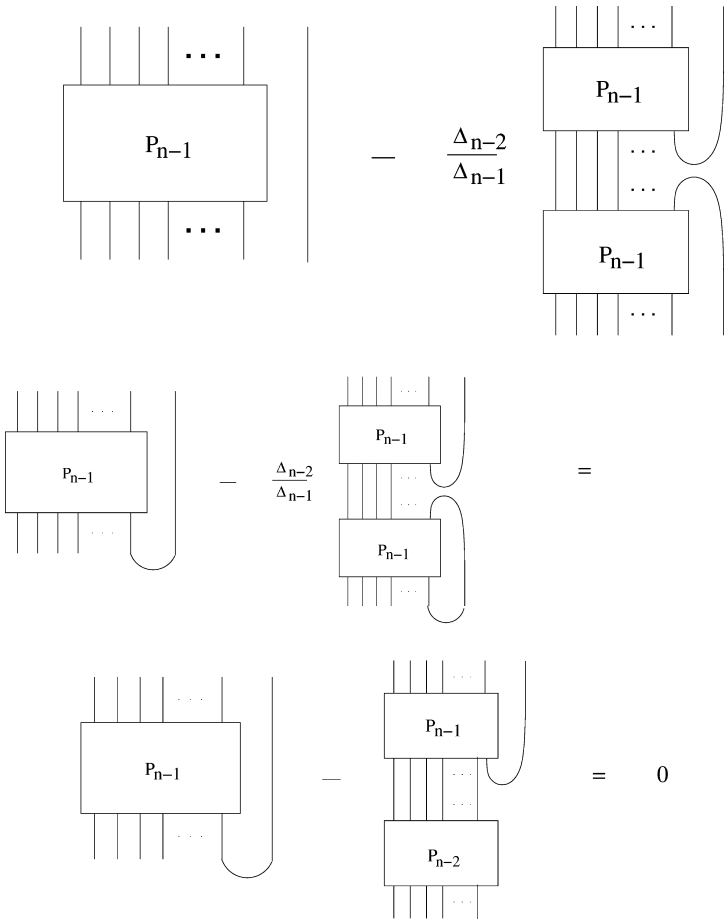

Fig. 15. In the top panel, the Jones–Wenzl projector P_n is written in terms of P_{n-1}. In the bottom two panels, it is shown how P_n annihilates a turnaround on the last two curves.

ensuing relation must vanish. If we divide the $2n$ endpoints into n incoming ones and n outgoing ones then it must not only be true that any turnaround attached to the top or the bottom or bottom must be annihilated, but also that connecting the first top endpoint to first bottom one or the last top endpoint to the last bottom one yields zero. Imposing the latter requirement on P_n implies that $dP_{n-1} - \frac{\Delta_{n-2}}{\Delta_{n-1}} P_{n-1} = 0$ or, simply $\Delta_{n-2}(d) = d\Delta_{n-1}(d)$. We have written the d dependence explicitly in order to emphasize that this condition restricts d. Only those ds which satisfy this relation can lead to a consistent n-curve relation. By taking the trace of our equation for P_n, we see that $\Delta_n = d\Delta_{n-1} - \Delta_{n-2}$. Hence, our condition is simply that the trace of P_n vanishes. If we pick a d such that Δ_n vanishes but $\Delta_{n-1}, \Delta_{n-2}$, etc. don't vanish (which will, in general, be possible) then we have the desired n-curve relation.

The equation $\Delta_n = d\Delta_{n-1} - \Delta_{n-2}$ is the recursion relation for the Chebyshev polynomials at $X = d/2$. Only at the roots of the n^{th} Chebyshev polynomial is there a consistent n-curve relation. Fortunately, the non-trivial roots of the $(k+1)^{\text{th}}$ Chebyshev polynomial are precisely the numbers $d_k = \pm 2\cos(\frac{\pi}{k+2})$ which arise in $SU(2)_k \times \overline{SU(2)_k}$ Chern–Simons gauge theory.

Thus, there is only one possible relation which we can impose in the level-k theory, and it involves $k+1$ strands. We are 'lucky' that we could even impose one. For other values of d, there is not even one such consistent relation. Let us write our $k+1$-strand relation as

$$1 + c_1 f_1 + c_2 f_2 + \cdots + c_q f_q = 0, \tag{113}$$

where $1, f_1, f_2, \ldots, f_q$ are elements of TL_{k+1}. Consider a 1-manifold α in which $k+1$ parallel strands have been brought close together. (In order to state this precisely, we may need to temporarily introduce a metric, define 'close,' and then show that, as a result of isotopy invariance, the final answer is independent of the metric.) Then let us define $f_j \cdot \alpha$ as the 1-manifold which is obtained by replacing the $k+1$ parallel strands by the element f_j of the Temperley–Lieb algebra TL_{k+1}.

Using this notation, we can write the Hilbert space of $SU(2)_k \times \overline{SU(2)_k}$ Chern–Simons gauge theory as the vector space of functionals of isotopy classes of loops which satisfy

$$\begin{aligned} &\Psi[\{\alpha \cup \bigcirc\}] - d_k \Psi[\{\alpha\}] = 0, \\ &\Psi[\{\alpha\}] + \sum_j c_j \Psi[\{f_j \cdot \alpha\}] = 0, \end{aligned} \tag{114}$$

where $d_k = 2\cos(\frac{\pi}{k+2})$, and $\Psi[\cdot]$ is defined for multi-curves with over- and under-crossings according to Eq. (87).

Applying the second relation, we see that we can relate $\Psi[\cdot]$ evaluated on an arbitrary isotopy class to $\Psi[\cdot]$ evaluated on isotopy classes with winding numbers less than or equal to k. Thus, the Hilbert space on the torus is $(k+1)^2$-dimensional.

Let us consider, for the sake of concreteness, the case $k = 2$. From the above discussion, we see that the Jones–Wenzl projector P_3 can be constructed from $P_2 =)(-\frac{1}{d}\smile\!\frown$, using $\Delta_2 = d^2 - 1$, $\Delta_1 = d$, and $d = \sqrt{2}$. It is displayed in Fig. 16. There are five terms in P_3, and the corresponding relation satisfied by states in the Hilbert space of $SU(2)_2 \times \overline{SU(2)_2}$ Chern–Simons theory is shown in Fig. 16.

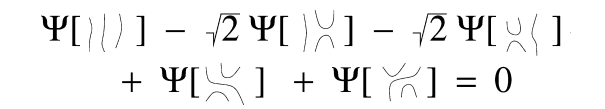

Fig. 16. The condition imposed on states in the Hilbert space of $SU(2)_2 \times \overline{SU(2)_2}$ Chern–Simons theory by the Jones–Wenzl projector P_3.

Thus, the Hilbert space on the torus is given by states $\Psi_{(n,m)}[\{\alpha\}]$ which vanish on all multi-curves α except those which are d-isotopic to the multi-curves (n,m) depicted in Fig. 17. For these multi-curves, the d-relation and isotopy can be used to determine their value relative to the value of $\Psi_{(n,m)}[.]$ evaluated on some fixed state in the d-isotopy class. One might have expected the states to correspond to the set of (n,m) with $n,m = 0,1,2$, but instead of (2,2), we have (−1,1) since $\Psi_{(2,2)} = 2\sqrt{2}\Psi_{(0,0)} - \Psi_{(2,0)} - \Psi_{(0,2)}$.

5.6. Quasiparticles in $SU(2)_k \times \overline{SU(2)_k}$ Chern–Simons gauge theory

Let us now consider quasiparticles and their braiding statistics. As in the Abelian cases which we considered earlier, quasiparticles are modeled as interior boundaries or punctures at which curves can terminate. Thus, quasiparticles are simply the allowed states on the annulus. We need only consider configurations with at most k curves extending from the quasiparticle to the outer boundary. The value of any physical wavefunctional on a configuration with more than k such curves can be related using P_{k+1} to its value on a configuration with fewer than $k+1$ such curves together with some

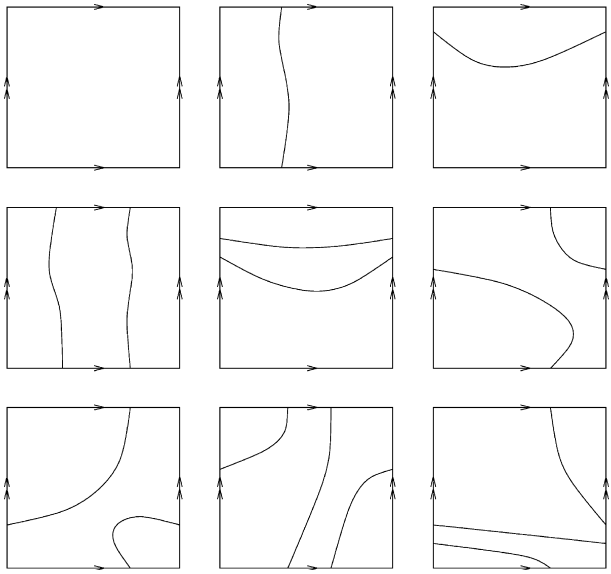

Fig. 17. The nine ground states of the $k = 2$ theory on the torus. Opposite sides of the square are identified, as indicated by the arrows.

curves which begin and end at the inner boundary of the annulus (and some which do the same at the outer boundary). The latter are edge excitations, which we ignore for now and discuss later. Thus, we can classify quasiparticles according to the number n of curves which terminate at them. In the level-k theory $n = 0, 1, \ldots, k$.

