Instanton Density Operator in Lattice QCD from Higher Category Theory

Jing-Yuan Chen

Institute for Advanced Study, Tsinghua University, Beijing, 100084, China

Abstract

A natural definition for instanton density operator in lattice QCD has been long desired. We show this problem is, and has to be, solved by higher category theory. The problem is solved by refining at a conceptual level the Yang-Mills theory on lattice, in order to recover the homotopy information in the continuum, which would have been lost if we put the theory on lattice in the traditional way.

The refinement needed is a generalization—through the lens of higher category theory—of the familiar process of Villainization that captures winding in lattice XY model and Dirac quantization in lattice Maxwell theory. The apparent difference is that Villainization is in the end described by principal bundles, hence familiar, but more general topological operators can only be captured on the lattice by more flexible structures beyond the usual group theory and fibre bundles, hence the language of categories becomes natural and necessary. The key structure we need for our particular problem is called multiplicative bundle gerbe, based upon which we can construct suitable structures to naturally define the 2d Wess-Zumino-Witten term, 3d skyrmion density operator and 4d hedgehog defect for lattice S^3 (pion vacua) non-linear sigma model, and the 3d Chern-Simons term, 4d instanton density operator and 5d Yang monopole defect for lattice $SU(N)$ Yang-Mills theory.

In a broader perspective, higher category theory enables us to rethink more systematically the relation between continuum quantum field theory and lattice quantum field theory. We sketch a proposal towards a general machinery that constructs the suitably refined lattice degrees of freedom for a given non-linear sigma model or gauge theory in the continuum, realizing the desired topological operators on the lattice.

Contents

1 Introduction

Quantum chromodynamics (QCD), which describes the strong interaction between quarks and gluons, is a theory that has a simple and elegant form but from which extremely rich dynamics emerges. The dynamics is so non-trivial that most substantial computations of interest are out of the reach of usual analytical means. Wilson pioneered the development of lattice QCD [1], which puts QCD on a spacetime lattice of Euclidean signature, so that, at the fundamental level, the quantum path integral of the theory receives a non-perturbative, UV complete definition, while at the practical level, many problems of interest can henceforth be computed numerically [2, 3]. In this sense, in many practical scenarios lattice QCD is the essential embodiment of QCD.

One of the most important aspects in the richness of QCD is the existence of *instanton* [4], a topological configuration of the Yang-Mills gauge field, whose presence leads to significant consequences in the observed properties of QCD [5–7]. Yet a curious problem then arises. While the instanton configurations are well-defined in the continuum, and moreover it is intuitive that in lattice QCD these configurations must have been somehow effectively captured in the fluctuations of the lattice Yang-Mills path integral, *there is no lattice operator that can be defined in an unambiguous, mathematically natural manner to explicitly represent the instanton.* Yet such an operator is desired, if we want to compute the correlations of instantons among themselves or with other operators, or to study further formal, non-perturbative problems. This problem has been well-known for over four decades [8]. It has a simple origin, which we will review below along with its current workaround solutions [9].

The primary goal of this work is to solve this problem. We find we must understand more deeply what it really means to "put a continuum path integral onto the lattice". We are naturally brought to the use of *higher category theory*, which returns us a conceptually refined definition of lattice Yang-Mills path integral which represents the continuum Yang-Mills theory, especially its topological aspects, better than the traditional definition does. Based upon this lesson, our more general goal—though not fully achieved within the present work—is to establish a machinery that does the following: Given a continuum quantum field theory of interest—think of a non-linear sigma model or gauge theory whose field takes continuous values and has topological configurations—construct the suitable field contents on the lattice so that the topological aspects of the continuum theory are adequately captured.

Our goal of the present work is to introduce the new concepts and principles. An immediate numerical implementation is beyond the scope of the present work. However there would be no fundamental obstacle, and indeed, a more explicit technical description will be presented in a subsequent work [10]. We do anticipate that, using our newly introduced concepts, actual numerical computations that involve explicit instanton operators can be implemented and carried out in the near future.

We stress that being able to define topological operators on the lattice is not only useful for numerical purposes, but also important for analytical studies as well as fundamental understandings. For early examples, being able to define the vortex operator in S^1 non-linear sigma model on the lattice led to the discovery of the Berezinskii-Kosterlitz-Thouless transition in 2d [11–13] and allowed an explicit lattice derivation for the 2d boson-vortex duality [14] (with T-duality [15] being its special case); while being able to define the monopole

operator in lattice $U(1)$ gauge theory allowed explicit lattice derivations for the 3d boson-vortex duality and the 4d electro-magnetic duality [16–18]. As we will see, these previous examples played a crucial role in motivating our present work. Later, lattice construction has also found an important position in the developments of topological quantum field theory, from both the high energy [19, 20] and the condensed matter perspective [21–23]. The thoughts from topological quantum field theory have also deeply influenced our present work, even though the theories we consider, including QCD and others, are not purely topological and contain interesting dynamics at various energy scales. Therefore, in addition to the potential application to the numerics of lattice QCD, theoretical appeals is in itself a major motivation of this work, in hope to facilitate future analytical studies, and to deepen the understanding of the theories themselves by placing the problem in a broader context.

