Two-dimensional quantum Yang–Mills theory with corners

2008 J. Phys. A: Math. Theor. 41 135401

(http://iopscience.iop.org/1751-8121/41/13/135401)

IOP PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND THEORETICAL

J. Phys. A: Math. Theor. **41** (2008) 135401 (20pp) doi:10.1088/1751-8113/41/13/135401

Two-dimensional quantum Yang–Mills theory with corners

Robert Oeckl

Instituto de Matemáticas, UNAM, Campus Morelia, CP 58190, Morelia, Michoacán, Mexico

E-mail: robert@matmor.unam.mx

Received 8 November 2007
Published 14 March 2008
Online at stacks.iop.org/JPhysA/41/135401

Abstract
The solution of quantum Yang–Mills theory on arbitrary compact two-manifolds is well known. We bring this solution into a TQFT-like form and extend it to include corners. Our formulation is based on an axiomatic system that we hope is flexible enough to capture actual quantum field theories also in higher dimensions. We motivate this axiomatic system from a formal Schrödinger–Feynman quantization procedure. We also discuss the physical meaning of unitarity, the concept of vacuum, (partial) Wilson loops and non-orientable surfaces.

PACS numbers: 11.15.−q, 02.40.Ma, 03.70.+k
Mathematics Subject Classification: 81T13, 57R56, 81P15, 81T40

1. Introduction

The subject of two-dimensional quantum Yang–Mills theory is an old one. Solvability of the theory was already shown by Migdal [1], using the lattice approach. The advent of topological quantum field theory (TQFT) and related ideas generated interest in this two-dimensional theory from a new perspective. The theory was formulated and solved on arbitrary (compact) two-manifolds with boundaries [2–5]. In the present paper we wish to carry this one step further, namely by allowing generalized manifolds which may have corners. Roughly, this means that the boundaries are not necessarily closed, but may have boundaries themselves. While there is already considerable work on TQFT with corners, usually following Walker [6], this generally does not extend to the situation where the vector spaces associated with boundaries are infinite dimensional. We show that two-dimensional quantum Yang–Mills theory provides a realization for a TQFT-type system of axioms that admits both infinite dimensionality of vector spaces and manifolds with corners.

Another motivation for the present work comes from the general boundary formulation of quantum mechanics [7–10]. The inclusion of corners in this framework, which is based on a specific TQFT-type system of axioms, is an outstanding problem; see the discussion in [10]. Two-dimensional quantum Yang–Mills theory being an actual quantum mechanical system should thus serve as an example of a 'physically correct' implementation of corners. In particular, this implementation must be compatible with the probability interpretation as outlined in [10]. Indeed, the version of two-dimensional quantum Yang–Mills theory constructed in this paper is unitary in the extended sense of [10] and hence compatible with this probability interpretation. In particular, we show how corners play a role in the deformation of regions and allow us to formulate the associated probability conservation condition.

The axioms presented in this paper are a direct generalization of those presented in [10]. We comment both on the mathematical as well as the physical motivation for the specific type of generalization we perform. As in the case without corners, the physical justification for the axioms comes from a simple quantization prescription, combining the Schrödinger representation with the Feynman path integral.

The additional topics we cover are the axiomatization and realization of the concept of vacuum, the inclusion of Wilson loops and the extension to non-orientable manifolds.

Section 2 is concerned with the axiomatic system, its mathematical and physical motivations, and a comparison to the axiomatic system of [10]. Section 3 elaborates on the two-dimensional case, identifying elementary data in this context. Section 4 then develops two-dimensional quantum Yang–Mills theory as a realization of the axioms. Section 5 extends this by including the concept of vacuum, Wilson loops and non-orientable surfaces. A closing section presents a brief outlook.

2. The axiomatic system

2.1. Mathematical motivation

We provide motivations for a specific set of axioms (to be introduced subsequently) that might broadly be identified as describing a type of topological quantum field theory. Note that the attribute 'topological' does not necessarily mean that we consider topological manifolds only, although we shall initially do so.

Recall the basic setup of a topological quantum field theory (TQFT) [11]. We associate finite-dimensional vector spaces $\mathcal{H}_\Sigma$ with $(n-1)$-dimensional manifolds Σ and maps between these vector spaces with n-dimensional cobordisms. An n-dimensional cobordism is an n-dimensional manifold M with a boundary Σ so that the boundary is the disjoint union of two $(n-1)$-manifolds $\Sigma = \Sigma_1 \cup \Sigma_2$. Thus we associate with M a linear map $\mathcal{H}_{\Sigma_1} \to \mathcal{H}_{\Sigma_2}$, declaring Σ_1 to be the 'in'-component and Σ_2 to be the 'out'-component of the boundary. A key requirement is then that the map associated with a cobordism that arises as the gluing of two cobordisms is the composition of the maps associated with the glued cobordisms. This allows a functorial formulation of TQFT, i.e., as a functor from the category of $(n-1)$-manifolds and n-cobordisms to the category of finite-dimensional vector spaces and linear maps.

This manner of axiomatizing TQFT faces limitations once we wish to consider infinite-dimensional vector spaces (as becomes necessary if we want to describe real quantum field theories). For example, let $n = 2$ and consider a cylinder as a cobordism between two circles. Associated with this is a map from the vector space for a circle to itself. This map is in fact the identity[1]. Hence, gluing the two circles together to form a torus yields the trace of the identity map. This is the dimension of the vector space and hence infinite if it is infinite

[1] Note that this does not follow from what we have said so far. In general, this map would be a projector. However, without loss of information we might restrict the state space to its domain, making it the identity. This is generally done in TQFT.

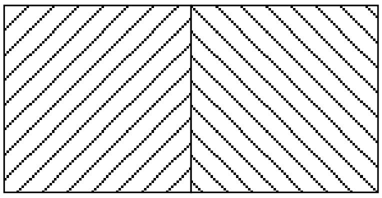

Figure 1. Gluing two rectangles to form a new rectangle.

dimensional. We could avoid this kind of problem by restricting the class of admissible closed n-manifolds.

However, a related problem cannot be eliminated in this way. Namely, in general there are many ways to arrange the connected components of the boundary of an n-manifold into an 'in'- and an 'out'-boundary. Not all of these would generally lead to well-defined maps. In particular, choosing the whole boundary to be 'out' will generally not lead to a well-defined map (due to the same type of infinities as above). We avoid this problem by always taking the whole boundary to be 'in'. Of course, this means that we have to reformulate the correspondence between gluing and composition, loosing its simple functorial formulation.

We also require both the n-manifolds and $(n-1)$-manifolds under consideration to be oriented. Furthermore, reversal of orientation of an $(n-1)$-manifold corresponds to dualization of the associated vector space. Hence, we want the bi-dual space to be isomorphic to the original one. To achieve this in the infinite-dimensional situation we add structure and consider Hilbert spaces. The use of Hilbert spaces has another essential reason, namely the applicability of the physical interpretation in terms of quantum mechanics and probabilities.

The points discussed so far are implemented in the axiomatic system presented in [10] together with an extended quantum mechanical probability interpretation. Furthermore, it was shown in [12, 13] that the Klein–Gordon quantum field theory satisfies these axioms and the associated physical interpretation at least for certain special classes of manifolds. A crucial element that is missing so far, but is desirable from a physics point of view (see the discussion in [10]) are *corners*.

Let us mention first that the manifolds in question may carry structure in addition to being topological manifolds. For the moment we shall only be interested in the case of no additional structure and that of differentiable structure. Before we proceed we change our terminology slightly: in the following we refer to the oriented cobordisms or n-manifolds as *regions* and to the oriented $(n-1)$-manifolds as *hypersurfaces*.

Now imagine that we want to glue two regions with the shape of solid rectangles along one side to form a new region with the shape of a solid rectangle (see figure 1). This situation is not covered by the standard TQFT axioms. Consider first the differentiable situation: the regions in question are not even differentiable manifolds. To admit them we have to generalize the definition of a region. Rather than being manifolds with a boundary, they will be manifolds with boundaries and *corners*. In the present example it is quite clear what this means, but we deliberately avoid a precise general definition here. On the other hand, the boundary of a rectangle is still a valid differentiable 1-manifold and seen as a part of it the corners become 'invisible'. However, they do play a role in the gluing as only a part of the boundary is glued. Moreover, this part is not a connected component as required by standard axioms. Rather it is itself bounded by two corners. If we consider topological manifolds, the rectangles are homeomorphic to discs and as such perfectly well-defined topological 2-manifolds with a boundary. However, the corners still make their appearance as boundaries of parts of boundaries along which regions are glued. Indeed, without them it would be impossible to glue two discs to a disc.