For concreteness, we focus on the level $k = 2$ theory, but the extension to other k is straightforward. Either $n = 0, 1$, or 2 curves can terminate at the inner boundary of the annulus. Let us choose two preferred points on the inner boundary and two more on the outer boundary where curves can terminate. We furthermore specify that it is preferable for curves to terminate at one of these points; a curve will terminate at the other preferred point only when the first one is already taken. This is not the most natural boundary condition, but it is convenient for counting states and drawing pictures. We will not allow bigons, curves which have both of their endpoints at the same boundary. As promised earlier, a discussion of boundary conditions and edge excitations will follow in a later section.

The allowed states on the annulus are depicted in Fig. 18. Note that there are three species of excitations (including the vacuum) in which there are no curves terminating at the quasiparticle, four species of excitations in which there is one curve terminating at the quasiparticle, and two species of excitation in which there are two curves terminating at the quasiparticle.

Let us denote the three states in Fig. 18A as $|0\rangle, |0_r\rangle, |0_{r^2}\rangle$. We denote the four states in Fig. 18B as $|1\rangle, |1_T\rangle, |1_{T^{-1}}\rangle, |1_{T^2}\rangle$, and the two states in Fig. 18C as $|2\rangle, |2_H\rangle$. The spin and two-particle braiding eigenstates are linear combinations of these:

$$\begin{aligned}
&\text{spin}0 : \sqrt{2}|0_r\rangle \pm |0_{r^2}\rangle, 2|0\rangle - |0_{r^2}\rangle,\\
&\text{spin} \pm \frac{1}{2} : |2\rangle \mp \mathrm{i}|2_H\rangle,\\
&\text{spin}\epsilon\left(\frac{4 \pm 1}{16}\right) : \pm \mathrm{e}^{\epsilon\pi\mathrm{i}/8}|1\rangle - \mathrm{e}^{\epsilon\pi\mathrm{i}/4}|1_T\rangle \mp \mathrm{e}^{-\epsilon\pi\mathrm{i}/8}|1_{T^{-1}}\rangle + |1_{T^2}\rangle\epsilon = \pm 1.
\end{aligned} \tag{115}$$

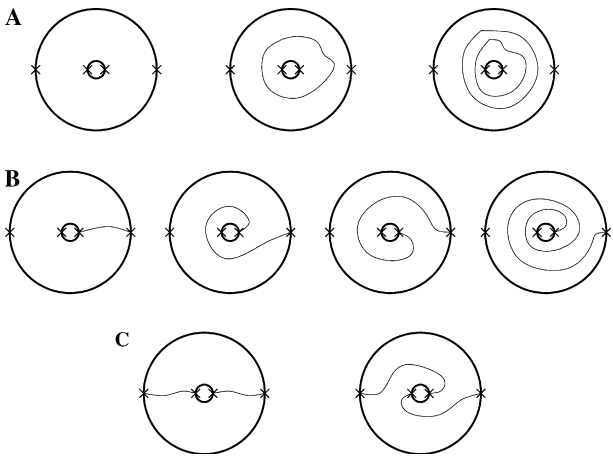

Fig. 18. The nine quasiparticle species (including the vacuum, or 'trivial particle') of the $k = 2$ theory. See text for an explanation.

(N.B. By 'spin' s, we mean here that the state has eigenvalue $e^{2\pi i s}$ under a spatial rotation by 2π. This is entirely distinct from the $SU(2) \times SU(2)$ representation which it might carry.)

These quasiparticles can be associated with corresponding $SU(2)$ representations. Let us assume that every particle carries two $SU(2)$ quantum numbers (j_1, j_2), with $j_i = 0, \frac{1}{2}, 1$. Let us decompose the product of these quantum numbers: $j_1 \otimes j_2 = |j_1 - j_2|, |j_1 - j_2| + 1, \ldots, \max(j_1 + j_2, 1)$. This is almost what we would expect for the Lie group $SU(2)$, except for the upper cutoff $k/2$ which, in this case, is 1. In the next section, we will see the origin of this upper cutoff in the language of Kac–Moody algebras. Now consider the 9 allowed combinations (j_1, j_2) and decompose their $j_1 \otimes j_2$ products into irreducible representation of total j. Four of these products are $j = \frac{1}{2}$. These correspond to the four states with a single curve terminating at the inner boundary, in Fig. 18B. There are also two $j = 0$ quasiparticles and three $j = 1$ quasiparticles. These correspond, respectively, to the first two states in Fig. 18A and the two states in Fig. 18C together with the third one in 18A.

For general k, the $SU(2)_k$ theory is expected to only have the representations $j = 0, \frac{1}{2}, \ldots, \frac{k}{2}$, and the $(k+1)^2$ quasiparticle species can be organized in the manner just described for $k = 2$. in the next section, we will show how this can be facilitated with some results from conformal field theory. Incidentally, this is why there are only four states of two $j = 1/2$ quasiparticles in $SU(2)_k$ theory: pairwise, they must form $j = 0$ or $j = 1$.

Let us now consider the braiding statistics of these quasiparticles. If we have two quasiparticles, their braiding statistics is determined by the quasiparticle spins. Note that on a compact manifold, if there are only two quasiparticles, then they must fuse to something topologically trivial. On the annulus, they could fuse to form a non-trivial quasiparticle because the outer boundary can compensate. For instance, suppose we have two $n = 1$ quasiparticles. On the sphere, they must fuse to form $n = 0$ (in fact, they must fuse to form the trivial particle); on the annulus, they can fuse to form $n = 2$ since the two curves can terminate at the outer boundary, which must also have $n = 2$. In either case, performing a braid cannot change n of the composite. In fact, we could braid one quasiparticle around another and then fuse them or simply fuse them and rotate the fused particle. This must lead to the same result, which is clearly just a phase in the second approach.

Suppose we have four $n = 1$ quasiparticles in the $k = 2$ theory. There are five possible states of four quasiparticles on the sphere. To see this, note that fusing two of the quasiparticles can lead to one of the three $n = 0$ quasiparticles or one of the two $n = 2$ quasiparticles. The other two quasiparticles must fuse to form the same excitation since the aggregate of all four quasiparticles must be topologically trivial. These five states are depicted in Fig. 19. (This is just one basis of the five states. One could easily draw a different set of five pictures but they would be related to these using P_3.) It is clear from the figure that taking particle 2 around particles 3 and 4 transforms the first state into the second. It is also clear that this does not commute with, say, exchanging particles 2 and 3. Thus, these particles exhibit non-Abelian statistics. The $k = 1$ case is special because P_2 allows us to collapse all of these states into one state. In this special case, the statistics is Abelian, since there is only

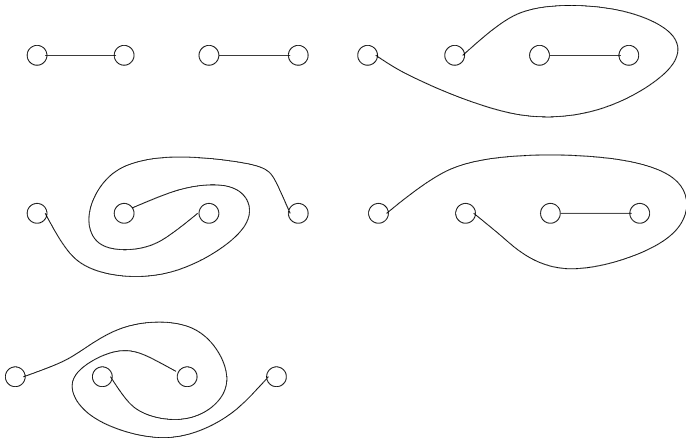

Fig. 19. A complete, linearly independent (but not orthogonal) set of five states of two $n = 1$ quasiparticles in the $k = 2$ theory.

one p-quasiparticle state, so it can at most acquire a phase. For higher k, P_{k+1} is simply not as restrictive. The braiding matrices for these theories can be computed simply by drawing pictures and applying the projector relation P_{k+1}.

6. Relation to 2*D* conformal field theory

At first glance, $2D$ (or $1 + 1 - D$) conformal field theories would seem to be ill-suited to describe topological phases of 2+1-dimensional physical systems. Obviously, the dimensionality is wrong. Furthermore, conformal field theories describe critical points, at which the spectrum is gapless, not stable phases with a gap. However, upon closer inspection, conformal field theories do, in fact, have many of the necessary ingredients for exotic braiding properties. Consider the holomorphic (or right-moving) part of a conformal field theory. Since the spectrum is gapless, the two-point correlation function of a field Φ varies as a power law: $\langle \Phi(z)\Phi(0)\rangle \sim 1/z^{2h}$. If h is not a half-integer, then this correlation function is multi-valued. A phase is acquired as one field encircles another, analogous to what occurs when one particle is taken around another in 2+1 dimensions. While a chiral two-point function will suffer from at most a phase ambiguity on the plane or sphere, it can have more severe non-uniqueness on higher-genus manifolds as can the chiral part of a correlation function, of four or more fields (i.e., a conformal block) on the plane itself [52]. In general, there will be a vector space (which is finite-dimensional, in the case of rational conformal field theories) of conformal blocks. Taking one field around another will lead to monodromy matrices rotating the conformal blocks into each other.