1.1 Problem and vague ideas

Let us first introduce the origin of the problem, and sketch some intuitive but vague ideas towards a solution.

We are interested in $SU(N)$ Yang-Mills gauge field in the continuum in 4d. The instanton density and the total instanton number (on an oriented closed 4d manifold $\mathcal{M}$, or an oriented infinite 4d manifold $\mathcal{M}$ with decaying field strength towards infinity) are given by

$$\mathcal{I} := \frac{1}{2} \mathbf{tr} \left[\frac{F}{2\pi} \wedge \frac{F}{2\pi} \right], \qquad I := \int_{\mathcal{M}} \mathcal{I} \in \mathbb{Z} . \tag{1}$$

The instanton number I is the second Chern number of the $SU(N)$ principal bundle of the gauge field over $\mathcal{M}$, and can be non-zero when the principal bundle is topologically non-trivial. In the quantum path integral of a gauge theory, all possible principal bundles are to be summed over.

We want to realize such topological configurations in lattice gauge theory. In the traditional lattice gauge theory [1], a lattice gauge field is to assign to each (oriented) lattice link l an element from the gauge group G, so the total configuration space is $\prod_{\text{links } l} G_l$. (We emphasize an important conceptual point: Gauge redundancy does not require any extra treatment on the lattice, because it is merely $\prod_{\text{vertices } v} G_v$, i.e. an element from G at each vertex, which is a locally finite size space for finite dimensional, compact G, and hence only leads to a product of *local constant factors*—hence unimportant—in the partition function [1]. At the level of observables, the Elitzur's theorem [24] means we do not need to demand any observable to be gauge invariant, since the gauge non-invariant part will essentially automatically vanish anyways.) Thus, in our case, to assign an instanton number to a lattice gauge configuration is to have a function

$$\prod_{\text{links } l} SU(N)_l \to \mathbb{Z} . \tag{2}$$

But the configuration space on the left-hand-side is connected. Thus, if we want to map the configurations to different values of instanton numbers, regardless of how we do so in details, we must encounter discontinuities in the assignment, which is unnatural.

4

From this simple argument it is easy to see the same problem occurs in more general cases, whenever we want to define the lattice counterpart of "topological configurations" for continuous-valued fields in the continuum. Such cases mainly include non-linear sigma models, whose traditional lattice realizations map each vertex to a point on a "vacua" target manifold, and gauge theories, whose traditional lattice realizations map each link to an element in a Lie group.

In the context of lattice QCD, the current solutions (see e.g. [9] for a review) are to allow discontinuous assignments, as long as the discontinuities are designed to only occur at field configurations of small weights in the Euclidean lattice path integral. There are several ways to do so. An early way is to forbid those lattice field configurations which appear "highly non-smooth", thus cutting the connected configuration space into disconnected pieces containing "smooth enough" configurations only, and then assign an instanton number to each piece by a procedure of interpolation to the continuum [8]. Another way, close in spirit but much more efficient in practice, is to design a procedure to flow those apparently "highly non-smooth" lattice configurations to more smooth ones, so that the interpolation to a continuum field configuration becomes obvious [25, 26]; discontinuities occur at where the flow bifurcates. Another direction of development is to define suitable Dirac operators on the lattice, and use a lattice version of the Atiyah-Singer index theorem to define the instanton number as the computed index [27–29], which may jump when the field configuration varies.

These methods to define instanton number on the lattice have all been studied deeply. The flow based methods and the Dirac operator based methods are both practically used for computing the topological susceptibility, $\langle I^2 \rangle / V$, the variance of the instanton number per spacetime volume. On the other hand, these definitions have important unsatisfactory aspects. On the practical side, if we want to compute correlations that involve local instanton densities at given spacetime positions, as opposed to the total instanton number, it seems the current methods are not sufficient to give an adequate local lattice definition (perhaps except for the first kind of method, which nonetheless has the disadvantage that too large a portion of the configuration space is forbidden, and hence not often used in practical computations). On the fundamental side, the problem is even more apparent—discontinuities indicate that these definitions are not sufficiently mathematically natural, and therefore it is hard to anticipate the aforementioned deepened understanding of the theory itself or the facilitation towards future analytical studies; moreover, the Dirac operator based methods have the additional problem of requiring an extra structure on the spacetime, the spin structure, i.e. fermion boundary conditions, which should not have been needed for defining the $SU(N)$ instanton configurations.

What can be done to solve the problem, then? There are two ideas to explore:

1. If the lattice theory has discrete degrees of freedom to begin with, then we can use their values to define discrete topological numbers without encountering discontinuity. Moreover, the definition should have some local expression so that the local density of the topological number is also defined.