We will refer to corners that are not explicitly visible in the differentiable or topological structure as *virtual corners*. Hence, in a topological context, to which we restrict ourselves in the following, all corners in the boundaries of regions are virtual. We try to implement them in a minimalistic fashion, complicating the axioms as little as possible. To this end we introduce a concept of *decomposition* of a hypersurface into hypersurfaces as follows. (One might conversely think of this as providing a notion of gluing of hypersurfaces.) Hypersurfaces are thus oriented topological manifolds of dimension $n-1$ with a boundary, and regions are oriented topological manifolds of dimension n with a boundary. A decomposition of a hypersurface Σ is the presentation of Σ as a finite union of hypersurfaces $\Sigma_1, \ldots, \Sigma_n$ with the following properties. Each Σ_i is closed in Σ and the intersection of any Σ_i with any Σ_j is the intersection of their boundaries. Note that this definition of decomposition includes as a special case a decomposition into disjoint components. In addition to the manifest corners of a hypersurface, namely its boundary, a decomposition provides additional virtual corners. These are the boundaries of the pieces of the decomposition which are not already contained in the boundary of the original hypersurface.

Before presenting the generalization of the system of axioms of [10], we mention another modification which was already discussed in [10]. Namely, to simplify the axioms it is convenient to introduce *empty regions*. Essentially, these are oriented topological manifolds of dimension $n-1$ with a boundary, as are the hypersurfaces. However, they should be thought of as regions completely contracted to their boundary. Hence, the boundary of an empty region is defined to be the union of two copies of it as a hypersurface, but with opposite orientations. Furthermore, these copies are glued along their boundaries, providing a hypersurface with decomposition.

2.2. *Core axioms*

We are now ready to list the axioms. If a hypersurface is denoted by Σ, its oppositely oriented version is denoted by $\bar{\Sigma}$.

- (T1) Associated with each hypersurface Σ is a complex separable Hilbert space $\mathcal{H}_\Sigma$, called the *state space* of Σ. We denote its inner product by $\langle\cdot,\cdot\rangle_\Sigma$.
- (T1b) Associated with each hypersurface Σ is an antilinear isomorphism $\iota_\Sigma : \mathcal{H}_\Sigma \to \mathcal{H}_{\bar{\Sigma}}$. This map is an involution in the sense that $\iota_{\bar{\Sigma}} \circ \iota_\Sigma$ is the identity on $\mathcal{H}_\Sigma$.
- (T2) Suppose the hypersurface Σ decomposes into a union of hypersurfaces $\Sigma_1 \cup \cdots \cup \Sigma_n$. Then, there is a bounded surjective map of state spaces $\tau : \mathcal{H}_{\Sigma_1} \otimes \cdots \otimes \mathcal{H}_{\Sigma_n} \to \mathcal{H}_\Sigma$. Furthermore, the restriction of τ to the orthogonal complement of its kernel preserves the inner product, i.e., is an isomorphism of Hilbert spaces. If the decomposition is disjoint τ is also injective. The composition of the maps τ associated with two consecutive decompositions is identical to the map τ arising from the resulting decomposition.
- (T2b) The involution ι is compatible with the above decomposition. That is, $\tau \circ (\iota_{\Sigma_1} \otimes \cdots \otimes \iota_{\Sigma_n}) = \iota_\Sigma \circ \tau$.
- (T4) Associated with each region M is a linear map from the state space of its boundary Σ (with induced orientation) to the complex numbers, $\rho_M : \mathcal{H}_\Sigma \to \mathbb{C}$. This is called the *amplitude* map.
- (T3x) Suppose M is an empty region. Then its boundary can be decomposed into two components that are identical up to orientation, $\bar{\Sigma} \cup \Sigma$. The extended amplitude map $(\cdot,\cdot)_\Sigma := \rho_M \circ \tau$ defines a bilinear pairing $\mathcal{H}_{\bar{\Sigma}} \otimes \mathcal{H}_\Sigma \to \mathbb{C}$. We require this

pairing to be compatible with the involution and Hilbert space structure in the sense that $\langle\cdot,\cdot\rangle_\Sigma = (\iota_\Sigma(\cdot),\cdot)_\Sigma$.

(T4b) Suppose M is a region with a boundary Σ, decomposable into the union of two components, $\Sigma = \Sigma_1 \cup \Sigma_2$. Suppose the extended amplitude map $\rho_M \circ \tau : \mathcal{H}_{\Sigma_1} \otimes \mathcal{H}_{\Sigma_2} \to \mathbb{C}$ gives rise to an isomorphism of vector spaces $\tilde{\rho}_M : \mathcal{H}_{\Sigma_1} \to \mathcal{H}_{\bar{\Sigma}_2}$. Then we require $\tilde{\rho}_M$ to preserve the inner product, i.e., be *unitary*.

(T5) Let M_1 and M_2 be two regions such that the union $M = M_1 \cup M_2$ is again a region and the intersection is a hypersurface Σ. The boundary of M_1 may be decomposed into $\Sigma_1 \cup \Sigma$ and the boundary of M_2 into $\Sigma_2 \cup \bar{\Sigma}$. Let $\{\xi_i\}_{i\in I}$ be an ON-basis of $\mathcal{H}_\Sigma$. If

$$\sum_{i\in I} \rho_{M_1} \circ \tau_1(\cdot \otimes \xi_i)\rho_{M_2} \circ \tau_2(\cdot \otimes \xi_i^*)$$

exists then we require it to be equal to $\rho_M \circ \tau(\cdot \otimes \cdot)$.

The strange seeming numbering of the axioms is provided merely for easier comparison to [10]. Let us briefly perform such a comparison. As already mentioned, we admit empty regions here which allows us to remove axioms (T3) and (T3b) of [10] and replace them with a single axiom (T3x). This appears now after (T4), however, as it requires the amplitude map to be defined. A further change is the formulation of the gluing (T5) via the insertion of an ON-basis instead of a functorial formulation (conditional on existence). This avoids the artificial introduction of an ‘in’/‘out’ splitting of boundaries. We also require consecutive decompositions of hypersurfaces to yield the same map τ as the resulting decomposition. While this was implicitly understood it is now written explicitly in (T2).

We now turn to the modification of the axioms mandated by the implementation of corners. This modification is effected by replacing the concept of decomposition of hypersurfaces from the more special one admitting only decompositions into connected components to the more general one defined above. Hence, the modification of the axioms is mostly implicit. This is the case for axioms (T2b), (T4b) and (T5). The single axiom where the implementation of corners is more explicit is axiom (T2). Here, we encounter the novel possibility that the map τ associated with a decomposition is not generally an isomorphism. Indeed, it is only an isomorphism in the case that the decomposition is into disjoint components.

If we restrict the concept of decomposition to the disjoint one, we recover the old axioms (although with the changes unrelated to corners mentioned above). In that case (T2) simplifies a little bit, but no other explicit change appears. (One might remove the explicit mention of τ in (T4b) and (T5), but this is more a matter of aesthetics.) In this sense our implementation of corners may be said to be minimalistic. Of course, another implicit change is in the definition of hypersurfaces. In the case without corners we would not admit hypersurfaces to have boundaries.

Besides mathematical minimalism our proposed implementation of corners is motivated from the physical requirement to match actual quantum field theories. In this direction our proposal arises from a Schrödinger–Feynman quantization prescription. This is the subject of the next section.

2.3. *Schrödinger–Feynman quantization and corners*

Recall that topological quantum field theory, even though it may be considered a purely mathematical subject, arose out of methods of quantum field theory and conformal field theory [14]. Roughly speaking, the vector spaces associated with boundaries were thought of as analogs of state spaces of a quantum mechanical system, while the maps between these vector spaces were thought of as analogs of time evolution operators.

The axioms presented in [10] and refined in the previous section may be seen as a late attempt to bring these ideas back into physics, by which we mean here to ordinary quantum field theory. One part of this endeavor is to develop the formal framework, i.e., the set of axioms together with their physical interpretation. Another part of this is to provide concrete theories fitting the framework. Usually, quantum theories are obtained through a process of quantization. Unfortunately, at present there exists no fully satisfactory quantization prescription for the framework considered here. However, a Schrödinger–Feynman quantization (which motivated TQFT originally) works on a formal level and can be made to work at least in some situations of physical interest [8, 12, 13]. We recall its most essential elements from the perspective of the application to our axiomatic system. We give special emphasis to the implementation of corners in this context. For more details (without corners), we refer the reader to [10] or [13].