A second propitious feature of conformal field theories is the existence of an operator product expansion. When two fields are viewed from a distance much larger

than their separation, they appear much as a single 'composite' field would or, rather, as a linear superposition of the different possible 'composite' fields:

$$\phi_i(z)\phi_j(w) = \sum_k c_{ijk}(z-w)^{h_k-h_i-h_j}\phi_k(w) \tag{116}$$

so a number of different ϕ_ks can appear on the right-hand-side of this equation. This is analogous to the 'fusion' of two particles in 2+1 dimensions: when two particles are brought close together, they can fuse to form a single composite particle. There are different possibilities for this particle, so the result of fusion is, in fact, a linear superposition of composite particles.

This is a particularly convenient decomposition for discussing braiding. When a counter-clockwise exchange of z with w is performed, the coefficient of ϕ_k changes by a phase factor $e^{\pi i(h_k-h_i-h_j)}$. In other words, the sum on the right-hand-side is over eigenvectors of a counter-clockwise exchange. If we consider the four-point correlation function, there will, in general, be several conformal blocks, which transform according to monodromy matrices as a result of exchange operations. These matrices have eigenvalues $e^{\pi i(h_k-h_i-h_j)}$.

The resemblance between conformal field theory and Chern–Simons theory is not coincidental, but reflects an underlying relationship between the theories. It is easiest to see the connection between the two by considering Chern–Simons theory for some semi-simple Lie group, G, on the $2+1-D$ manifold $D^2\times R$, where R is identified with the time direction. In Coulomb gauge, $a_0^a = 0$ (where a is a Lie algebra index), the Hamiltonian vanishes and there is only the constraint $f_{\mu\nu}^a = 0$. In the functional integral formulation, this is expressed as follows. The field a_0^a only appears in the action linearly, so the functional integral over a_0^a may be performed, yielding a δ-function:

$$\int Da e^{k/4\pi\int_{D^2\times R}\epsilon^{\mu\nu\lambda}\left(a_\mu^a\partial_\nu a_\lambda^a+\frac{2}{3}f_{abc}a_\mu^a a_\nu^b a_\lambda^c\right)} = \int Da_i\delta(f_{ij}^a)e^{\frac{k}{4\pi}\int_{D^2\times R}\epsilon^{ij}a_i^a\partial_0 a_j^a}, \tag{117}$$

where $i,j=1,2$ run over spatial indices. The constraint imposed by the δ-function can be solved by taking

$$a_i^a = \partial_i U U^{-1}, \tag{118}$$

where U is a single-valued function taking values in the Lie group. Substituting this into the right-hand-side of (117), we find that the action which appears in the exponent in the functional integral takes the form

$$\begin{aligned}
S &= \frac{k}{4\pi}\int_{D^2\times R}\epsilon^{ij}\mathrm{tr}\left(\partial_i UU^{-1}\partial_0\left(\partial_j UU^{-1}\right)\right)\\
&= \frac{k}{4\pi}\int_{D^2\times R}\epsilon^{ij}\left[\mathrm{tr}\left(\partial_i UU^{-1}\partial_0\partial_j UU^{-1}\right)+\mathrm{tr}\left(\partial_i UU^{-1}\partial_j U\partial_0 U^{-1}\right)\right]\\
&= \frac{k}{4\pi}\int_{D^2\times R}\epsilon^{ij}\left[\partial_j\mathrm{tr}\left(\partial_i U^{-1}\partial_0 U\right)+\mathrm{tr}\left(\partial_i UU^{-1}\partial_j U\partial_0 U^{-1}\right)\right]\\
&= \frac{k}{4\pi}\int_{S^1\times R}\mathrm{tr}\left(\partial_1 U^{-1}\partial_0 U\right)+\frac{k}{12\pi}\int_{D^2\times R}\epsilon^{\mu\nu\lambda}\mathrm{tr}\left(\partial_\mu UU^{-1}\partial_\nu UU^{-1}\partial_\lambda UU^{-1}\right).
\end{aligned} \tag{119}$$

The Jacobian which comes from the δ-function $\delta(f_{ij}^a)$ is cancelled by that associated with the change of integration variable from Da to DU. In the final line, we have integrated by parts the first term so that the integral is over the boundary, which is parameterized by the coordinate x_1. The second term has an extra factor of 1/3 resulting from the more symmetrical form in which we have written it. Though it appears to be an integral over the $3D$ manifold, it only depends on the boundary values of U. This action is the chiral WZW action, as we will see momentarily.

The equation of motion is simply

$$\partial_0\left(U^{-1}\partial_1 U\right) = 0. \tag{120}$$

The equation of motion only deals with the restriction of U to the boundary, $S^1 \times R$, which is one way of seeing that the action is independent of the continuation from $S^1 \times R$ to $D^2 \times R$. This equation is a slightly obscured version of the chiral WZW equation of motion. The Chern–Simons action has a vanishing Hamiltonian which is why the current $U^{-1}\partial_1 U$ is independent of time. With a more complicated boundary condition, we could impose non-trivial dynamics at the boundary, which would lead to a similar "off-diagonal" derivative structure in the equation of motion, but with $\partial_0 \to \partial_t - \partial_1$, $\partial_\phi \to \partial_0 + \partial_1$:

$$\partial_-\left(U^{-1}\partial_+ U\right) = 0, \tag{121}$$

which is the usual equation of motion of the right-handed part of the WZW model.

Either of these equations, (120) or (121), states that the currents

$$J^a = \mathrm{tr}\left(T^a U^{-1}\partial_1 U\right) \tag{122}$$

are free, where the T^as are the generators of the Lie algebra in the adjoint representation. Consequently, they obey a Kac–Moody algebra, which allows us to algebraically compute the braiding properties of primary fields. The OPE of the currents yields the Kac–Moody algebra:

$$J^a(x_1)J^b(0) = \frac{k\delta^{ab}}{x_1^2} + \frac{f^{abc}J^c(0)}{x_1} + \cdots, \tag{123}$$

where f^{abc} are the structure constants of the Lie algebra. The energy-momentum tensor of this theory is of the Sugawara form:

$$T = \frac{1/2}{k + C_A} : J^a J^a :, \tag{124}$$

where C_A is the quadratic Casimir in the adjoint representation if the highest root is normalized to length 1. A field $\varphi_{(r)}$ which transforms in representation r of the group G and is primary under the Kac–Moody algebra,

$$J^a(x_1)\varphi_{(r)}(0) = \frac{T^a_{(r)}\varphi_{(r)}(0)}{x_1} + \cdots \tag{125}$$

has, according to (124), dimension

$$T(x_1)\varphi_{(r)}(0) = \frac{1}{x_1^2}\frac{T^a_{(r)}T^a_{(r)}/2}{k + C_A} + \cdots = \frac{1}{x_1^2}\frac{C_r}{k + C_A} + \cdots \tag{126}$$

For the case of $SU(2)_k$, this means that a spin j primary field has dimension $h_j = j(j+1)/(k+2)$.

A restriction on the allowed j's in the $SU(2)_k$ theory can be found by expanding $J^a(x_1)$ in modes

$$J^a(x_1) = \sum_m J^a_m \mathrm{e}^{-\mathrm{i}(m+1)x_1} \tag{127}$$

Then the operators $I^a \equiv J^a_0$ form an $su(2)$ Lie algebra. Hence, $2J^3_0$ has integer eigenvalues in any finite-dimensional unitary representation. Similarly, $\tilde{I}^1 \equiv (J^1_1 + J^1_{-1})/\sqrt{2}$, $\tilde{I}^2 \equiv (J^2_1 + J^2_{-1})/\sqrt{2}$, $\tilde{I}^3 \equiv \frac{1}{2}k - J^3_0/\sqrt{2}$ also form an $su(2)$ Lie algebra. Consider a spin j highest weight state $|j, m = j\rangle$, with $I^3|j, m = j\rangle = j|j, m = j\rangle$. Then

$$\begin{aligned} 0 &\leqslant \langle j, m = j|\tilde{I}^+\tilde{I}^-|j, m = j\rangle \\ &= \langle j, m = j|[\tilde{I}^+, \tilde{I}^-]|j, m = j\rangle \\ &= \langle j, m = j|k - 2I^3|j, m = j\rangle = k - 2j. \end{aligned} \tag{128}$$

Hence, the $SU(2)_k$ theory has particles which transform under the $j = 1/2, 1, \ldots, k/2$ representations. These particles have 'spins' $h_j = j(j+1)/(k+2)$, which determine how they transform under a spatial rotation. They have fusion rule $j_1 \otimes j_2 = \oplus j_3$ with $|j_1 - j_2| < j_3 < \min(j_1 + j_2, k - j_1 - j_2)$ (the derivation may be found in [53]). Together, these completely determine the braiding properties of the particles.

A spin-j primary field in the $SU(2)_k$ conformal field theory corresponds to a Wilson loop in the spin-j representation in the $SU(2)_k$ Chern–Simons theory. Thus, we can calculate the braiding properties of Wilson lines in Chern–Simons theory using the OPE in the associated conformal field theory. Consider the spin-1/2 representation of $SU(2)_k$. Then,

$$\phi^\alpha_{1/2}(z)\phi^\beta_{1/2}(0) = z^{-3/(2k+4)}\epsilon^{\alpha\beta} + Cz^{5/(2k+4)}\epsilon^{\alpha\delta}\tau^\beta_\delta \cdot \phi_1(0), \tag{129}$$

where α, β, δ are spinor indices, τs are Pauli matrices, and C is a known constant. Thus, there is a 2×2 braid matrix associated with an exchange and its eigenvalues are $-\mathrm{e}^{3\pi\mathrm{i}/(2k+4)}$, $\mathrm{e}^{5\pi\mathrm{i}/(2k+4)}$. Note, however, that there is an extra half-twist or a deficit of a half-twist, respectively, associated with each of these operations (i.e., the framing associated with them is different from the standard one which we chose earlier), so that the desired eigenvalues are actually $-\mathrm{e}^{\pi\mathrm{i}/(2k+4)}$, $\mathrm{e}^{3\pi\mathrm{i}/(2k+4)}$.