2. Suppose the lattice degrees of freedom are still continuous-valued. But instead of assigning a discrete topological number to a field configuration, we assign a probability profile of the topological number:
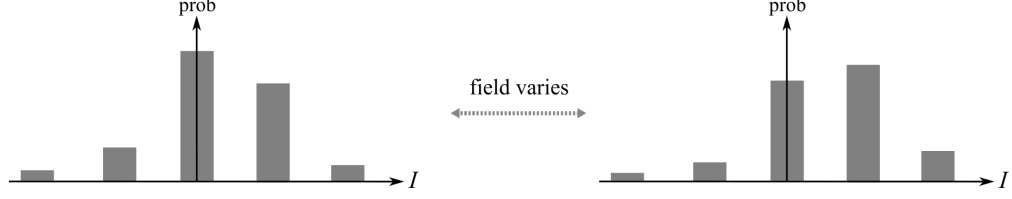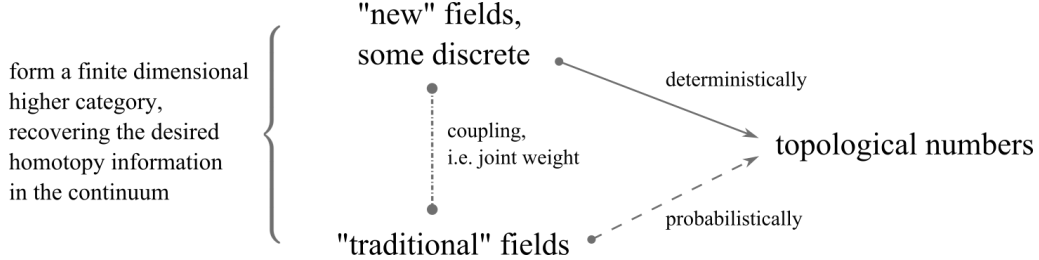
As we continuously vary the field on the lattice, the probability profile changes continuously. Nevertheless, the most probable topological number can jump when the two largest probabilities cross over at some "highly non-smooth" configurations, and this is intuitively how this idea is related to the previous methods that allow discontinuities. Moreover, the assignment of the probability profile to a given field should have some local expression, so that we can also define the probability profile of the local density of the topological number.

At the level of classical action, these ideas seem *ad hoc* and diverted from the original continuum theory, in which there seems to be no discrete-valued local fields, and $\mathcal{I}$ depends on F deterministically. However, we are in a quantum theory. The apparent degrees of freedom we use to present the path integral are nothing fundamental, as they are to be integrated out anyways. So these concerns raised around the classical action might not be relevant. On the very contrary, quantum mechanically there are good arguments in support of both ideas:

1. The very problem itself, that we should somehow get discrete topological numbers on the lattice, suggests that it is a good idea to find a presentation of our theory that involve discrete-valued degrees of freedom on the lattice. In fact, even in the continuum, the summation over different principal bundles is a discrete degree of freedom, though seemingly not manifested locally in the classical action.

2. If we intuitively think of the field on the lattice as some kind of "sampling" of the field in the continuum, then something deterministic in the continuum becoming probabilistic on the lattice is natural, because from a "sampling" we should not expect a deterministic inference of the "full original data", but a probabilistic inference. And this especially rings in the context of Euclidean path integral.

Most interestingly, these two ideas are not mutually exclusive, nor orthogonal, but complementary. Let us start from the first idea, i.e. we want to find such a presentation for our theory of interest on the lattice, that not only involves the "traditional" continuous-valued fields, but also some "new" fields, some of which are discrete-valued. In the Euclidean path integral, the "traditional" fields and the "new" fields are coupled, i.e. integrated over with a joint weight which depends on the fields smoothly. For a given configuration of all these fields, a topological operator density has an explicit local expression, such that the associated total topological number is only determined by the discrete-valued fields, hence there is no discontinuity. On the other hand, for a given configuration of the "tranditional" fields only, we can integrate out the "new" fields, and since those "new" fields are weighted probabilistically conditioned on the given "tranditional" fields, so will be the topological operator density and hence the total topological number.

6

The question, then, becomes how to naturally find such "new" fields and joint weights, given a continuum theory of interest. This is where the lessons from the previous examples [11–14, 16–18] and the power of category theory come in. The purpose is to build a natural correspondence between the lattice and the continuum. The idea can be summarized as

form a finite dimensional higher category, recovering the desired homotopy information in the continuum
{
"new" fields, some discrete
coupling, i.e. joint weight
"traditional" fields
deterministically
topological numbers
probabilistically
.

Here the "homotopy information" means how a field changes gradually from one place to another in the continuum; this is an infinite dimensional piece of information, but since a lot of details are unimportant, the topological part of the information can be effectively reduced to finite dimensional by category theory (these kinds of mathematical problems were indeed one major motivation why category theory was invented and developed in the first place), to be used as the lattice degrees of freedom. The "traditional" lattice fields are in no sense more "fundamental" than the "new" lattice fields, only that they are the lowest order topological approximation to the continuum in a suitable sense.