Consider a classical field theory. Hence, a space K_Σ of (field) configurations is associated with each hypersurface Σ together with a measure on this configuration space. We define the state space $\mathcal{H}_\Sigma$ as the space of complex square integrable functions, called *wave functions*, on K_Σ with the inner product

$$\langle\psi,\psi'\rangle_\Sigma := \int_{K_\Sigma} \mathcal{D}\varphi \overline{\psi(\varphi)}\psi'(\varphi). \tag{1}$$

The involution of axiom (T1b) is simply the complex conjugation of wave functions,

$$(\iota_\Sigma(\psi))(\varphi) := \overline{\psi(\varphi)} \qquad \forall\,\psi\in\mathcal{H}_\Sigma, \varphi\in K_\Sigma. \tag{2}$$

Suppose we have a decomposition of a hypersurface Σ into two components, Σ_1 and Σ_2. This yields an injective map $K_\Sigma \to K_{\Sigma_1}\times K_{\Sigma_2}$ between the associated state spaces by simply forgetting parts of the configuration data. (This supposes the configuration data to be local in a suitable sense.) Hence, we obtain an induced surjective linear map $\tau : \mathcal{H}_{\Sigma_1}\otimes\mathcal{H}_{\Sigma_2}\to\mathcal{H}_\Sigma$ between the associated spaces of wave functions. Explicitly,

$$(\tau(\psi\otimes\eta))(\varphi) = \psi(\varphi|_{\Sigma_1})\eta(\varphi|_{\Sigma_2}) \qquad \forall\,\psi\in\mathcal{H}_{\Sigma_1}, \eta\in\mathcal{H}_{\Sigma_2}, \varphi\in K_\Sigma. \tag{3}$$

If the decomposition is disjoint, no configuration data are 'forgotten' and the maps are bijective. Indeed, the latter is then an isomorphism of Hilbert spaces since the inner product on the tensor product is induced from the inner products on the components. The latter case provided the motivation for the old version of axiom (T2), which is standard in TQFT, while the new version is motivated by the general case with corners implemented through generalized decompositions.

The amplitude (T4) for a region M is given by an integral of the wave function over boundary configurations weighted by a kernel Z_M, called the *field propagator*:

$$\rho_M(\psi) := \int_{K_\Sigma} \mathcal{D}\varphi\, \psi(\varphi) Z_M(\varphi) \qquad \forall\,\psi\in\mathcal{H}_\Sigma. \tag{4}$$

The field propagator in turn is defined through an action S_M, which is a function on a space of configurations K_M on M,

$$Z_M(\varphi) := \int_{K_M,\phi|_\Sigma=\varphi} \mathcal{D}\phi\, \mathrm{e}^{\mathrm{i}S_M(\phi)} \qquad \forall\,\varphi\in K_\Sigma. \tag{5}$$

Supposing the above definitions can be made rigorous, the axioms are then automatically satisfied (except for unitarity). In particular, axiom (T5) follows from formal gluing properties of the path integral. In the case without corners this is detailed in [10] and [13]. In the case with corners, for axiom (T2) this was explained above. The only other axiom were the corners make an essential difference is (T5). Let us thus briefly explain why (T5) still holds.

Assume the context of axiom (T5). Denote the configuration spaces on Σ, Σ_1 and Σ_2 by K, K_1 and K_2, respectively. Denote the configuration space on $\Sigma_1 \cup \Sigma_2$ by K_{12}. Also, denote the configuration spaces on $\Sigma \cup \Sigma_1$ and on $\Sigma \cup \Sigma_2$ by K_{1s} and K_{2s}, respectively. Furthermore, we denote the corners forming the intersection of Σ_1 and Σ_2 by c. Then, by gluing properties of the path integral we have

$$Z_M(\varphi_{12}) = \int_{K,\varphi|_c=\varphi_{12}|_c} \mathcal{D}\varphi Z_{M_1}(\varphi_{12}|_{\Sigma_1} \cup \varphi) Z_{M_2}(\varphi_{12}|_{\Sigma_2} \cup \varphi), \tag{6}$$

where $\varphi_{12} \in K_{12}$ and $\cup$ denotes the joining of configuration data. Note the formal property of the ON-basis $\{\xi_i\}_{i\in I}$ of $\mathcal{H}_\Sigma$,

$$\sum_{i\in I} \xi_i(\varphi_a)\xi_i^*(\varphi_b) = \delta(\varphi_a, \varphi_b) \qquad \forall \varphi_a, \varphi_b \in K. \tag{7}$$

Here ξ_i^* denotes the dual basis element of ξ_i in $\mathcal{H}_{\bar{\Sigma}}$. Combining those properties with the identity

$$\begin{aligned} &\int_{K_{1s}} \mathcal{D}\varphi_{1s} \int_{K_{2s}} \mathcal{D}\varphi_{2s} \delta(\varphi_{1s}|_\Sigma, \varphi_{2s}|_\Sigma) f(\varphi_{1s}, \varphi_{2s}) \\ &\qquad = \int_{K_{12}} \mathcal{D}\varphi_{12} \int_{K,\varphi|_c=\varphi_{12}|_c} \mathcal{D}\varphi f(\varphi_{12}|_{\Sigma_1} \cup \varphi, \varphi_{12}|_{\Sigma_2} \cup \varphi) \end{aligned} \tag{8}$$

for arbitrary $f : K_{1s} \times K_{2s} \to \mathbb{C}$ yields the formula of axiom (T5) as required.

3. Two-dimensional TQFT with corners

We now specialize to the case of two dimensions. For the moment we remain in the setting of topological manifolds and also restrict them to be compact.

Hence, a connected component of a hypersurface is either an oriented closed interval or an oriented circle. For simplicity, we refer to the former object as an *open string* and to the latter as a *closed string* (not mentioning the orientation explicitly). These are the only *elementary hypersurfaces*. A general hypersurface is then simply a finite disjoint union of open and closed strings.

Let us denote the Hilbert spaces associated with the open and closed string by $\mathcal{H}_\mathrm{O}$ and $\mathcal{H}_\mathrm{C}$, respectively. We denote their inner products by $\langle\cdot,\cdot\rangle_\mathrm{O}$ and $\langle\cdot,\cdot\rangle_\mathrm{C}$, respectively. We denote the involutions of (T1b) by $\iota_\mathrm{O} : \mathcal{H}_\mathrm{O} \to \mathcal{H}_{\bar{\mathrm{O}}}$ and $\iota_\mathrm{C} : \mathcal{H}_\mathrm{O} \to \mathcal{H}_{\bar{\mathrm{C}}}$, respectively. The bar indicates that we are considering the strings with opposite orientations.

Let us consider the concept of decomposition of a hypersurface. Axiom (T2) tells us that associated with a decomposition is a map τ. The composition of the maps τ of consecutive decompositions of a hypersurface is the same as the map τ associated with the resulting decomposition. Hence, we only need to consider decompositions with at most two components. What is more, in the case of disjoint decompositions the map τ is simply an identification of the state space with the tensor product of the state spaces of the components. Thus, it is enough to specify τ for decompositions of connected hypersurfaces.

Indeed, it is easy to see that there are merely two *elementary decompositions*. The first one is the decomposition of an open string into two open strings. In that case there are two corners at the two ends of the string to be decomposed and one corner at the point where the string is to be cut. The latter appears in both of the component strings and is a virtual corner with respect to the original string. We denote the induced map by $\tau_\mathrm{OO} : \mathcal{H}_\mathrm{O} \otimes \mathcal{H}_\mathrm{O} \to \mathcal{H}_\mathrm{O}$. Note that the decomposition property implies that τ_OO is *associative*, making $\mathcal{H}_\mathrm{O}$ into an associative algebra.

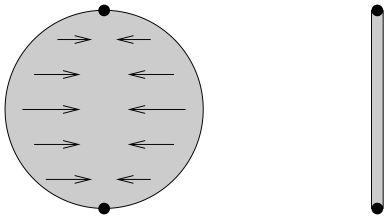

Figure 2. Empty disc and boundary: the boundary of a disc is decomposed into two open strings and the disc is 'squeezed' (left) until the two boundary components coincide (right).

The other elementary decomposition is that of a closed string into an open string. In this case we mark one point of the closed string as a virtual corner and cut it open there. The resulting open string has two copies of this corner as its endpoints. We denote the induced map by $\tau_{\mathrm{OC}} : \mathcal{H}_{\mathrm{O}} \to \mathcal{H}_{\mathrm{C}}$. Note that the composition $\tau_{\mathrm{OC}} \circ \tau_{\mathrm{OO}}$ must be *commutative* due to the lack of natural ordering of the two pieces in the process of cutting a closed string into two pieces.

Compact connected orientable manifolds with a boundary are Riemann surfaces with holes. Thus, a region is simply a finite union of oriented Riemann surfaces with holes. A connected region is then characterized by two non-negative integers, the genus g and the hole number n. In view of the gluing axiom for regions (T4), however, and taking into account the fact that virtual corners allow us to perform rather arbitrary gluings, there is only one *elementary region*. This is the disc. All other regions can be obtained by gluing discs together. We denote the associated amplitude by $\rho_D : \mathcal{H}_{\mathrm{C}} \to \mathbb{C}$.