This result can be used to calculate the value of a contractible, unknotted Wilson loop, according to the derivation of the previous section:

$$d = 2\cos\left(\frac{\pi}{k+2}\right). \tag{130}$$

7. Edge excitations

In our previous discussion of the Hilbert space of doubled Chern–Simons theory in the presence of a manifold with boundaries or in the presence of quasiparticles, we explicitly forbid bigons with endpoints at the same boundary. Now, however, it is

time to let bigons be bigons. We also remove the constraint which fixed the endpoints of curves to lie at marked points. As we will see, the boundary excitations can be understood in terms of the conformal field theories of the previous section.

In general, the endpoint of a curve which terminates at a boundary will be described by a wavefunction $\psi(\theta)$. Previously, we required $\psi(\theta) = \delta(\theta - \theta_0)$. This can be viewed as the correct ground state in the presence of a particular boundary Hamiltonian. The most important terms in the boundary Hamiltonian will be of the form

$$H = \sum_i \left(\beta L_i^2 + \eta\right). \tag{131}$$

In other words, there will be an energy penalty η for each endpoint and a kinetic energy proportional to the square of the angular momentum, $L_i = \partial/\partial\theta_i$. Thus, the eigenfunctions in the sector with n endpoints will be angular momentum eigenstates

$$\psi(\theta_1, \theta_2, \ldots, \theta_n) = \mathrm{e}^{\mathrm{i}m_1\theta_1 + \mathrm{i}m_2\theta_2 + \cdots + \mathrm{i}m_n\theta_n}, \tag{132}$$

where $\theta_1, \theta_2, \ldots, \theta_n$ are the angular positions of the n endpoints. These may be endpoints of bigons or of curves which ultimately terminate at other boundaries.

The allowed values m_i are determined by consistency with the surgery relation of the theory. In general, they are not integers. Let us consider, as examples, $k = 1, 2$, corresponding to $d = -1, \sqrt{2}$. First, consider states with a single endpoint in the $d = -1$ theory. Using P_2, we see that the wavefunction $\psi(\theta) = \mathrm{e}^{\mathrm{i}m\theta}$ must satisfy the relation $\psi(\theta) = -\psi(\theta + 4\pi)$. Hence, $m \in Z \pm \frac{1}{4}$. States with more than one endpoint are equivalent to states with either zero or one endpoint and some number of bigons. Under a 2π rotation, the bigons are unchanged, so the total angular momentum $M = \sum_i m_i$ must satisfy $m \in Z$ or $m \in Z \pm \frac{1}{4}$, respectively, for an even or odd number of endpoints.

Now, consider the $k = 2$ theory. For a single endpoint, P_3 dictates that $\psi(\theta + 2\pi) = \mathrm{e}^{2\pi\mathrm{i}s}\psi(\theta + 2\pi)$, where s takes one of the four values $s = \pm(\frac{4\pm1}{16})$, as in Eq. (115). In other words, $m \in Z + s$. For two endpoints, it dictates that $s = \pm\frac{1}{2}$. States with more than two endpoints can always be related using P_3 to a state with 0,1, or 2 endpoints and some number of bigons. The bigons are, of course, invariant under $\theta \to \theta + 2\pi$, so the constraint on the angular momentum is the same as above, depending on the number of endpoints modulo 3.

Thus, we see that the boundary states of the doubled Chern–Simons theories can be understood as massive deformations of the corresponding achiral conformal field theories which we discussed in the previous section. If we were dealing with undoubled chiral topological theories, the corresponding edge excitations would be chiral and, therefore, necessarily gapless. Achiral theories must be tuned—so that $\eta = 0$ in (131)—in order to be gapless. The basic state counting is the same, however. In the $SU(2)_1 \times \overline{SU(2)_1}$ case, we have the spin 0 sector—or states with an even number of endpoints—corresponding to the towers of states:

$$J^L_{-n_1} \cdots J^L_{-n_m} J^R_{-l_1} \cdots J^R_{-l_q} |0\rangle, \tag{133}$$

$$J^L_{-n_1} \cdots J^L_{-n_m} J^R_{-l_1} \cdots J^R_{-l_q} \phi^R_{1/2} \phi^L_{1/2} |0\rangle. \tag{134}$$

The n_i, l_i correspond to the angular momenta differences between the two endpoints of the associated bigons. $\phi^R_{1/2}$, $\phi^L_{1/2}$ are the $SU(2)$ doublet primary fields. We also have the spin $\pm\frac{1}{4}$ sectors—states with an odd number of endpoints—corresponding to the towers of states:

$$J^L_{-n_1}\cdots J^L_{-n_m}J^R_{-l_1}\cdots J^R_{-l_q}\phi^R_{1/2}|0\rangle, \tag{135}$$

$$J^L_{-n_1}\cdots J^L_{-n_m}J^R_{-l_1}\cdots J^R_{-l_q}\phi^L_{1/2}|0\rangle. \tag{136}$$

The fields $\phi^R_{1/2}$, $\phi^L_{1/2}$ have spins $\pm\frac{1}{4}$, confirming the correspondence.

$SU(2)_2\times\overline{SU(2)_2}$ has the states with zero endpoints modulo 3:

$$J^L_{-n_1}\cdots J^L_{-n_m}J^R_{-l_1}\cdots J^R_{-l_q}|0\rangle, \tag{137}$$

$$J^L_{-n_1}\cdots J^L_{-n_m}J^R_{-l_1}\cdots J^R_{-l_q}\phi^R_{1/2}\phi^L_{1/2}|0\rangle, \tag{138}$$

$$J^L_{-n_1}\cdots J^L_{-n_m}J^R_{-l_1}\cdots J^R_{-l_q}\phi^R_{1}\phi^L_{1}|0\rangle. \tag{139}$$

The states with one endpoint modulo 3:

$$J^L_{-n_1}\cdots J^L_{-n_m}J^R_{-l_1}\cdots J^R_{-l_q}\phi^R_{1/2}|0\rangle, \tag{140}$$

$$J^L_{-n_1}\cdots J^L_{-n_m}J^R_{-l_1}\cdots J^R_{-l_q}\phi^L_{1/2}|0\rangle, \tag{141}$$

$$J^L_{-n_1}\cdots J^L_{-n_m}J^R_{-l_1}\cdots J^R_{-l_q}\phi^L_{1}\phi^R_{1/2}|0\rangle, \tag{142}$$

$$J^L_{-n_1}\cdots J^L_{-n_m}J^R_{-l_1}\cdots J^R_{-l_q}\phi^R_{1}\phi^L_{1/2}|0\rangle \tag{143}$$

have spins with are an integer plus $\pm 3/16$, $\pm 1/16$ since $\phi^{R,L}_{1/2}$ has spin $\pm 3/16$ and $\phi^{R,L}_{1}$ has spin $\pm 1/2$. The states with two endpoints modulo 3:

$$J^L_{-n_1}\cdots J^L_{-n_m}J^R_{-l_1}\cdots J^R_{-l_q}\phi^R_{1}|0\rangle \tag{144}$$

$$J^L_{-n_1}\cdots J^L_{-n_m}J^R_{-l_1}\cdots J^R_{-l_q}\phi^L_{1}|0\rangle \tag{145}$$

have spins $\pm 1/2$.

Continuing in this way, we could construct the edge excitations for any of the $SU(2)_k\times\overline{SU(2)_k}$ theories: they are simply massive theories corresponding to the associated achiral conformal field theories. The allowed weights of the primary fields follow from an application of the projector relation P_{k+1}, as we have shown explicitly for the cases of $k=1,2$.

8. Towards microscopic model hamiltonians

Our formulation of doubled Chern–Simons theories in terms of Hilbert spaces spanned by configurations of multi-curves on surfaces is compact and elegant.

However, its virtues are not purely esthetic, but also include its natural connection to microscopic models which give rise to these phases. In Section 3, we showed how they could arise from mean-field solutions of a variety of microscopic models of interacting electrons, but did not show that these were the true ground states of any particular Hamiltonians. Here, we take a different tack and take some steps towards a more direct connection between microscopic models and the loop space formulation of doubled Chern–Simons theories. As we observe in this section, many systems admit a loop space description.

Consider a system of $s = 1/2$ spins on a triangular lattice. Let us work in the S_z basis in which every spin takes a definite value $\uparrow, \downarrow$. Let us represent these basis states in terms of the domain walls which separate clusters of up- and down-spins. For every configuration of domain walls, there are two spin configurations which are related by a reversal of all spins. The domain walls lie on the links of the dual honeycomb lattice and the spins sit at the centers of the faces of the honeycomb lattice. It is clear that these domain walls can neither terminate nor cross. Thus the Hilbert space of a triangular lattice spin system is of precisely the desired form, as depicted in Fig. 20. However, only very special Hamiltonians will lead to a ground state which obeys relations such as (114).