1.2 Why category theory

The idea sketched above has been realized before, though only in limited examples, and not organized into such a general perspective. It first appeared in what is now known as the Villain model [11, 12, 30], which we will review in details in Section 2. Briefly speaking, this is a lattice construction for S^1 non-linear sigma model, but such that, in addition to the "traditional" angular variable θ_v on the vertices, there is also an integer variable m_l on the links. They have a joint weight in the Euclidean path integral so that, summing out the integers m_l, we will retrieve a theory that resembles the traditional lattice S^1 non-linear sigma model (XY model) in terms of $e^{i\theta_v}$ only. At first, the Villain model was simply seen as an approximation to the seemingly "more actual" XY model. However, it was soon realized that, in the Villain model, we can explicitly define the the topological observables of winding number and vortex in terms of the integer variable m_l, and this played a crucial role in a lot of analytical studies of the S^1 non-linear sigma model [11–14, 16]. This suggests that there is something more profound to the Villain model than simply being an approximation to the XY model.

What are the lessons to be extracted from the Villain model? There are two directions of thinking, and both will lead to higher category theory if we dig deep enough. In fact, the two directions of thinking splice back again in the language higher category theory, whence bring us a natural solution for our goal.

The more geometrical direction of thinking is to first understand the "continuum meaning" of the integer variable m_l on the links in the Villain model. Think of the lattice as

being embedded in the continuum. Then, starting from the angular variable $e^{i\theta_v}$ at vertex v, moving in the continuum along the path traced out by the link l towards a neighboring vertex v', the integer m_l can be thought of as parametrizing the winding of $e^{i\theta(x)}$ around the S^1 before reaching $e^{i\theta_{v'}}$:

$$\text{lattice } \theta_{v'} - \theta_v + 2\pi m_l \quad \longleftrightarrow \quad \text{continuum } \int_{x \in l} d\theta(x) . \tag{3}$$

In this sense, the Villian model *topologically refines* the XY model: it captures more information of the continuum theory than the traditional XY model does; in particular, it recovers the homotopy information, so that winding and vortex can be explicitly defined.

This suggests that more generally, when the desired continuum theory has continuous-valued fields, the traditionally defined lattice theory misses the homotopy information from the continuum; but we can refine the lattice theory by suitably including more lattice fields in order to capture the essential homotopy information of interest from the continuum. This is admittedly vague, but category theory is what it takes to make this program substantial. Category theory is the mathematical language that deals with relations, relations between relations, essential contents, and so on, in a manner that is highly general, flexible but at the same time rigorous. It is therefore the natural language to help us rethink what it really means to "essentially capture" the continuum theory onto the lattice.

Let us now turn to the other, more algebraical direction of thinking, as we speak of the "essential information of interest", which is the topological information in our context. Soon after the invention of Villain model, it was understood that the Villainization process of introducing the integer variable is, mathematically, to implement the universal cover $\mathbb{Z} \to \mathbb{R} \to S^1$ over S^1, so that the fundamental group $\pi_1(S^1) \cong \mathbb{Z}$—the topological characterization of winding and vortex—is explicitly captured into the newly introduced $\mathbb{Z}$ variables, $\pi_1(S^1) \xrightarrow{\sim} \pi_0(\mathbb{Z}) \cong \mathbb{Z}$. With this understanding, the Villainization process has soon been generalized to lattice gauge theories, with the target space S^1 above replaced by Lie groups such as $U(1)$ or others with non-trivial π_1 [17, 18, 31], so that the monopole operators can be explicitly defined and worked with. We will review these ideas and these known constructions in details in Section 2.

For $SU(N)$ Yang-Mills theory, Villainization would not help, as $SU(N)$ already has trivial π_1 and is its own universal cover; meanwhile the instanton configurations in 4d comes from $\pi_3(SU(N)) \cong \mathbb{Z}$. It turns out that there is a mathematical notion called *3-connected cover*, which is to π_3 just like the universal cover (1-connected cover) is to π_1. This seems to be what we might need. However, in basically all cases of interest, the 3-connected covers are infinite dimensional spaces, and are hence contradictory to the very purposes of defining lattice theories, especially the purpose of performing numerical computations.

Category theory comes to rescue. In the recent years, Villainization has been reformulated as realizing the universal cover into a category [32–34]. With this perspective in mind, instead of realizing the 3-connected cover as a single infinite dimensional space, one has the new option of realizing it as a higher category, which involves multiple "layers" of spaces relating to one another via suitable maps, and moreover each layer can be chosen to be a finite dimensional space [35]. This higher category realization of the 3-connected cover, of which the key part is known as a *multiplicative bundle gerbe* [36], is what we need to put on

the lattice in order to capture the π_3 of the field in the continuum theory, and describing how this works is indeed the primary purpose of this paper.

Most interestingly, this is also where the geometrical and the algebraical directions of thinking splice back together. In the geometrical direction of thinking, we are led to consider the paths, surfaces and so on in the target space. It turns out that, these geometrical objects precisely form a choice of the higher categorical realization of the 3-connected cover [37,38]— albeit that, in this particular choice, infinite dimensional spaces are involved. But in the categorical sense, or say the algebraic sense, this choice of higher categorical realization is not unique, and there are realizations that are essentially equivalent, but with each layer being finite dimensional [35] and hence suitable for lattice theory. Therefore, the language of higher category theory indeed unifies the different directions of inspirations that can be drawn from the Villain model, and thereby solves our problem.