In the present topological setting a disc is the same as an empty disc. Hence, we can use axiom (T3x) to relate the inner product on the open string state space to the disc amplitude. Namely, we insert two virtual corners on the closed string boundary of the disc and decompose the boundary into two open strings which are then identified up to orientation (squeezing the interior of the disc) (see figure 2). Axiom (T3x) then implies

$$\langle \psi, \eta \rangle_{\mathrm{O}} = \rho_D \circ \tau_{\mathrm{OC}} \circ \tau_{\mathrm{OO}}(\iota_{\mathrm{O}}(\psi) \otimes \eta) \qquad \forall\, \psi, \eta \in \mathcal{H}_{\mathrm{O}}. \tag{9}$$

Note a subtlety here: Formally, the domain of τ_{OO} is $\mathcal{H}_{\mathrm{O}} \otimes \mathcal{H}_{\mathrm{O}}$. However, when shrinking the disc we view the two open string components of the boundary as oppositely oriented and hence denote this domain by $\mathcal{H}_{\bar{\mathrm{O}}} \otimes \mathcal{H}_{\mathrm{O}}$ (composing with ι_{O} then recovers a domain $\mathcal{H}_{\mathrm{O}} \otimes \mathcal{H}_{\mathrm{O}}$). This apparent ambiguity between $\mathcal{H}_{\mathrm{O}}$ and $\mathcal{H}_{\bar{\mathrm{O}}}$ is merely a shortcoming of our notation. To identify a hypersurface as oppositely oriented to an identical copy makes only sense when the two copies are geometrically identified, i.e., 'occupy the same space'.

The inner product on $\mathcal{H}_{\mathrm{C}}$ is completely determined by that on $\mathcal{H}_{\mathrm{O}}$ in combination with the map τ_{OC}; see axiom (T2). Alternatively, the inner product of $\mathcal{H}_{\mathrm{C}}$ is related via (T3x) to the amplitude of an empty cylinder. The latter can be obtained in turn by gluing the empty disc to itself in a suitable way.

In the present topological setting axiom (T4b) is automatically satisfied for the disc with its boundary decomposed into two open strings. Indeed, this follows directly from axiom (T3b) discussed above, using again that the disc is the same as the empty disc. Similarly, axiom (T5) is automatic for gluing two discs to a new disc. This involves again decomposing the boundaries into two open strings (see figure 3). Inserting an ON-basis times its dual into the pair of extended disc amplitudes yields again the extended disc amplitudes. This is obvious from interpreting the extended amplitude as the bilinear form of (T3x). Then, everything descends from extended amplitudes to non-extended amplitudes, yielding (T5).

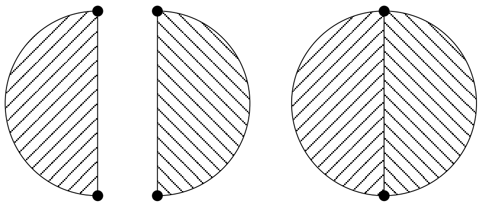

Figure 3. Gluing two discs to one: each disc boundary is decomposed into two open strings (left), then the discs are glued by identifying one of the open strings from each (right).

We have identified elementary data that completely determine a theory satisfying the axioms in the case of two-dimensional compact topological manifolds. However, these data are not free, but subject to several conditions, some of which we have identified.

4. Two-dimensional quantum Yang–Mills theory

Two-dimensional quantum Yang–Mills theory on arbitrary compact surfaces was solved in the early 1990s [2–5]. It is only a small step from there to an explicitly TQFT-like formulation. Our main interest here, however, is the additional step to extend this to the case with corners. As we shall see, this provides a realization of the axioms introduced in section 2.2. To expose the novel aspects of our treatment in detail we proceed in an essentially self-contained fashion.

4.1. Path integral and gauge symmetry

Let G be a compact, connected and simply connected Lie group. The field of classical Yang–Mills theory is given by a connection 1-form A for a principal G-bundle on a manifold M with metric. (We restrict ourselves to the case of the principal bundle being trivial.) The Yang–Mills action is

$$S_M[A] = -\frac{1}{\gamma^2}\int_M \mathrm{tr}(F \wedge \star F), \tag{10}$$

where F is the curvature 2-form of A and γ the (classically irrelevant) coupling constant. If M is two-dimensional, the dependence of $S_M[A]$ on the geometry of M is merely through the area form of the metric[2].

It turns out that the naive version of the Schrödinger–Feynman approach to quantization of section 2.3 is not quite appropriate as we have to take into account gauge symmetry. We shall see how this can be accomplished through a suitable modification of the procedure.

Consider a manifold M with the topology of a disk. Its boundary ∂M is a closed string and we denote a field configuration on it by A_∂. Naively implementing a Schrödinger representation, A_∂ is simply a connection 1-form on ∂M. Formally, the field propagator (5) is thus the path integral

$$Z_M[A_\partial] = \int_{A|_{\partial M}=A_\partial} \mathcal{D}A\, \mathrm{e}^{\mathrm{i}S_M[A]}. \tag{11}$$

Here, the integral is over connection 1-forms A in M which restrict on the boundary to A_∂. Note that the argument of the exponential is imaginary as it should be in quantum theory. Indeed, this will be essential for unitarity and the probability interpretation[3].

[2] Note that this applies whether the metric is Riemannian or pseudo-Riemannian.
[3] This has nothing to do here with the choice of the signature of the metric.

By gauge invariance Z_M can only depend on the holonomy g of the boundary connection A_∂ around the boundary. What is more, Z_M can only depend on the conjugacy class of g (which also makes the choice of the starting point for the holonomy irrelevant). On the other hand, Z_M should be invariant under orientation preserving diffeomorphisms of M that map the boundary to itself. Note that the restriction of such a diffeomorphism to the boundary leaves the conjugacy class of the holonomy g of the boundary connection A_∂ invariant. Hence, the only geometric information relevant to the value of Z_M can be in the total area of M.[4]

Obviously, the value of Z_M should not change whether we introduce a corner (in the differentiable sense) in its boundary or smooth it off, preserving the area. Hence, the differentiable structure of M is expendable as well. This leaves us in an almost topological setting. Thus, a region is a pair (M, s) of a two-dimensional compact oriented topological manifold M with a boundary and a non-negative real number s, the area. If $s = 0$ the region is an empty region. If we glue two regions (M, s) and (N, t), the area of the new region is the sum of the areas, i.e., we get $(M \cup N, s + t)$. A hypersurface does not carry any additional structure. Hence it is simply a one-dimensional compact oriented topological manifold with a boundary.

4.2. Hypersurfaces and state spaces

The above analysis of gauge symmetry tells us that the appropriate reduced configuration space for the closed string is the space of conjugacy classes of G. Thus, the state space $\mathcal{H}_C$ should be the space of functions on it. Before returning to this space, let us consider more general functions on G. The group G has a unique normalized and invariant measure, the Haar measure. This yields the inner product

$$\langle \psi, \eta \rangle = \int \mathrm{d}h \overline{\psi(h)} \eta(h) \tag{12}$$

for functions ψ and η. The set of square integrable functions $\mathcal{C}(G)$ on G becomes a Hilbert space with this inner product. We denote by $\mathcal{C}_{\text{class}}(G)$ the closed subspace of class functions, i.e., functions ψ that are invariant under conjugation, $\psi(g) = \psi(hgh^{-1})$ for all $g, h \in G$. Hence, this subspace can be thought of as a space of functions on the conjugacy classes of G. This provides $\mathcal{H}_C$ and its inner product.

What is the state space for an open string? One way to think about this is to consider the above example of the disc propagator, but think of the boundary as decomposed into several open strings. Obviously, if we know the holonomy along each open string we can calculate the total holonomy and that is all we need. Hence, it is sufficient to associate a group element with each open string representing this holonomy. There are gauge transformations that change the values of these group elements. However, they necessarily change several group elements simultaneously (except if there is only a single one). In considering a single open string alone such gauge transformations cannot be permitted as we do not know about other strings we might want to attach to it. In other words, for determining the configuration space associated with an open string only gauge transformations are relevant that act identical on the ends of the string. Thus, the associated configuration space is the space of elements of G and the state space is the space of functions on it. Summarizing, we obtain

$$\mathcal{H}_O = \mathcal{C}(G), \qquad \mathcal{H}_C = \mathcal{C}_{\text{class}}(G),$$

with the inner product given in both cases by (12). We thus have specified the realization of axiom (T1) for elementary (in the sense of section 3) hypersurfaces.

[4] A diffeomorphism can transform a disc with an area form to any other disc with a given area form if and only if the total areas are equal.