For the initial members of our sequence of theories, $d = \pm 1$, such Hamiltonians can be written in a simple form. Consider theory B, the $d = 1$ theory. The Hamiltonian $H = h \sum_i S_i^x$ requires every spin to point in the x-direction in spin space or, in other words, in an equal linear superposition of the states $|\uparrow\rangle$ and $|\downarrow\rangle$. Flipping a spin causes the dark and light bonds of the corresponding hexagonal plaquette to be exchanged in Fig. 20. This can either create a new contractible loop, erase a loop of minimal size, deform a loop, or perform the surgery operation $)(\rightarrow \smile\!\frown$.

Thus, we have found that a system of spins in a magnetic transverse field is trivially equivalent to the simplest of our topological theories. This is a special case ($J = 0 \cong K = \infty$) of the duality between the transverse field Ising model

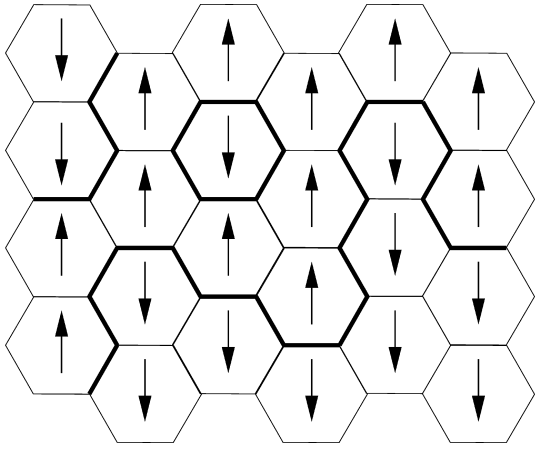

Fig. 20. S_z basis states of $s = 1/2$ spins on the triangular lattice can be represented in terms of loops on the honeycomb lattice.

$$H = -J \sum_{\langle i,j \rangle} S_i^z S_j^z + h \sum_i S_i^x \tag{146}$$

and the Z_2 gauge theory [54].

However, this is only the 'even' part of the theory. Domain walls cannot terminate in the bulk of the system and they can only terminate in pairs at a circular boundary. Thus, the above Hamiltonian does not admit excitations such as the one depicted in Fig. 5. However, the following simple model due to Kitaev [35] contains both the even and odd parts of the theory. There is a spin-1/2 degree of freedom $\sigma_z = \pm 1$ on each *link* of a lattice. The lattice is arbitrary, but let's consider a honeycomb lattice, for the sake of concreteness.

$$H = J_1 \sum_i A_i - J_2 \sum_p F_p, \tag{147}$$

where

$$\begin{aligned} A_i &\equiv \Pi_{\alpha \in \mathcal{N}(i)} \sigma_\alpha^z \\ F_p &\equiv \Pi_{\alpha \in p} \sigma_\alpha^x. \end{aligned} \tag{148}$$

These operators all commute,

$$\left[F_p, F_{p'}\right] = [A_i, A_j] = [F_p, A_j] = 0 \tag{149}$$

so the model can be solved exactly by diagonalizing each term in the Hamiltonian: the ground state $|0\rangle$ satisfies $A_i|0\rangle = -|0\rangle, F_p|0\rangle = |0\rangle$. If we represent $\sigma_z = 1$ by colored bonds and $\sigma_z = -1$ by uncolored bonds, then $A_i|0\rangle = -|0\rangle$ requires chains of bonds to never end, while $F_p|0\rangle = |0\rangle$ requires the ground state to contain an equal superposition of any configuration with one obtained from it by creating a new contractible loop, erasing a loop of minimal size, isotopically deforming a loop, or performing the surgery operation $)(\to \smile\frown$. This is clearly the same as the transverse field spin model above, except that curves can now terminate, albeit with an energy cost $2J_1$.

In a similar way, we can formulate a model which gives rise to the $d = -1$ theory.

$$H = J_v \sum_v A_v - J_i \sum_p F_p^i + J_{d,s} \sum_p F_p^{s,d}, \tag{150}$$

where

$$\begin{aligned} A_v &\equiv \Pi_{\alpha \in \mathcal{N}(v)} \sigma_\alpha^z, \\ F_p^i &\equiv \sigma_1^+ \sigma_2^+ \sigma_3^+ \sigma_4^+ \sigma_5^+ \sigma_6^- + \text{h.c.} + \sigma_1^+ \sigma_2^+ \sigma_3^+ \sigma_4^+ \sigma_5^- \sigma_6^- + \text{h.c.} + \sigma_1^+ \sigma_2^+ \sigma_3^+ \sigma_4^- \sigma_5^- \sigma_6^- + \text{h.c.} \\ &\quad + \sigma_1^+ \sigma_2^- \sigma_3^+ \sigma_4^- \sigma_5^+ \sigma_6^- + \text{h.c.} + \text{cyclic permutations}, \\ F_p^{s,d} &\equiv \sigma_1^x \sigma_2^x \sigma_3^x \sigma_4^x \sigma_5^x \sigma_6^x - F_p^i. \end{aligned} \tag{151}$$

The first term in the Hamiltonian is the same as in the $d = 1$ theory: it selects a low-energy subspace in which chains of up-spins never terminate. Hence, the configurations of this low-energy subspace can be represented as closed multi-curves. In the $d = 1$ theory, isotopy, the condition $d = 1$, and the surgery relation can all be

implemented with a single plaquette term. For $d = -1$, however, the ground state must contain an equal superposition of isotopic configurations but superpositions with opposite signs of configurations which are related through surgery or the addition of a contractible loop. These conditions are enforced by the second and third terms in the Hamiltonian. (The fourth term in F_p^i is actually the result of two surgeries and, therefore, carries a factor of $d^2 = 1$, the same as the isotopy moves with which we have grouped it.) The vertex and plaquette terms in the Hamiltonian commute with themselves and each other. Hence, this Hamiltonian is exactly soluble. Following steps similar to those of the previous paragraph, we see that this Hamiltonian implements the $d = -1$ theory and supports semionic excitations.

We expect that other physical systems can give rise to this type of structure and, with some luck, topological phases corresponding to doubled Chern–Simons theories. Dimer models have a natural representation in terms of multi-curves. In these models, it is assumed that that there are spins located at the sites of a lattice and that each spin forms a singlet dimer with one of its nearest neighbors. Thus, there are dimers on the links of the lattice which satisfy the following condition: there is one and only one dimer touching each site. A dimer covering is not, of course, composed of closed curves, but the *transition graph* between two dimer covering is: one considers some fixed reference dimer covering R and superposes it on the dimer covering of interest C. Where R and C coincide, we erase the dimers (or, perhaps, think of them as a minimum size closed loop). The remaining dimers form closed loops, with dimers from R and C alternating as one travels along the loop, as shown in Fig. 21. A dynamics is now needed which assigns a weight d to contractible loops, allows the loops to deform isotopically, and enacts P_{k+1}. The Kivelson-Rokhsar [21] Hamiltonian on

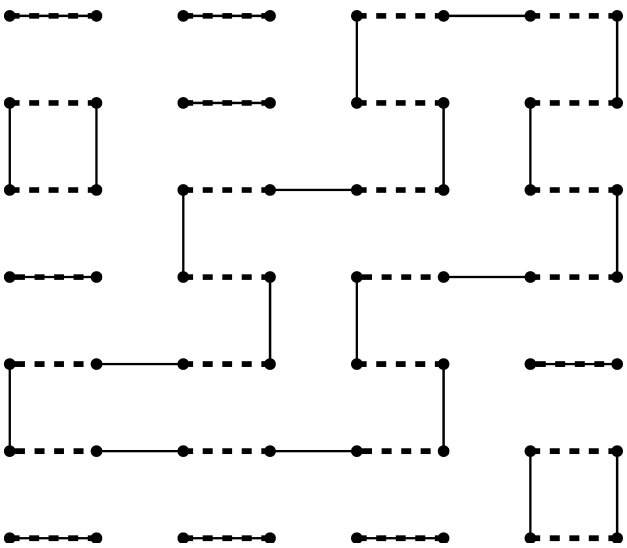

Fig. 21. By superimposing a dimer configuration (dark lines) on a reference dimer configurations (dotted lines), it can be represented by the depicted multi-loop which consists of alternating solid and dotted lines.

the triangular lattice [22] does this for the $d = 1, k = 1$ theory. Some ideas about implementing the higher-k theories in dimer models are discussed in [66].

Josephson junction models [39,65] admit a similar description. In these models, superconducting islands (which, in the simplest incarnation, live on the links of a lattice) are connected by tunneling junctions. The charge is effectively restricted to take the values 0,1, which correspond to the presence or absence of a curve on that link. The Hamiltonian further requires these curves to be connected and non-terminating. More elaborate models allow for charges $1, 2, \ldots, N$, which can be mapped to labeled curves. These may be useful for the implementation of $k > 1$ theories since a curve carrying the label n might be a way of representing n curves.

Of course, finding a model with a loop gas representation is only the first step. The model must also: (a) assign a value of d to each closed contractible loop and (b) enforce the associated Jones–Wenzl projector relation. It might seem improbable that any realistic Hamiltonian would impose precisely the right value of d, much less impose the corresponding Jones–Wenzl projector, which can be quite complicated, as may be seen from Fig. 16. However, if a model incorporates the correct value of $d = 2\cos(\pi/k + 2)$, the uniqueness of the corresponding Jones–Wenzl projector implies that a generic perturbation stands a good chance of driving the system into the corresponding $SU(2)_k \times \overline{SU(2)_k}$ topological phase.