In a broader scope, our work directs towards a framework that turns the problem of "how to 'discretize' a continuum quantum field theory (QFT) onto the lattice while retaining the topological operators for the continuous-valued fields" into a well-posed mathematical problem. The general framework is only a sketched one at this stage (though our current limited development is already sufficient for our primary goal), as we will discuss in Section 6, and we believe it can be made more complete in the future. For non-linear sigma models, our proposal can be schematically (not precisely) summarized into a diagram:

$$
\mathcal{M} \to \mathcal{T} \quad \Longrightarrow \quad
\begin{array}{cccccccc}
\mathcal{L}_d & & \mathcal{P}^d\mathcal{M} & & \mathcal{P}^d\mathcal{T} & & \mathbf{ET}_d & \\
\Downarrow & & \Downarrow & & \Downarrow & & \Downarrow & \\
\cdots & & \cdots & & \cdots & & \cdots & \\
\Downarrow & \xrightarrow{\sim} & \Downarrow & \longrightarrow & \Downarrow & \xrightarrow[\text{what we care}]{\text{equiv up to}} & \Downarrow & \\
\mathcal{L}_2 & & \mathcal{P}^2\mathcal{M} & & \mathcal{P}^2\mathcal{T} & & \mathbf{ET}_2 & \\
\Downarrow & & \Downarrow & & \Downarrow & & \Downarrow & \\
\mathcal{L}_1 & & \mathcal{P}\mathcal{M} & & \mathcal{P}\mathcal{T} & & \mathbf{ET}_1 & \\
\Downarrow & & \Downarrow & & \Downarrow & & \Downarrow & \\
\mathcal{L}_0 & & \mathcal{M} & & \mathcal{T} & & \mathcal{T} &
\end{array}
\quad . \tag{4}
$$

The left of the " $\Rightarrow$ " describes a field in the continuum—simply a smooth function from the spacetime manifold $\mathcal{M}$ to some target manifold $\mathcal{T}$. The right is what we need for the lattice: Briefly speaking, the second and third columns (higher categories) are the continuum spacetime and the target space, where $\mathcal{P}$ means taking the space of all paths, which will in general give infinite dimensional spaces. The first column is the lattice, with the subscript labelling the dimension of the cells; this column is discrete, but nonetheless captures the essential information of the second column in the intuitive way—the lattice just fills up the continuum. The mathematical problem that becomes well-posed is to find the last column: we want a finite dimensional structure (in general a weak higher category) that is nonetheless topologically equivalent—up to whatever topological information that we care about—to the infinite dimensional third column. The horizontal arrows between columns are suitably defined maps (higher anafunctors between higher categories). The map from the first column to the last column represents a field on the lattice. Remarkably, this process of considering the higher path spaces in the continuum and then looking for categorical equivalence resonate

9

with the historical development of the subject of higher homotopy theory itself [39, 40]. The proposal for gauge theories is similar,

$$
\begin{array}{ccccccc}
 & & & \mathcal{L}_d & \mathcal{P}^d\mathcal{M} & \mathcal{P}^d|BG| & \mathbf{BEG}_d \\
 & & & \Downarrow & \Downarrow & \Downarrow & \Downarrow \\
 & & & \cdots & \cdots & \cdots & \cdots \\
\mathcal{PM} & G & & \Downarrow \xrightarrow{\sim} & \Downarrow \rightarrow & \Downarrow \xrightarrow[\text{what we care}]{\text{equiv up to}} & \Downarrow \\
\Downarrow \rightarrow \Downarrow & & \Longrightarrow & \mathcal{L}_2 & \mathcal{P}^2\mathcal{M} & \mathcal{P}^2|BG| & \mathbf{BEG}_2 \\
\mathcal{M} & * & & \Downarrow & \Downarrow & \Downarrow & \Downarrow \\
 & & & \mathcal{L}_1 & \mathcal{PM} & \mathcal{P}|BG| & G \\
 & & & \Downarrow & \Downarrow & \Downarrow & \Downarrow \\
 & & & \mathcal{L}_0 & \mathcal{M} & |BG| & *
\end{array}
\tag{5}
$$

where $|BG|$ is the classifying space of G, and the last column, i.e. the structure to be found, can be thought of as related to that in the non-linear sigma model case via the categorical process of delooping. Here we are only posting these diagrams for a schematic (not precise) summary. They will be explained in Section 6.

Prior to the present work, in the recent years higher category theory is already becoming important in theoretical physics, especially in the context of classification of phases of matter using generalized global symmetries (e.g. [32–34, 41–43]); higher gauge theories have also been proposed to describe exotic field theories [44–46]; moreover, some of these studies indeed have an emphasis on lattice theories [33, 41, 43, 45]. In the present work, however, the way higher category theory appears has some notable differences with the previous works:

- Physically, the present work is not a study of the low energy, universal properties of phases, but a study of the dynamics of particular theories at generic energy scales. Moreover, the theories we study are by no means "exotic". They are familiar quantum field theories (pion effective theory and Yang-Mills theory in QCD) that describe fundamental particle physics, even though there is no obvious involvement of higher categories in their familiar continuum presentations.