A suitable orthogonal basis for $\mathcal{C}(G)$ is given through the Peter–Weyl decomposition by matrix elements of irreducible representations. We denote these by t_{ij}^V, where V is the representation and j and i are indices for a basis of V and its dual, respectively. A suitable orthogonal basis for the subspace $\mathcal{C}_{\text{class}}(G)$ is given by the characters associated with irreducible representations, $\chi^V = \sum_i t_{ii}^V$. Recall the identities

$$\overline{t_{ij}^V(g)} = t_{ji}^V(g^{-1}) \qquad \text{and} \qquad \int \mathrm{d}g t_{ij}^V(g^{-1}) t_{mn}^W(g) = \delta_{V,W}\delta_{i,n}\delta_{j,m}\frac{1}{\dim V}.$$

Hence, the inner product (12) with respect to these basis elements is

$$\left\langle t_{ij}^V, t_{mn}^W \right\rangle = \delta_{V,W}\delta_{i,m}\delta_{j,n}\frac{1}{\dim V}, \qquad \langle \chi^V, \chi^W \rangle = \delta_{V,W}. \tag{13}$$

According to (2), we might expect the antilinear involution of axiom (T1b) to be simply complex conjugation of the wave function. However, this is not the case here because the configuration data are sensitive to the orientation of a hypersurface. More, precisely, a holonomy g along an open string (for example) becomes a holonomy g^{-1} if the look at the string 'the other way round', i.e. change its orientation. This was not taken into account in section 2.3. Hence, the antilinear involution is really given by

$$(\iota_{\mathrm{O}}(\psi))(g) = \overline{\psi(g^{-1})}, \qquad (\iota_{\mathrm{C}}(\psi))(g) = \overline{\psi(g^{-1})}. \tag{14}$$

In terms of matrix elements this is

$$\iota_{\mathrm{O}}\left(t_{ij}^V\right) = t_{ji}^V, \qquad \iota_{\mathrm{C}}(\chi^V) = \chi^V. \tag{15}$$

We now turn to axiom (T2) describing hypersurface decompositions. As explained in section 3, there are only two elementary ones. We consider the decomposition of an open string into two open strings first. Gauge symmetry means that we do not actually have a map $G \to G \times G$ that expresses the splitting of configuration data. Indeed there are many ways a holonomy g could be split into a product $g_1 g_2$. However, since we are really dealing with functions on configuration data we can solve this problem by integrating over all such splittings. Hence,

$$(\tau_{\mathrm{OO}}(\psi \otimes \eta))(g) = \int \mathrm{d}h \psi(gh)\eta(h^{-1}) = \int \mathrm{d}h \psi(h)\eta(h^{-1}g). \tag{16}$$

Note that this looks quite different from (3). Indeed, the integral appearing in τ_{OO} may be seen as an averaging over gauges. These are precisely the 'missing' gauge transformations at the endpoints of open strings that we can only perform once we attach the open string to something else. Note also the analogy to gauge transformations in lattice gauge theory. These are performed at a vertex and affect the holonomies associated with all edges connected to the vertex. In terms of matrix elements,

$$\tau_{\mathrm{OO}}(t_{ij}^V \otimes t_{mn}^W) = \delta_{V,W}\delta_{j,m}\frac{1}{\dim V}t_{in}^V. \tag{17}$$

This map is indeed associative as required for consistency (see section 3). It makes $\mathcal{H}_{\mathrm{O}}$ into an associative algebra which is commutative only if G is abelian. Note that the algebra product is quite different from the commutative algebra product of $\mathcal{H}_{\mathrm{O}}$ as an algebra of functions.

The only other elementary hypersurface decomposition is that of a closed string into an open string. On the level of configuration data we want to recover a group element from its conjugacy class. Again, (in the non-abelian case) there are many possible group elements which yield the same conjugacy class. Hence, we integrate over them,

$$(\tau_{\mathrm{OC}}(\psi))(g) = \int \mathrm{d}h \psi(hgh^{-1}). \tag{18}$$

Again, this may be seen as an averaging over gauges at the endpoint where we glue the open string to itself. In terms of matrix elements,

$$\tau_{\mathrm{OC}}\big(t^V_{ij}\big) = \delta_{i,j}\frac{1}{\dim V}\chi^V. \tag{19}$$

Indeed, this is an orthogonal projection operator from the Hilbert space $\mathcal{H}_{\mathrm{O}}$ to its subspace $\mathcal{H}_{\mathrm{C}}$. If G is abelian, $\mathcal{H}_{\mathrm{O}} = \mathcal{H}_{\mathrm{C}}$ and τ_{OC} is simply the identity. The composition

$$\tau_{\mathrm{OC}} \circ \tau_{\mathrm{OO}}\big(t^V_{ij} \otimes t^W_{mn}\big) = \delta_{V,W}\delta_{j,m}\delta_{i,n}\frac{1}{(\dim V)^2}\chi^V \tag{20}$$

is commutative as required for consistency (see section 3). It is also straightforward to check axiom (T2b) explicitly.

4.3. The disc region and amplitudes

Let us return to the propagator (11). As we have seen, this depends only on the conjugacy class of the holonomy g around the boundary. Hence we, can expand it in characters. The dependence on the geometry of the disc manifold M is only through its area s. Therefore, the expansion coefficients depend only on this real number s and we can write

$$Z_M[g] = \sum_V \dim V \alpha_V(s)\chi^V(g), \tag{21}$$

where the sum is over the finite-dimensional irreducible representations V of G. Without knowing the exact nature of the functions $\alpha_V(s)$ we can write the amplitude map (4) for the disc D with area s as

$$\rho_{(D,s)} : \mathcal{H}_{\mathrm{O}} \to \mathbb{C} \qquad \rho_{(D,s)}(\psi) = \sum_V \dim V\alpha_V(s)\int \mathrm{d}g\,\chi^V(g)\psi(g). \tag{22}$$

In terms of matrix elements this is simply

$$\rho_{(D,s)}(\chi^V) = \dim V\alpha_V(s). \tag{23}$$

Recall from section 3 that the disc is the only elementary region out of which we can construct any other region by gluing. Hence, we need not specify any other amplitude *a priori*. However, we have several consistency conditions. In particular, recall from section 3 that axiom (T3x) implies the identity (9). We start by considering the extended amplitude map $\rho_{(D,s)} \circ \tau_{\mathrm{OC}} \circ \tau_{\mathrm{OO}}$ arising from decomposing the boundary of a disc into two open strings. Composing (20) with the amplitude yields

$$\rho_{(D,s)} \circ \tau_{\mathrm{OC}} \circ \tau_{\mathrm{OO}}\big(t^V_{ij} \otimes t^W_{mn}\big) = \delta_{V,W}\delta_{j,m}\delta_{i,n}\frac{\alpha_V(s)}{\dim V}. \tag{24}$$

When shrinking the disc (recall figure 2), the two open strings coincide up to orientation and the above map (with $s = 0$) is interpreted as a pairing $(\cdot,\cdot)_{\mathrm{O}} : \mathcal{H}_{\bar{\mathrm{O}}} \otimes \mathcal{H}_{\mathrm{O}} \to \mathbb{C}$. Using (13) and (15) the requirement $\langle\cdot,\cdot\rangle_{\mathrm{O}} = (\iota_{\mathrm{O}}(\cdot),\cdot)_{\mathrm{O}}$ is then seen to be equivalent to the condition that $\alpha_V(0) = 1$ for all irreducible representations V. Thus, the pairing is explicitly given by

$$\big(t^V_{ij}, t^W_{mn}\big)_{\mathrm{O}} = \delta_{V,W}\delta_{j,m}\delta_{i,n}\frac{1}{\dim V}. \tag{25}$$

In particular, taking $\big\{t^V_{ij}\big\}_{V,i,j}$ as a basis of $\mathcal{H}_{\mathrm{O}}$, the dual basis of $\mathcal{H}_{\bar{\mathrm{O}}}$ is given by $\{\dim V t^V_{ji}\}_{V,i,j}$. This satisfies equation (7), modified to take into account the orientation dependence of the holonomies:

$$\sum_{V,i,j} \dim V t^V_{ij}(g)t^V_{ji}(h) = \delta(g, h^{-1}). \tag{26}$$

Without working out the amplitude for the empty cylinder at this point, we know from (13) and (15) that the pairing $(\cdot,\cdot)_{\mathrm{C}} : \mathcal{H}_{\bar{\mathrm{C}}} \otimes \mathcal{H}_{\mathrm{C}} \to \mathbb{C}$ must be given by

$$(\chi^V, \chi^W)_{\mathrm{C}} = \delta_{V,W}. \tag{27}$$

Hence, taking $\{\chi^V\}_V$ as a basis of $\mathcal{H}_{\mathrm{C}}$ the dual basis of $\mathcal{H}_{\bar{\mathrm{C}}}$ is given by $\{\chi^V\}_V$.