9. A plasma analogy using loop gases

Some insight into the physics of the wavefunctions discussed above, their representation of the topological structures discussed in this paper, and the difficulties associated with imposing the Jones–Wenzl projector on them with a local perturbation can be gained by constructing a "plasma analogy" for the ground state wavefunctions of our topological field theories. As in the case of Laughlin's plasma analogy for Abelian quantum Hall states, the idea is to map the squared norm of the quantum ground state in a 2+1-dimensional system to the partition function of a classical 2-dimensional system. Equal-time correlation functions in the former will be correlation functions in the latter. However, unlike in Laughlin's plasma analogy, in which the squared norms of ground states were mapped to the partition functions of plasmas, we map them to loop gases. In order to effect this mapping, we will need to consider classical models on lattices. Such a short-distance regularization may seem unnatural and unnecessary from the point of view of the preceding discussion of doubled Chern–Simons theories. However, it will prove to be very natural when we turn to 'microscopic' lattice models which support topological phases.

Let us consider the unnormalized ground state on the sphere for a given value of d. It is given by

$$|0\rangle = \sum_{\{\gamma\}} d^{n_c(\gamma)} |\{\gamma\}\rangle, \tag{152}$$

where $n_c(\gamma)$ is the number of loops in the configuration γ. Its norm is

$$\langle 0|0\rangle = \sum_{\{\gamma\}} d^{2n_C(\gamma)} \langle\{\gamma\}|\{\gamma\}\rangle. \tag{153}$$

In other words, it is a sum over all possible loop configurations, weighted by a loop fugacity d^2.

Two types of classical statistical mechanical models have very similar sums appearing in their partition functions, the ferromagnetic q-state Potts model and the O(n) model. The details of these mappings can be found in [55] and (the former) in [56], nevertheless we shall briefly review them here.

We consider first a set of $O(n)$ models which give rise to loop gases [57]. These models have partition function

$$Z = \int \prod_i \frac{\mathrm{d}\hat{S}_i}{K_N} \mathrm{e}^{-\beta H}, \tag{154}$$

where the $\hat{S}_i$s are N-dimensional unit vectors, K_N is the surface area of the N-sphere, and

$$-\beta H = \sum_{\langle i,j\rangle} \ln(1 + x\hat{S}_i \cdot \hat{S}_j) \tag{155}$$

or, simply,

$$\mathrm{e}^{-\beta H} = \prod_{\langle i,j\rangle} (1 + x\hat{S}_i \cdot \hat{S}_j). \tag{156}$$

While the Hamiltonian can take unphysical imaginary values for $x > 1$, the model is still perfectly well-defined. Let us work with this model defined on the honeycomb lattice.

The high-temperature series expansion of this model is obtained by expanding the product (156). Consider any given term in this expansion. For each link on the lattice, there will either be the factor 1 or $x\hat{S}_i \cdot \hat{S}_j$. If it's the latter, then this term will vanish upon integration over $\hat{S}_i$ and $\hat{S}_j$ unless there is another factor in the term which contains $\hat{S}_i$ and a factor containing $\hat{S}_j$. Suppose the factor fulfilling the former requirement is $x\hat{S}_i \cdot \hat{S}_k$. Then a factor containing $\hat{S}_k$ must also be present. Continuing in this way, we see that if there are any bonds of the lattice for which factors of the form $x\hat{S}_i \cdot \hat{S}_j$ appear rather than 1, then these bonds must form closed loops or else the term will vanish. Clearly, there is no requirement of close-packing. There is one non-vanishing term in the expansion in which there are no bonds. By choosing the honeycomb lattice, we have ensured that the loops cannot cross.

For each vertex on such a loop, we have a factor of

$$\int \frac{\mathrm{d}S_i^\alpha}{K_N} S_i^\alpha S_i^\beta = \frac{1}{N} \delta^{\alpha\beta}. \tag{157}$$

Thus, for each closed loop, we obtain a factor of

$$x^k \sum_{\alpha_1,\alpha_2,\ldots,\alpha_k} \frac{1}{N} \delta^{\alpha_1\alpha_2} \cdot \frac{1}{N} \delta^{\alpha_2\alpha_3} \cdots \frac{1}{N} \delta^{\alpha_k\alpha_1} = \left(\frac{x}{N}\right)^k N. \tag{158}$$

Hence, the partition function is

$$Z = \sum_G \left(\frac{x}{N}\right)^b N^\ell, \tag{159}$$

where b is the number of bonds and ℓ is the number of loops. In this way, for $x = N$ and $N = d^2$, we have a statistical mechanical model whose partition function is the squared norm of the ground state of $SU(2)_k \times \overline{SU(2)_k}$ Chern–Simons Gauge Theory.

The q-state Potts models also have loop gas representations, but, as we will see, the loops are fully packed in this case. With no wiggle room, isotopy is impossible. It is conceivable that this makes life easier since there is no need to impose isotopy invariance as a condition on low-energy states. Thus, it may prove useful to consider microscopic models of this form.

The Hamiltonian of the ferromagnetic q-state Potts model is given by

$$-\beta H = J \sum_{\langle i,j \rangle} \delta_{\sigma_i \sigma_j}, \tag{160}$$

where $\sigma_i = 0, 1, \ldots q-1, J > 0$. With the help of the identity $\exp(J\delta_{\sigma_i\sigma_j}) = 1 + v\delta_{\sigma_i\sigma_j}$ where $v = \exp(J) - 1$, the partition function for for this model can be written as follows:

$$Z_{\text{Potts}} \equiv \sum_{\{\sigma\}} \mathrm{e}^{-\beta H} = \sum_{\{\sigma\}} \prod_{\langle i,j \rangle} \left(1 + v\delta_{\sigma_i\sigma_j}\right). \tag{161}$$

Expanding the product in (161) can be interpreted graphically: every time $v\delta_{\sigma_i\sigma_j}$ is chosen for a pair of neighboring sites i and j, the corresponding bond is colored; the choice of 1 results in an empty bond. Due to Kronecker δ-symbols, all sites belonging to the same cluster must have identical values of spins σ. Summing over all possible spin configurations $\{\sigma\}$ we then obtain

$$Z_{\text{Potts}} = \sum_{\mathscr{G}} v^b q^c, \tag{162}$$

where b is the total number of occupied bonds and c is the number of clusters (including isolated sites). The sum is now performed over all configurations $\mathscr{G}$ of such clusters. This is a so-called Fortuin–Kasteleyn or random cluster representation [58].

A so-called polygon decomposition [59] lets us relate this to a loop gas on the surrounding lattice (vertices of the surrounding lattice are the midpoints of the original bonds). We then think of an occupied bond as a double-sided mirror placed at the site of the surrounding lattice. If a bond is not occupied, then its dual bond is considered a mirror. Thus every site of the surrounding lattice gets one of the two possible mirrors. We then use these mirrors to construct paths as shown in Fig. 22. Since these paths have no sources or sinks, they always form loops that either surround the clusters or are contained inside clusters (in the latter case, the loops can be thought of as surrounding *dual* clusters). The number of loops ℓ is given by $\ell = c + f$ where f is the number of faces, i.e., the minimum number of occupied bonds which have to be cut in order to make each cluster tree-like (essentially the number of "voids" which are completely contained within clusters). If we use the Euler relation,

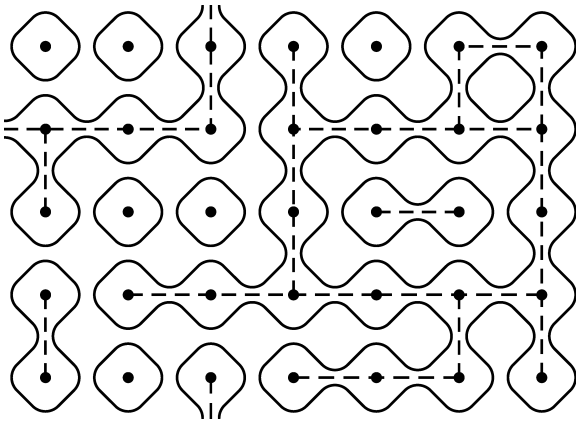

Fig. 22. A typical cluster configuration for the Potts model is shown by dashed lines. Spins belonging to the same cluster take the same value, which must be summed over the q possible values, as described in the text. Clusters can be represented by loops on the surrounding lattice, shown by solid lines.

$$b + c - f = \text{const.} \tag{163}$$

then $\ell = 2c + b$, or $c = (\ell - b)/2$. Hence, we can rewrite Z_{Potts} as

$$Z_{\text{Potts}} = \sum_G v^b q^{(\ell-b)/2} = \sum_G \left(\frac{v}{\sqrt{q}}\right)^b \cdot (\sqrt{q})^\ell. \tag{164}$$

If $v = \sqrt{q}$—i.e., if the Potts model is at its self-dual point, then

$$Z_{\text{Potts}}^{\text{Self-Dual}} = \sum_G (\sqrt{q})^\ell. \tag{165}$$

This appears to be a lattice regularization of the norm of the ground state wavefunction above if $d^2 = \sqrt{q}$, apart from the full-packing condition: that is to say that every bond of the surrounding lattice belongs to a loop, no bonds are left empty, as seen in Fig. 22. Since there is no corresponding notion in the continuum, it is not completely clear what the connection is to our earlier discussion of the ground states of doubled Chern–Simons theories.