- Lattice theories with discrete higher categories are way much better studied than those with continuous ones. The present work deals with continuous ones, and as we have seen, the very reason that higher categories appear is to rescue continuity. The key mathematical feature needed for handling continuous higher categories is the use of simplicial weak categories and anafunctors. This point seems not to have been well appreciated in the theoretical physics context before.

- The categories involved in the present work are not inherently equipped with a linear structure, unlike those used in the the classification of low energy phases. Here the quantum mechanical linearity simply results from the fact that in the end we are building a well-defined path integral.

Note however, that if we apply the categorical formalism in this work to discrete groups, we will straightforwardly recover the previously developed group cohomology based lattice models [19, 23]; the Turaev-Viro model [20] beyond group theory can also be covered. Therefore,

we view the present work as (potentially) a more general framework that can encompass the study of topological aspects in both the UV physics and the IR physics.

The previous literature which could somehow hint our present work is [36], which introduced the higher categories we need, i.e. multiplicative bundle gerbes, in the context of Wess-Zumino-Witten terms and Chern-Simons terms in the continuum. The surprise, however, is that the seemingly overkilling mathematical formality there in the continuum becomes natural and necessary on the lattice. And of course the crucial advantage of the lattice over the continuum is that the path integral measure is explicitly locally well-defined. Moreover, the systematic topological relation found in our present work between quantum field theories in the continuum and on the lattice will allow us to work on more general problems, with more general mathematical structures, in the future.

In a recent work [47] that appeared as this manuscript was being finalized, bundle gerbe techniques have been employed to compute the Wess-Zumino-Witten integral in lattice non-linear sigma model. However, the degrees of freedom there are the traditional fields on vertices, hence the result still has the discontinuity problem. By contrast, the main point of the present work is that the degrees of freedom themselves form a bundle gerbe structure.

Some other recent works [48, 49], which appeared during the course of preparation of this manuscript, used bundle gerbe (without multiplicative structure) on the lattice for a very different physical context. The goal there is to study the higher Berry phase [50, 51] on 1d spatial lattice using matrix product states, and the bundle gerbe realizes an element in $H^3(X; \mathbb{Z}) \xleftarrow{\sim} H^2(\Omega_* X, \mathbb{Z})$, where X is the parameter space (which *a priori* has nothing multiplicative) at each point on the 1d spatial system. By contrast, in our present work, the multiplicative bundle gerbe realizes an element (the generator) of $H^4(|BG|; \mathbb{Z})$, which can transgress $H^4(|BG|; \mathbb{Z}) \to H^3(|G|, \mathbb{Z})$ if we forget about the multiplicative structure on G. While their physical context and hence the categorical structure are different from our present work, the purpose to introduce finite dimensional higher categorical structures on the lattice is the same: to keep the lattice problem locally finite dimensional meanwhile capturing the essential homotopy information from the continuum. This coincidence shows that such categorical way of thinking might be becoming broadly useful in tackling traditionally difficult problems in different branches of theoretical physics.

This work is organized as the following. In Section 2, we review in details the known examples of lattice theories with well-defined topological operators for continuous-valued fields; they include variants of the Villain model and the spinon decomposition. In Section 3, we explain the fundamental difficulty to go beyond the known examples if we stick with the familiar toolbox of group theory and/or fibre bundles. In Section 4, we introduce our main constructions for 1) lattice pion effective theory with skyrmion operator and 2) lattice QCD with instanton operator, respectively, using an intuitive explanation rather than the systematic language of category theory. In Section 5, we first cast the previously known examples in the language of strict higher categories, and then explain how the picture can be generalized to more flexible higher categories and lead to our main constructions. In Section 6, we sketch our more general proposal towards systematically connecting continuum QFT and lattice QFT. Finally, Section 7 contains our further, scattered thoughts.

2 Known Examples

We begin by reviewing the known examples of lattice QFTs in which the topological operators of continuous-valued fields are naturally defined after making suitable refinements on the lattice. These known examples belong to two kinds: (generalized) Villainization, and spinon-decomposition ($\mathbb{C}P^1$ representation). We will extract the common rationale behind these constructions, putting them into an organized picture. While the examples themselves are familiar, in our review we will make special emphasis on some conceptual points which are not commonly discussed but will become important. This will help us understand why no more example can be (and indeed, has been) found along this rationale, and henceforth think about how to step back and then reach beyond.

2.1 Villainized S^1 non-linear sigma model: winding and vortex

The first example is S^1 non-linear sigma model (nlσm). In 1d, there is the topological configuration of winding; in 2d and above, there is the topological defect of vortex, where a winding occurs around the vortex core. They are characterized by $\pi_1(S^1) \cong \mathbb{Z}$.