Consistency requires that gluing the disc to another disc satisfies axiom (T5). This is no longer automatic as in the purely topological context of section 3. Rather, 'composing' the amplitude for a disc with area s_1 with the amplitude of a disc with area s_2 should yield the amplitude for a disc with area $s_1 + s_2$. Geometrically, we need to decompose the boundary of each of the two discs to be glued into two open strings. One open string of one disc is then glued to an oppositely oriented open string of the other disc; recall figure 3. Algebraically, this yields the identity

$$\rho_{(D,s_1+s_2)} \circ \tau_{\mathrm{OC}} \circ \tau_{\mathrm{OO}}(\psi \otimes \eta) \\ = \sum_{V,i,j} \dim V \rho_{(D,s_1)} \circ \tau_{\mathrm{OC}} \circ \tau_{\mathrm{OO}}\big(\psi \otimes t^V_{ij}\big)\rho_{(D,s_2)} \circ \tau_{\mathrm{OC}} \circ \tau_{\mathrm{OO}}\big(t^V_{ji} \otimes \eta\big). \tag{28}$$

Note that we have used the basis $\mathcal{H}_{\mathrm{O}}$ and dual basis of $\mathcal{H}_{\bar{\mathrm{O}}}$ as determined above. Evaluating this on matrix elements using (24) yields the series of identities $\alpha_V(s_1 + s_2) = \alpha_V(s_1)\alpha_V(s_2)$ for all V. Together with the condition $\alpha_V(0) = 1$ derived above and assuming continuity of the functions α_V we find that they must be exponentials of the area s. That is, $\alpha_V(s) = \exp(\beta_V s)$ for unknown constants β_V.

The only axiom we have not used so far is the unitarity axiom (T4b). This stands apart as it is much more related to the physical interpretation of the formalism than to its mathematical coherence. The simplest context for its application is given by decomposing the boundary of a disc into two open strings and converting the extended amplitude map into a map between the state spaces associated with the open strings. Concretely, we have to dualize one tensor component in the domain of the map (24). This yields a linear map $\tilde{\rho}_{(D,s)} : \mathcal{H}_{\mathrm{O}} \to \mathcal{H}_{\mathrm{O}}$ given by

$$\tilde{\rho}_{(D,s)}\big(t^V_{ij}\big) = \exp(\beta_V s) t^V_{ij}. \tag{29}$$

This defines obviously an isomorphism on the vector space of matrix elements which is dense in $\mathcal{H}_{\mathrm{O}}$. We declare that it should in fact be a vector space isomorphism on the whole space $\mathcal{H}_{\mathrm{O}}$. Then axiom (T4b) requires unitarity. This means the condition $|\exp(\beta_V s)| = 1$ for any irreducible representation V and any area s. If we equipped all regions with zero area this would be automatically satisfied. Indeed, this would make the theory topological and we have already seen in section 3 how this implies unitarity. In general, however, we find that the constants β_V must be imaginary to satisfy this condition.

4.4. General regions

As already mentioned, the amplitude for any region can be obtained via gluing from that of the disc. What is more, the amplitude for any connected region can be obtained by taking a single disc, decomposing its boundary suitably and then gluing pieces of this boundary together in suitable ways. Gluing a region to itself is not explicitly mentioned in axiom (T5), but it is implicitly obtained by gluing with an empty region.

We consider explicitly here only the cylinder. To obtain it, we decompose the boundary of a disc into four open strings and glue two non-adjacent ones together:

$$\rho_{(\mathrm{cyl},s)} \circ (\tau_{\mathrm{OC}} \otimes \tau_{\mathrm{OC}})(\psi \otimes \eta) \\ = \sum_{V,i,j} \dim V \rho_{(D,s)} \circ \tau_{\mathrm{OC}} \circ \tau_{\mathrm{OO}} \circ (\tau_{\mathrm{OO}} \otimes \tau_{\mathrm{OO}})\big(\psi \otimes t^V_{ij} \otimes \eta \otimes t^V_{ji}\big). \tag{30}$$

Concretely, $\rho_{(\mathrm{cyl},s)}$ can be expressed as

$$\rho_{(\mathrm{cyl},s)}(\psi \otimes \eta) = \sum_V \dim V \exp(\beta_V s) \int \mathrm{d}g\, \mathrm{d}h\, \chi^V(g)\psi(gh^{-1})\eta(h), \tag{31}$$

or, in terms of characters,

$$\rho_{(\mathrm{cyl},s)}(\chi^V \otimes \chi^W) = \delta_{V,W} \exp(\beta_V s). \tag{32}$$

At this point, we may easily verify axiom (T3x) for the empty cylinder.

Any connected region is a Riemann surface with holes. We classify it by genus g and hole number n. To obtain the amplitude for any Riemann surface we may, for example, first work out the amplitude for a sphere with $n + 2g$ holes and then glue g pairs of holes together. The result is

$$\rho_{\mathrm{g,n,s}}(\chi^{V_1} \otimes \cdots \otimes \chi^{V_n}) = \delta_{V_1,\ldots,V_n} \exp(\beta_{V_1} s)(\dim V_1)^{2-2g-n}. \tag{33}$$

In the special case with no hole we obtain also a sum over representations,

$$\rho_{\mathrm{g,0,s}} = \sum_V \exp(\beta_V s)(\dim V)^{2-2g}. \tag{34}$$

These amplitudes reproduce well-known formulae (obtained without corners) (see [2, 5]).

4.5. The constants β_V

So far we have not really exploited the actual form of the Yang–Mills action. We have only used (a) the fact that the action depends only on an area form, (b) gauge symmetry, (c) general properties of the path integral (11), and (d) the unitarity requirement. This has allowed us to determine the theory completely up to a set of unknown imaginary numbers β_V, one for each irreducible representation V of G.

A more detailed analysis of the path integral (11) shows that β_V takes the form

$$\beta_V = \frac{\mathrm{i}}{4}\gamma^2 C_V, \tag{35}$$

where C_V is the value of the quadratic Casimir operator on the representation V. This is well known in lattice gauge theory; see, e.g., [1]. For a simple derivation using essentially the same conventions as here (except for the i), see [15].

Unsurprisingly, exchanging the i in the path integral (11) for a -1 (as customary for example in lattice gauge theory) effects the same change in β_V, making it real. What is more, this makes the sums over irreducible representations appearing above generally convergent. While we have not mentioned this explicitly so far, those sums are not guaranteed to converge in our setting. However, amplitudes generally do converge if the boundary states are matrix elements. For example, we really have defined the amplitude map for the disc only on the dense domain of the Hilbert space spanned by matrix elements. Indeed, one can check that with β_V defined by (35) this amplitude map is unbounded and hence cannot be continuously extended to the whole Hilbert space. However, we may still formally satisfy axiom (T4) if we find a non-continuous extension to the whole Hilbert space. The physical relevance of this is questionable though and we do not pursue this point further.

Other amplitudes that may be ill-defined are those for closed regions given by (34). Notably, the amplitudes for the sphere ($g = 0$) and the torus ($g = 1$) will be ill-defined. The behaviors of amplitudes for higher genus surfaces depend on the group. For an abelian group all these amplitudes will diverge, while, for example, for $SU(2)$ the ones for $g \geqslant 2$ will converge. Most of these problems can be avoided if we make β_V real. However, in the case of zero area $s = 0$, i.e., for empty regions, they will persist.

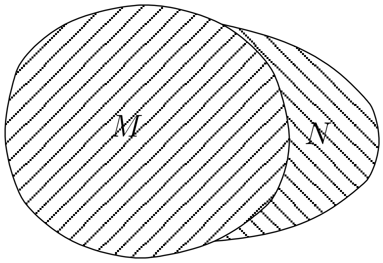

Figure 4. A disc M with an adjacent smaller disc N, deforming it.

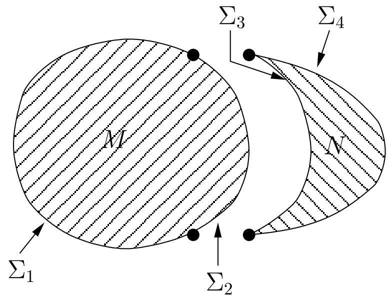

Figure 5. Decomposition of the boundaries of M and N into open string hypersurfaces.

4.6. Unitarity and probability conservation

The main reason for insisting on imaginary β_V is the physical interpretation. In quantum mechanics, unitarity is essential for a sensible probability interpretation. However, this is usually thought to make sense only for transitions between spacelike hypersurfaces. In the present context, we could for example take a cylindrical spacetime with space being the circle and time an interval. In this example axiom (T4b) has indeed the usual meaning of unitarity as guaranteeing probability conservation.

It was shown in [10] that a more general probability interpretation is possible which is not restricted to spacelike hypersurfaces. (For an actual example with timelike hypersurfaces, see [13].) The unitarity axiom (T4b) then acquires the physical meaning of probability conservation in more general situations.

In particular, consider a region M which we take to be a disc here. Now, enlarge the disc M through a 'small' outward deformation N to a new region $M \cup N$. For simplicity we let N and $M \cup N$ be (topological) discs as well (see figure 4). Denote the state spaces associated with the boundaries of M and $M \cup N$ by $\mathcal{H}_M$ and $\mathcal{H}_{M\cup N}$. Probability is then conserved for measurements associated with the boundary of M relative to measurements associated with the boundary of $M \cup N$ if the map $\mathcal{H}_{M\cup N} \to \mathcal{H}_M$ induced by the amplitudes is unitary. As was already remarked in [10], this setup necessarily involves corners. Indeed, this provides a major reason for our interest in corners here.