The $O(N)$ and self-dual Potts models are quite different, but they share one curious feature, which is of interest to us: for $d > 2$, which corresponds to $N > 2$ and $q > 4$, respectively, all correlation functions are short-ranged. The underlying physical reasons are unrelated: the $O(N > 2)$ models do not have an ordered phase by the Mermin–Wagner theorem, so their correlations are short-ranged. In the ordered phase, loops become long and wander around the system, as happens for $N = 1$, $x = 1$, which is deep within the ordered state of the Ising model. For $N = 2$, there is an algebraically ordered phase, which includes $x = 2$, so loops are still able to wander about the system. The $q > 4$ self-dual Potts models, on the other hand, are short-ranged because they have a first-order phase transition. The ordered and disordered phases have only small loops, and loops can grow large and wander only at a

second-order phase transition point, as occurs for $q \leqslant 4$. For $q > 4$, the phase transition is first order, and loops never grow large [60].

The basic phenomenon is that for d large, it is highly favorable to have as many contractible loops as possible. On the lattice, the proliferation of small loops takes up all of the space, leaving no room for loops to stretch and fluctuate. There is no corresponding effect in the continuum, where there is no shortest length.

This makes it difficult for any local term on the lattice to enact the Jones–Wenzl relation on essential loops, i.e., long loops which encircle non-trivial topological features such as handles or quasiparticles or terminate at quasiparticles. Such loops have very low probability of ever approaching each other. In order to circumvent this difficulty, we need to specify the short-distance dynamics of a candidate lattice model so that it more faithfully represents the topological relations of the continuum theory.

10. Lattice models, chiral anomalies, and doubled theories

From the preceding sections, we see that there is potentially a natural relation between doubled Chern–Simons theories and lattice models whose configurations can be represented in terms of loops. In this section, we note, on general grounds, that lattice models should yield doubled, rather than undoubled theories.

Consider a one-dimensional electron system at finite density. Such a system is a metal. In the continuum, the numbers of left- and right-moving electrons are independently conserved at low energies. Their sum is the total charge, ρ, which must be conserved under all circumstance. Their difference is the current, j, which suffers from the chiral anomaly,

$$\partial_\mu j^\mu_D = E, \tag{166}$$

where $j^\mu_D = (j, \rho)$ and E is the electric field. Since an electric field generates a current, the numbers of left- and right- moving electrons are not independently conserved in the presence of an electric field. In particular, in a homogeneous state,

$$\frac{\partial j}{\partial t} = E. \tag{167}$$

In other words, the chiral anomaly is simply the statement that a metal conducts electricity. To be completely precise, it is more than that, since it implies perfect conductivity, but only a little more since in a translationally invariant system in which all particles have the same charge, a metal must be a perfect metal.

In a crystal, which has only discrete translational symmetry, this equation must be modified by the formation of bands. Instead, we have,

$$\begin{aligned} &\frac{\mathrm{d}k}{\mathrm{d}t} = E, \\ &v(k) = \nabla_k \epsilon(k) \end{aligned} \tag{168}$$

with j given by a summation of $v(k)$ over occupied states. Since $\epsilon(k)$ is a periodic function of k, the current cannot increase linearly with E. Averaged over (sufficiently long) time, it is, in fact, constant. Said differently, if we follow the motion of the spectrum, then as right-moving states flow off to the right, just as many left-moving states flow in from the left because the number of these states is conserved and there is no net spectral flow. However, in the continuum, there is an infinite Dirac sea, and right-moving states can flow to the right (and be replenished by the infinite Dirac sea) without a compensating flow of left-moving states.

This is not a great concern in the case of real metals, which are not perfect metals because they are not translationally invariant. They have finite conductivity as a result of scattering by impurities, phonons, etc., so they could never satisfy (166) anyway. However, there is a serious problem in chiral systems, since they must truly be perfect metals. Chirality implies that their currents can not be degraded by backscattering. However, we have just seen that it is not possible to have a perfect metal, in the sense of (166), on the lattice. This implies that a system on a lattice cannot be a chiral metal.

Since there is no chirality in 2+1-dimensions [67], these arguments do not apply directly. However, according to the arguments of Section 6, the ground state wavefunctions of the TQFTs discussed here are related to correlation functions in associated conformal field theories. For chiral TQFTs, these conformal field theories are chiral, and have chiral anomalies. However, such a theory cannot arise from a lattice model. Thus, we do not expect chiral TQFTs to arise from lattice models. Doubled theories are more natural.

This does not mean that chiral theories are impossible on the lattice, just that some way of evading the above logic is needed. In order to have a finite chiral anomaly, an infinite Dirac sea is needed. This is not so artificial in a $2+1-D$ model since the excited states (which go beyond the mapping of the ground state to a $2D$ theory) can serve as the necessary reservoir. Such degrees of freedom would seem to be missing from the pre-Hilbert spaces which we have been using in this paper, but a more general class of models might have the requisite structure. One simple possibility, motivated by the domain-wall fermion proposal for chiral lattice fermions [68], is if the system has a finite extent in a third spatial direction (which one is free to envision as an internal degree of freedom).

11. Discussion

Our purpose in this paper has been to explore the Hilbert spaces of a set of P,T-invariant topological field theories. These Hilbert spaces are most neatly understood in terms of a combinatorial construction which allows the distinct Hilbert spaces of the theories on different manifolds and with different quasiparticle numbers to be discussed in a unified formalism. The power of this construction derives from its reduction of the structure of these Hilbert spaces to a set of *local* rules for multi-curves on surfaces. Aside from its mathematical beauty, this is a Good Thing since it gives us some clues about what types of physical systems—which should, after all, be described by local rules—can manifest phases which are described by these topological

field theories. The two most salient features are that the system should admit a description in terms of configurations of multi-curves—e.g., domain walls, chains of spins, etc.—and that higher-level theories should have longer (but still finite) ranged interactions or larger building blocks (e.g., higher spins).

The essence of these phases is their ability to support excitations with non-trivial braiding statistics, which in almost all cases is non-Abelian. This property makes these phases relevant as a setting for topological quantum computation [35,36]. It also makes them difficult to detect experimentally. Ideally, one would like to create excitations and manipulate them in order to perform braiding operations and Aharonov–Bohm interferometry. A more indirect way may be through the observation of a broken symmetry phase which may be in close proximity to the topological phase of interest. If a finite density of Abelian anyonic excitations is created (by doping an insulator, say) then they will superconduct in zero magnetic field [27,28]. If the same is true with a system of non-Abelian anyonic excitations (see [69] for an argument that it is), then the observation of such an exotic superconducting state may reveal the existence of a nearby topological phase. (Depending on one's perspective, one might argue that one or the other is the more interesting phase.)

Acknowledgments

We thank Paul Fendley, Eduardo Fradkin, and David Thouless for discussions. M.F., C.N., and K.S. gratefully acknowledge the hospitality of the Aspen Center for Physics. C.N. was supported in part by the National Science Foundation under Grant DMR-9983544, the Alfred P. Sloan Foundation, and a visiting professorship at Nihon University.

References

[1] Bosons and fermions are the only one-dimensional representations of the permutation group. In principle, higher-dimensional representations are possible, namely parastatistics. These representations can be interpreted in terms of bosons or fermions with some additional 'hidden' *local* quantum numbers. There is no topological character to these representations, and since these local quantum numbers are presumably measurable, this would not appear to be a useful arena for quantum computation. See H.S. Green, Phys. Rev. 90 (1953) 270;
O. Steinmann, Nuovo Cimento 44 (1966) A755;
P.V. Landshoff, H.P. Stapp, Ann. Phys. 45 (1967) 72;
Y. Ohnuki, S. Kamefuchi, Phys. Rev. 170 (1968) 1279;
Ann. Phys. 51 (1969) 337;
S. Doplicher, R. Haag, J. Roberts, Commun. Math. Phys. 23 (1971) 199;
Commun. Math. Phys. 35 (1974) 49;
A.M.L. Messiah, O.W. Greenberg, Phys. Rev. B 136 (1964) 248;
Phys. Rev. B 138 (1965) 1155;
J.B. Hartle, J.R. Taylor, Phys. Rev. 178 (1969) 2043;
R.H. Stolt, J.R. Taylor, Phys. Rev. D 1 (1970) 2226;
Nucl. Phys. B 19 (1970) 1;
J.B. Hartle, R.H. Stolt, J.C. Taylor, Phys. Rev. D 2 (1970) 1759.