On the lattice, the traditional theory, known as the XY model, has an S^1 variable $e^{i\theta_v}$ at each vertex v. On the link l between v and v', the path integral is weighted by some positive increasing function $W_{XY}(e^{id\theta_l} + c.c)$, where $d\theta_l := \theta_{v'} - \theta_v$, so that configurations with better aligned $e^{i\theta}$ have higher weights. The partition function then reads

$$Z_{XY} = \left[\prod_{v'} \int_{-\pi}^{\pi} \frac{d\theta_{v'}}{2\pi} \right] \prod_l W_{XY}(e^{id\theta_l} + c.c) . \tag{6}$$

A usual choice for W_{XY} is $W_{XY}(x) = \exp[(x-2)/2T]$, where T can be interpreted as the temperature in statistical mechanics context, and $R = T^{-1/2}$ can be interpreted as the S^1 radius in QFT context. [1] However, minor quantitative changes in the detailed choice for the weight should not matter for long distance observables, in the sense of renormalization. The theory has a 0-form $U(1)$ global symmetry $e^{i\theta_v} \to e^{i\theta_v}e^{i\alpha_v}$ with α_v satisfying $e^{id\alpha_l} = 1$. [2] [3] (We do not say "$e^{i\alpha}$ is a constant" because if the spacetime has multiple pieces disconnected from each other, $e^{i\alpha}$ can take different values between the pieces, and this is sometimes referred to as "locally constant".)

For the general reason explained in Section 1.1, topological operators—windings and vortices—cannot be defined naturally in the XY model. For instance, consider a 1d lattice which forms a loop, with the $e^{i\theta}$ configuration indicated by the arrows; here we pictured two configurations:

---

[1] Here we assumed the lattice is uniform, since we have implicitly set each lattice length to be 1. Otherwise the weight on each link should depend on the length of the link in order for the physics to appear uniform. This consideration is understood in all the discussions below.

[2] Our use of S^1 versus $U(1)$ is based on whether it is thought of as a space, or as a Lie group with a special point being the identity.

[3] The global symmetry here is actually $U(1) \rtimes \mathbb{Z}_2 \cong O(2)$, where the $\mathbb{Z}_2$ part takes $e^{i\theta_v} \to e^{-i\theta_v}$. This $\mathbb{Z}_2$ part will not play a crucial role in our discussion below; we can explicitly break it and our key points below will not be altered.

.

We feel the configuration on the left should have a winding number $w = 1$. However, by turning each arrow individually (note, this cannot be done in the continuum), this configuration can be continuously deformed to the one on the right, which, we feel, should have $w = 0$. So a deterministic assignment of winding number would certainly run into discontinuities. To avoid this, we can, instead, say the two configurations, respectively, have high probabilities with $w = 1$ and $w = 0$, and the probabilities for different w crossover during the deformation process.

The Villain model is the natural refinement of the XY model that makes this concrete. Originally, on each link l, the variable under consideration is $e^{id\theta_l} \in U(1)$. In the Villain model, the link variable is extended to $\gamma_l \in \mathbb{R}$, with the constraint that $e^{i\gamma_l} = e^{id\theta_l}$. We will interpret γ_l below. If we choose a 2π range for θ, say $\theta \in (-\pi, \pi]$, then we can write $\gamma_l = d\theta_l + 2\pi m_l = \theta_{v'} - \theta_v + 2\pi m_l$, where $m_l \in \mathbb{Z}$; but the value of m_l itself is not physically meaningful, because if we change the 2π range for θ, the value of m_l will change accordingly to keep γ_l unchanged. Since the m part is not fixed by $e^{i\theta}$, it is an independent degree of freedom (d.o.f.) to be summed over in the path integral. The XY model is supposed to be the Villain model with m_l summed over, i.e.

$$W_{XY}(e^{id\theta_l} + c.c) \approx \sum_{m_l \in \mathbb{Z}} W_1(\gamma_l) \tag{7}$$
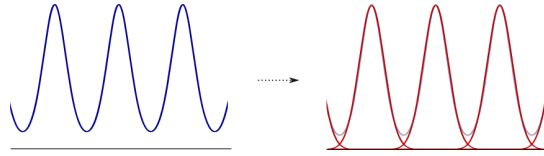
as a periodic function of $d\theta_l$, where W_1 is some positive even function decreasing with $|\gamma_l|$ (for each value of m_l, $W_1(\gamma_l)$ is pictured above as one red peak). Here the "$\approx$" is because, as we said before, the weight can change slightly without changing the physics, in the sense of renormalization; the usual choice $W_{XY}(e^{id\theta_l} + c.c) = \exp[(\cos d\theta_l - 1)/T]$ is often approximated by a sum of Gaussians, where $W_1(\gamma_l) = \exp[-\gamma_l^2/2T]$, with m_l controlling the center of the Gaussian. [4] The partition function of the Villain model is therefore [11, 12, 30]

$$Z = \left[\prod_{v'} \int_{-\pi}^{\pi} \frac{d\theta_{v'}}{2\pi} \right] \left[\prod_{l'} \sum_{m_{l'} \in \mathbb{Z}} \right] \prod_l W_1(\gamma_l) . \tag{8}$$

Now we need to understand the following questions:

---

[4]Sometimes this Gaussian approximation is said to be the motivation to perform Villainization. We emphasize that it is not. While bringing in many conveniences for further analytical studies (as we will see soon), the Gaussian approximation is not an important point at the fundamental level. The important point of Villainization is to make it possible to define topological operators [11, 12].