It is easy here to compute the map in question. Decompose the boundaries of M and N into two open strings each, so that they have one open string forming their common boundary. This is shown in figure 5. Now, consider the map $\tilde{\rho}_N : \mathcal{H}_{\Sigma_4} \to \mathcal{H}_{\Sigma_3}$ induced by the extended amplitude of the disc N. As we have seen, this is given by (29) with s being the area of N.

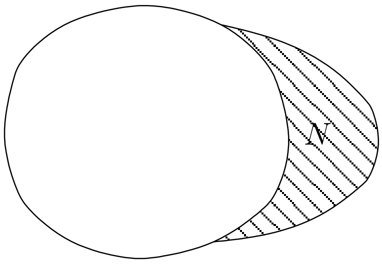

Figure 6. The region N as a degenerate cylinder.

The induced map from the decomposed boundary state space $\mathcal{H}_{\Sigma_1} \otimes \mathcal{H}_{\Sigma_4}$ of $M \cup N$ to the decomposed boundary state space $\mathcal{H}_{\Sigma_1} \otimes \mathcal{H}_{\Sigma_2}$ of M is simply the product $id_{\Sigma_1} \otimes \tilde{\rho}_N$. (Recall that Σ_3 and Σ_2 are identified.) It remains to obtain the corresponding map $\mathcal{H}_{M\cup N} \to \mathcal{H}_M$ between the undecomposed boundary state spaces. This amounts to completing the bottom line of the commutative diagram,

$$\begin{tikzcd} \mathcal{H}_{\Sigma_1} \otimes \mathcal{H}_{\Sigma_4} \arrow[r] \arrow[d] & \mathcal{H}_{\Sigma_1} \otimes \mathcal{H}_{\Sigma_2} \arrow[d] \\ \mathcal{H}_{M\cup N} \arrow[r] & \mathcal{H}_M \end{tikzcd}$$

where the vertical maps are given by (20). It is easy to see that the required map is given by $\chi_V \mapsto \exp(\beta_V s)\chi_V$, with s again the area of N. Imaginarity of the β_V implies unitarity of this map and hence probability conservation.

The above result is exactly the map one obtains from converting the cylinder amplitude (31) into a map between its two bounding closed string state spaces. This is not surprising. Indeed, we could avoid the use of corners above by thinking of N as a cylindrical region that surrounds M completely. This cylinder would be infinitely thin along parts of its boundaries, representing a kind of 'partially empty' region (see figure 6). Our above computation shows that both pictures are consistent and yield the same result as was in fact already anticipated in [10]. We refer the reader interested in a more general perspective on these issues to sections 4.3 and 9 of that paper.

5. Extensions

5.1. The vacuum

In [10] a proposal was made for the axiomatization of the concept of *vacuum*. This was successfully tested in the context of Klein–Gordon quantum field theory [12, 13]. It can be easily adapted and generalized to the case with corners.

(V1) For each hypersurface Σ there is a distinguished state $\psi_{\Sigma,0} \in \mathcal{H}_\Sigma$, called the *vacuum state*.

(V2) The vacuum state is compatible with the involution. That is, for any hypersurface Σ, $\psi_{\bar{\Sigma},0} = \iota_\Sigma(\psi_{\Sigma,0})$.

(V3) The vacuum state is compatible with decompositions. Suppose the hypersurface Σ decomposes into components $\Sigma_1 \cup \ldots \cup \Sigma_n$. Then $\psi_{\Sigma,0} = \tau(\psi_{\Sigma_1,0} \otimes \cdots \otimes \psi_{\Sigma_n,0})$.

(V5) The amplitude of the vacuum state is unity, $\rho_M(\psi_{\partial M,0}) = 1$.

Note that axiom (V4) of [10] is redundant here as it is implied by (V5) applied to empty regions. The only other change is the generalization of axiom (V3). This is formulated now with the generalized notion of decomposition and includes the map τ explicitly.

There is a unique realization of these axioms in two-dimensional quantum Yang–Mills theory. In both $\mathcal{H}_{\mathrm{C}}$ and $\mathcal{H}_{\mathrm{O}}$ the vacuum state is the state **1**. This is the constant function on G with value 1. In matrix element notation this is $\mathbf{1} = \chi^0 = t^0_{00}$, where 0 denotes the trivial representation. Note that the quadratic Casimir operator on the trivial representation is zero, $C_0 = 0$ and hence $\beta_0 = 0$ and $\alpha_0 = 1$. Verification of the vacuum axioms is elementary.

5.2. *Wilson loops*

It is easily possible to implement Wilson loops into the formalism. The most natural way to do this is via the introduction of additional labeled empty regions. Since we have two types of elementary hypersurfaces, there are correspondingly two types of elementary labeled regions. A labeled empty region with the shape of an open string is labeled by an element of $\mathcal{H}_{\mathrm{O}}$. We denote the corresponding extended amplitude map $\mathcal{H}_{\mathrm{O}} \otimes \mathcal{H}_{\bar{\mathrm{O}}} \to \mathbb{C}$ by $\rho_{\mathrm{O},\psi}$, where the label is $\psi \in \mathcal{H}_{\mathrm{O}}$. It is defined via

$$\rho_{\mathrm{O},\psi}(\eta \otimes \mu) = \int \mathrm{d}g\, \eta(g)\psi(g)\mu(g^{-1}). \tag{36}$$

Note that the orientation enters here in a special way. We can think of ψ as associated with an oriented piece of loop that points in the same direction as the orientation of the 'side' which carries η, but oppositely to the one that carries μ. This is reflected in how g enters in the arguments of the different functions either as g or as g^{-1}.

Similarly, we get labeled empty regions with the shape of a closed string. The label set is now the state space $\mathcal{H}_{\mathrm{C}}$. We denote the corresponding extended amplitude map $\mathcal{H}_{\mathrm{C}} \otimes \mathcal{H}_{\bar{\mathrm{C}}} \to \mathbb{C}$ by $\rho_{\mathrm{C},\psi}$, where the label is $\psi \in \mathcal{H}_{\mathrm{C}}$. It is defined with the same formula as (36), except that all functions must now be class functions.

Note that the newly defined empty regions generalize the empty disc and the empty cylinder. Indeed, these are recovered for the special choice of label $\psi = \mathbf{1}$, the constant function with value 1. However, the newly defined empty regions have in general no version of the amplitude with 'non-decomposed' boundary as the empty disc and cylinder have.

On first sight it may seem that our definitions have little resemblance to what one usually considers as Wilson loops. However, the new empty regions are precisely closed Wilson loops (closed string) or pieces thereof (open string). In the closed string case the label is a class function. In particular, we can choose a character. This recovers the usual labeling of Wilson loops by irreducible representations.

In the open case we have only a piece of a Wilson loop. To obtain a closed Wilson loop we have to glue pieces together. This gluing is not directly a gluing of the new empty manifolds at their endpoints as this notion does not exist. However, we may deduce such a notion by making more complicated gluings involving empty discs. The result is that the gluing of two labeled empty open string regions to a labeled empty open string region is given by formula (16). Similarly, the gluing of a labeled empty open string region to itself by joining the endpoints is given by formula (18). Hence, we obtain a new and completely different interpretation of the maps τ_{OO} and τ_{OC}. Instead of applying them to states we apply them to labels here. In particular, we see that as soon as we close a Wilson loop we get a label by a class function, or if we use matrix elements, by an irreducible representation.

It is relatively easy to see that the introduction of the new objects preserves the consistency and coherence of the axioms. Since our interest here is merely the implementation in principle, we abstain from performing concrete but straightforward calculations of amplitudes for surfaces with inserted Wilson loops. We refer the reader interested in this to some results obtained in [2, 5] (for closed Wilson loops).

5.3. Non-orientable surfaces

So far we have required regions to be orientable. However, this is not really essential. In axiom (T4) the orientation of the region is used to induce an orientation on its boundary. Instead, we may drop the orientation of the region, but explicitly specify an orientation on the boundary. The latter is essential and implies that regions must still have orientable boundaries. In axiom (T5) the matching orientations of the regions to be glued are used to ensure that the boundary components to be glued have opposite orientation. For non-oriented (including non-orientable) regions we may simply demand the latter property explicitly.

A problem occurs in so far as the gluing axiom is now less powerful than it should be. In particular, it will not be possible to obtain a non-orientable region out of orientable ones through gluing. Thus, in the two-dimensional case, the disc is no longer elementary in the sense that every other region can be obtained by gluing discs together. To remedy this we need to introduce gluings also along hypersurfaces with parallel instead of opposite orientation. This can be accomplished for example by inserting an orientation-changing map $\mathcal{H}_\Sigma \to \mathcal{H}_{\bar{\Sigma}}$ into the gluing. Note that this map must be linear and hence cannot be the antilinear involution ι.