[2] J. Leinaas, J. Myrheim, Nuovo Cimento B 37 (1977) 1;
G. Goldin, R. Menikoff, D. Sharp, J. Math. Phys. 22 (1981) 1664;
F. Wilczek, Phys. Rev. Lett. 48 (1982);
F. Wilczek, Phys. Rev. Lett. 49 (1982) 957, see also the discussion and several of the reprints in F. Wilczek (Ed.) Fractional Statistics and Anyon Superconductivity, World Scientific, Singapore 1990.
[3] D.C. Tsui, H.L. Stormer, A.C. Gossard, Phys. Rev. Lett. 48 (1982) 1559.
[4] R.B. Laughlin, Phys. Rev. Lett. 50 (1983) 1395.
[5] B.I. Halperin, Phys. Rev. Lett. 52 (1984) 1583.
[6] D. Arovas, J.R. Schrieffer, F. Wilczek, Phys. Rev. Lett. 53 (1984) 722.
[7] R.L. Willet, J.P. Eisenstein, H.L. Stormer, D.C. Tsui, A.C. Gossard, J.H. English, Phys. Rev. Lett. 59 (1987) 1779.
[8] W. Pan, J.-S. Xia, V. Shvarts, D.E. Adams, H.L. Stormer, D.C. Tsui, L.N. Pfeiffer, K.W. Baldwin, K.W. West, Phys. Rev. Lett. 83 (1999) 3530;
W. Pan, H.L. Stormer, D.C. Tsui, L.N. Pfeiffer, K.W. Baldwin, K.W. West, Sol. State Commun. 119 (2001) 641.
[9] G. Moore, N. Read, Nucl. Phys. B 360 (1991) 362.
[10] M. Greiter, X.G. Wen, F. Wilczek, Nucl. Phys. B 374 (1992) 567.
[11] C. Nayak, F. Wilczek, Nucl. Phys. B 479 (1996) 529.
[12] E.H. Rezayi, N. Read, Phys. Rev. B 56 (1996) 16864.
[13] E. Fradkin, C. Nayak, A. Tsvelik, F. Wilczek, Nucl. Phys. B 516[FS] (1998) 704.
[14] J.P. Eisenstein, K.B. Cooper, L.N. Pfeiffer, K.W. West, Phys. Rev. Lett. 88 (2002) 076801.
[15] E.H. Rezayi, N. Read, Phys. Rev. B 59 (1999) 8084;
V. Gurarie, E.H. Rezayi, Phys. Rev. B 61 (2000) 5473.
[16] E. Ardonne, K. Schoutens, Phys. Rev. Lett. 82 (1999) 5096;
E. Ardonne, S. Guruswamy, K. Schoutens, P. Bouwknegt, Phys. Rev. B 61 (2000) 10298;
E. Ardonne, N. Read, E. Rezayi, K. Schoutens, Nucl. Phys. B 607 (2001) 549.
[17] A. Cappelli, L.S. Georgiev, I.T. Todorov, Nucl. Phys. B 599 (2001) 499;
Commun. Math. Phys. 205 (1999) 657.
[18] X.G. Wen, Int. J. Mod. Phys. B 4 (1990) 239.
[19] X.G. Wen, Q. Niu, Phys. Rev. B 41 (1990) 9377.
[20] K. Walker, unpublished notes.
[21] S.A. Kivelson, D.S. Rokhsar, J.P. Sethna, Phys. Rev. B 35 (1987) 8865;
D.S. Rokhsar, S.A. Kivelson, Phys. Rev. Lett. 61 (1988) 2376;
S.A. Kivelson, Phys. Rev. B 39 (1989) 259.
[22] R. Moessner, S.L. Sondhi, Phys. Rev. Lett. 86 (2001) 1881.
[23] R. Jalabert, S. Sachdev, Phys. Rev. B 44 (1991) 686;
S. Sachdev, M. Vojta, J. Phys. Soc. Jpn. 69 (Suppl. B) (2000) 1.
[24] G. Kuperberg, Commun. Math. Phys 180 (1996) 109.
[25] S. Elitzur, G. Moore, A. Schwimmer, N. Seiberg, Nucl. Phys. B 326 (1989) 108.
[26] E. Witten, Commun. Math. Phys. 121 (1989) 351.
[27] R. Laughlin, Phys. Rev. Lett. 60 (1988) 2677;
Science 242 (1988) 525;
A. Fetter, C. Hanna, R. Laughlin, Phys. Rev. B 39 (1989) 9679.
[28] Y. Chen, F. Wilczek, E. Witten, B. Halperin, Int. J. Mod. Phys. B 3 (1989) 1001.
[29] L. Balents, M.P.A. Fisher, C. Nayak, Int. J. Mod. Phys. B 12 (1998) 1033;
Phys. Rev. B 60 (1999) 1654;
Phys. Rev. B 61 (2000) 6307.
[30] T. Senthil, M.P.A. Fisher, Phys. Rev. B 62 (2000) 7850.
[31] P.W. Anderson, Science 235 (1987) 1196.
[32] N. Read, B. Chakraborty, Phys. Rev. B 40 (1989) 7133.
[33] N. Read, S. Sachdev, Phys. Rev. Lett. 66 (1991) 1773;
N. Read, S. Sachdev, Int. J. Mod. Phys. B 5 (1991) 219.
[34] X.G. Wen, Phys. Rev. B 44 (1991) 2664.

[35] A. Kitaev, Ann. Phys. 303 (2003) 2.
[36] M.H. Freedman, quant-ph/0003128 and references therein.
[37] N.E. Bonesteel, quant-ph/0006092.
[38] C. Nayak, K. Shtengel, Phys. Rev. B 64 (2001) 064422.
[39] L.B. Ioffe, M.V. Feigel'man, A. Ioselevich, D. Ivanov, M. Troyer, G. Blatter, Nature 415 (2002) 503.
[40] V. Drinfeld, Quantum groups, in: Proceedings of ICM, vols. 1 and 2, Berkeley, CA, 1986, pp. 798–820, AMS, 1987.
[41] B.I. Halperin, Helv. Phys. Acta 56 (1983) 75.
[42] X.G. Wen, A. Zee, Phys. Rev. Lett. 69 (1992) 1811;
Phys. Rev. B 47 (1993) 2265.
[43] E. Fradkin, F.A. Schaposnik, Phys. Rev. Lett. 66 (1991) 276.
[44] *Isotopy* is similar to but not exactly the same as *homotopy*, which is a familiar concept from the theory of topological defects (see, for instance, N.D. Mermin, Rev. Mod. Phys. 51 (1979) 591). Two multi-curves, $x_1(t)$ and $x_2(t)$, embedded in the manifold Y—i.e., $x_{1,2}(t) \in Y$ for $t \in [0, 2\pi]$—are said to be homotopic if there is a family of maps $H_s(t)$ with $s \in [0, 1]$ such that $H_s(t) \in Y$, $H_0(t) = x_1(t)$, and $H_1(t) = x_2(t)$. If $H_s(t)$ satisfies the stronger condition that it is one-to-one, onto, continuous, and has a continuous inverse for all $s \in [0, 1]$ then $x_1(t)$ and $x_2(t)$ are said to be isotopic.
[45] Such a Hilbert space representation has been used in the context of quantum gravity. The similarity is greatest in 2+1 dimensions. See A. Ashtekar, V. Husain, C. Rovelli, J. Samuel, L. Smolin, Class. Quant. Grav. 6 (1989) 185;
See also L. Smolin, Mod. Phys. Lett. A 4 (1989) 1091.
[46] S. Lang, Algebra, third ed., Springer, Berlin, 2002, GTM 211.
[47] W. Goldman, Invent. Math. 85 (1986) 263.
[48] L.H. Kauffman, Topology 26 (1987) 395.
[49] C. Blanchet, N. Habegger, G. Masbaum, P. Vogel, Topology 34 (1995) 883.
[50] V.G. Turaev, Quantum Invariants of Knots and 3-Manifolds, Walter de Gruyter, Berlin, New York, 1994.
[51] M. Freedman, et al., in preparation.
[52] Physically measurable correlation functions in 2D statistical mechanical models always combine the conformal blocks of the holomorphic and antiholomorphic parts of the theory in just such a way that the result is single-valued.
[53] P. Goddard, D. Olive, in: M. Green, D. Gross (Eds.), Workshop on Unified String Theories, World Scientific, Singapore, 1986.
[54] J.B. Kogut, Rev. Mod. Phys. 51 (1979) 659.
[55] B. Nienhuis, in: C. Domb, J.L. Lebowitz (Eds.), Phase Transitions and Critical Phenomena, vol. 11, Academic Press, London, 1987.
[56] R.J. Baxter, Exactly Solved Models in Statistical Mechanics, Academic Press, New York, 1982.
[57] E. Domany, D. Mukamel, B. Nienhuis, A. Schwimmer, Nucl. Phys. 190 (1981) 279.
[58] C.M. Fortuin, P.W. Kasteleyn, Physica 57 (1972) 536.
[59] R.J. Baxter, S.B. Kelland, F.Y. Wu, J. Phys. A 9 (1976) 397.
[60] For the $q > 4$ Potts model this is easily remedied, however. A first order transition becomes continuous after the introduction of an arbitrary small disorder [61,62]. The size of the clusters diverges at the self-dual point [63,64] and therefore the surrounding loops remain long.
[61] M. Aizenman, J. Wehr, Phys. Rev. Lett. 62 (1989) 2503.
[62] K. Hui, A.N. Berker, Phys. Rev. Lett. 62 (1989) 2507.
[63] J. Cardy, J.L. Jacobsen, Phys. Rev. Lett. 79 (1997) 4063;
J. Cardy, cond-mat/9806355, 1998.
[64] L. Chayes, K. Shtengel, Commun. Math. Phys. 204 (1999) 353.
[65] T. Senthil, O. Motrunich, Phys. Rev. B 66 (2002) 205104;
O.I. Motrunich, T. Senthil, Phys. Rev. Lett. 89 (2002) 277004.
[66] M. Freedman, C. Nayak, K. Shtengel, cond-mat/0309120.
[67] However, these arguments do apply to four dimensions. If a magnetic field is imposed, Landau levels form, and there is one-dimensional dispersion in the direction parallel to the magnetic field, and one

can define right- and left-moving fields with respect to this direction. Therefore, an electric field parallel to the magnetic field causes a flow of current in this direction, and there is a chiral anomaly proportional to $E \cdot B$.
[68] D. Kaplan, Phys. Lett. B 288 (1992) 342.
[69] A. Cappelli, P. Valtancoli, Nucl. Phys. B 453 (1995) 727.