In the context of the two-dimensional quantum Yang–Mills theory it is completely clear what this map is. Namely, it corresponds to interpreting the same configuration data on a hypersurface with opposite orientation. Hence, it corresponds to the map $g \mapsto g^{-1}$ for holonomy data. This means on the level of state spaces the map $\psi \mapsto \psi'$ with $\psi'(g) = \psi(g^{-1})$.

We are now in a position to work out amplitudes for non-orientable surfaces. As a first example consider the Möbius strip. To obtain its amplitude we start with a disc and decompose its boundary into four open strings. We then glue two non-adjacent open strings together (similarly to the case of the cylinder), but such that their orientations are parallel. That is, in this gluing we have to insert the aforementioned map. The resulting amplitude is

$$\rho_{(\text{Möb},s)}(\chi^V) = \delta_{V,V^*} \exp(\beta_V s). \tag{37}$$

Here, V^* denotes the representation dual to V. In other words, this amplitude is non-zero only if the representation V is self-dual. For example, for the group $SU(2)$ this is always the case while for the group $U(1)$ this is only the case for the trivial representation.

As another example, the amplitude for the Klein bottle is

$$\rho_{(\text{Klein},s)} = \sum_V \delta_{V,V^*} \exp(\beta_V s). \tag{38}$$

(Glue the cylinder to itself with an orientation reversal inserted.) The appearance of a factor δ_{V,V^*} is a general feature of amplitudes of non-orientable surfaces.

Our method of dealing with non-orientable surfaces by inserting an orientation changing map is a direct generalization of the method used by Witten [2] to the case with corners. We also refer to Witten's article for more examples of amplitudes for non-orientable surfaces.

6. Outlook

Physical experiments are usually confined to finite regions of spacetime and independent of what goes on in other parts of spacetime. Hence, it is desirable to be able to describe a physical process through states and amplitudes associated with such a region. Furthermore, it is desirable to be able to describe what happens when we join two such processes and associated regions together. This is the motivation (in the context of quantum field theory) for a gluing axiom of the form (T5). However, for spacetime regions with generic topology like that of a 4-ball this can only work if we have the concept of corners at our disposal.

At the same time, this gluing with corners yields valuable consistency conditions. In the example of two-dimensional quantum Yang–Mills theory, we have seen this explicitly. Gluing the disc to itself to obtain a new disc yielded strong constraints on the possible form of the amplitude map. More precisely, the unknown functions $\alpha_V(s)$ could in this way be constrained to be of an exponential form $\alpha_V(s) = \exp(\beta_V s)$.

This same type of gluing could also be made to play a role in the procedure of renormalization. For example, we might apply the present formalism to lattice gauge theory. Instead of the often used toroidally compactified versions of spacetime, one would consider bounded hypercubic pieces of spacetime. These would carry a hypercubic lattice within them (and on their boundary) of given length scale. Comparing the amplitude of one such piece with the amplitude of several pieces glued together, but with the lattice scale changed to obtain the same physical dimensions would allow us to set up the corresponding renormalization group equation. Of course, there are a lot of open questions to be addressed before this could become viable, such as the identification of appropriate boundary states.

Spin foam models have recently become popular in approaches to quantum gravity (see [15–18] and references therein). These models are state sum models, but share TQFT-like features in that they are composed out of elementary building blocks (usually n-simplices with certain labels). It should be possible to bring some of these models into the axiomatic form of section 2.2. This would in turn allow the application of the probability interpretation proposed in [10] and possibly help resolve long standing problems regarding their physical interpretation.

Combining this with the relation between gluing and renormalization suggested above leads to an enhancement of the framework for the renormalization of spin foam models proposed in [19, 20] (see also [15]). We merely mention here that viewing two adjacent n-balls as part of the cellular decomposition of an n-manifold, their gluing in the sense of axiom (T5) becomes the (n, n)-move (or 'fusion move') in the terminology of [15].

The present framework has some similarities to so called open-closed TQFT in two dimensions (see [21] and references therein). It should be useful to perform a detailed comparison between the two. For example, open-closed TQFT also admits free boundaries that do not carry state spaces. Such boundaries could be introduced in the present framework by attaching fixed states to ordinary boundaries. These states would then be seen as labels on the free boundaries. An equivalent way to look at this would be as Wilson loop empty regions (defined as in section 5.2), but with only one side.

The treatment of two-dimensional quantum Yang–Mills theory in this work is meant merely as a first example of a quantum field theory with corners. On the one hand, more complex theories need to be considered. Remaining in the two-dimensional context, conformal field theory comes to mind. On the other hand, higher dimensional examples are of interest. The question arises in particular, whether the axioms proposed in section 2.2 are 'good enough' in higher dimensions or need to be further modified. This concerns in particular axiom (T2), as now a whole hierarchy of dimensions comes into play when decomposing boundaries.

A theory that should be relatively straightforward to work out is three-dimensional quantum gravity with corners. In that case, we merely need a topological context. The hypersurfaces will be Riemann surfaces with holes, while the regions (general compact orientable three-manifolds with a boundary) do not admit a simple classification. However, the latter fact should not pose any serious problem as there is again one elementary region, the three-ball, out of which all others can be obtained by gluing.

Acknowledgments

I would like to thank J A Zapata and E Bianchi for stimulating discussions. I would also like to thank I Runkel for carefully reading the manuscript and alerting me to several minor mistakes.

References

[1] Migdal A A 1976 Recursion equations in gauge field theories *Sov. Phys.—JETP* **42** 413–8
[2] Witten E 1991 On quantum gauge theories in two dimensions *Commun. Math. Phys.* **141** 153–209
[3] Fine D S 1991 Quantum Yang–Mills on a Riemann surface *Commun. Math. Phys.* **140** 321–38
[4] Rusakov B 1990 Loop averages and partition functions in U(N) gauge theory on two-dimensional manifolds *Mod. Phys. Lett.* A **5** 693–703
[5] Blau M and Thompson G 1992 Quantum Yang–Mills theory on arbitrary surfaces *Int. J. Mod. Phys.* A **7** 3781–806
[6] Walker K 1991 On Witten's 3-manifold Invariants, unpublished
[7] Oeckl R 2003 Schrödinger's cat and the clock: lessons for quantum gravity *Class. Quantum Grav.* **20** 5371–80 (*Preprint* gr-qc/0306007)
[8] Oeckl R 2003 A 'general boundary' formulation for quantum mechanics and quantum gravity *Phys. Lett.* B **575** 318–24 (*Preprint* hep-th/0306025)
[9] Conrady F, Doplicher L, Oeckl R, Rovelli C and Testa M 2004 Minkowski vacuum in background independent quantum gravity *Phys. Rev.* D **69** 064019 (*Preprint* gr-qc/0307118)
[10] Oeckl R 2008 General boundary quantum field theory: foundations and probability interpretation *Adv. Theor. Math. Phys.* (*Preprint* hep-th/0509122) at press
[11] Atiyah M 1989 Topological quantum field theories *Inst. Hautes Études Sci. Publ. Math.* **68** 175–86
[12] Oeckl R 2005 States on timelike hypersurfaces in quantum field theory *Phys. Lett.* B **622** 172–77 (*Preprint* hep-th/0505267)
[13] Oeckl R 2006 General boundary quantum field theory: timelike hypersurfaces in Klein–Gordon theory *Phys. Rev.* D **73** 065017 (*Preprint* hep-th/0509123)
[14] Segal G 2004 The definition of conformal field theory *Topology, Geometry and Quantum Field Theory* (Cambridge: Cambridge University Press) pp 421–577
[15] Oeckl R 2005 *Discrete Gauge Theory: from Lattices to TQFT* (London: Imperial College Press)
[16] Oriti D 2001 Spacetime geometry from algebra: spin foam models for non-perturbative quantum gravity *Rep. Prog. Phys.* **64** 1703–57 (*Preprint* gr-qc/0106091)
[17] Perez A 2003 Spin foam models for quantum gravity *Class. Quantum Grav.* **20** R 43 (*Preprint* gr-qc/0301113)
[18] Rovelli C 2004 *Quantum Gravity* (Cambridge: Cambridge University Press)
[19] Oeckl R 2003 Renormalization of discrete models without background *Nucl. Phys.* B **657** 107–38 (*Preprint* gr-qc/0212047)
[20] Oeckl R 2006 Renormalization for spin foam models of quantum gravity *Proc. 10th Marcel Grossmann Meeting on General Relativity (Rio de Janeiro, 2003)* ed M Novello and S Perez Bergliaffa (Singapore: World Scientific) pp 2296–300 (*Preprint* gr-qc/0401087)
[21] Lauda A and Pfeiffer H 2005 Open-closed strings: two-dimensional extended TQFTs and Frobenius algebras *Preprint* math.AT/0510664