1. How does Villainization enable us to define windings and vortices?

2. What is the rationale behind the extension from $e^{id\theta_l} \in U(1)$ to $\gamma_l \in \mathbb{R}$?

3. In what sense things are continuous/smooth in the Villain model?

to appreciate that the Villain model is useful and natural.

Geometrically, it is intuitive to understand the meaning of $\gamma_l \in \mathbb{R}$ in relation to the continuum S^1 nlσm. Think of the lattice as being embedded in the continuum. Then $e^{i\theta_v}$ at different vertices v are like samplings from $e^{i\theta(x)}$ with x generic points in the continuum. The lattice link l connecting v and v' is a path in the continuum. Along this path l, the field $e^{i\theta(x \in l)}$ interpolates and traces out a path in S^1 going from $e^{i\theta_v}$ to $e^{i\theta_{v'}}$.
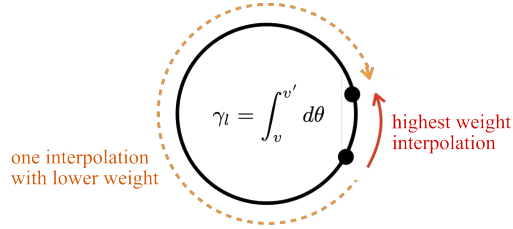

Then $\gamma_l \in \mathbb{R}$, satisfying $e^{i\gamma_l} = e^{id\theta_l}$, is nothing but the (signed) length of this path in S^1, $\gamma_l = \int_{x \in l} d\theta(x)$, with $m_l \in \mathbb{Z}$ describing the different winding choices for the interpolating path. It is then intuitive why the weight $W_1(\gamma_l)$ is chosen to be decreasing with $|\gamma_l|$.

With this understanding of γ, the natural definition for winding number in 1d is obviously

$$w := \oint_{1d} \frac{\gamma}{2\pi} := \sum_l \frac{\gamma_l}{2\pi} = \sum_l m_l \in \mathbb{Z} . \tag{9}$$

It is easy to confirm our intuition before. Given a $e^{i\theta}$ configuration, while the $2\pi\mathbb{Z}$ part of γ_l is not determined by $e^{i\gamma_l} = e^{id\theta_l}$, the weight W will prefer the choice that makes γ_l closest to 0. In this example of configuration
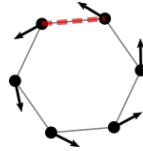
it amounts to the most probable choice being each $\gamma_l \approx 2\pi/(\text{number of links})$, and thus $w = 1$. In terms of m_l, since we have chosen $\theta \in (-\pi, \pi]$, we find $d\theta_l \gtrsim 0$ on most links, except for the indicated link, where $\theta_{v'} \gtrsim -\pi, \theta_v \lesssim \pi$ so that $d\theta_l \gtrsim -2\pi$, and therefore the most probable configuration for m is to have $m_l = 1$ at the indicated link and $m_l = 0$ elsewhere. Thus $w = 1$ is the most probable winding number given these $e^{i\theta_v}$ configurations, but other winding numbers are also possible. If we have chosen some other range $(a, a + 2\pi]$ for θ, then the most probable m_l configuration would change, but the physically meaningful γ_l and w do not depend on this choice.

In 2d and above, we can define the topological defect of vortex. The vorticity at a plaquette p is nothing but the winding number around p, so it is defined as

$$v_p := \frac{d\gamma_p}{2\pi} = dm_p \in \mathbb{Z} , \tag{10}$$

where $d\gamma_p$ is the lattice curl around the plaquette (it can be a square, or 2d cell of other shapes). Clearly it satisfies

$$\oint_{2d} v := \sum_p v_p = 0, \quad dv_c = 0 , \tag{11}$$

(here c labels a 3d lattice cube, or 3d lattice cell of other shapes) which means on a closed oriented 2d surface the total vorticity must be 0, [5] and in 3d or above the vortex forms a $(d-2)$-dimensional defect without boundary if viewed on the dual lattice.

Now that vortices are naturally defined on the lattice for $d \geq 2$, we can independently control their fugacity:

$$Z = \left[\prod_{v'} \int_{-\pi}^{\pi} \frac{d\theta_{v'}}{2\pi} \right] \left[\prod_{l'} \sum_{m_{l'} \in \mathbb{Z}} \right] \prod_l W_1(\gamma_l) \prod_p W_2(v_p) \tag{12}$$

where the subscripts on W denote the dimension of the lattice cells on which the weight is defined (so (7) will not longer take place exactly). A usual convenient choice is $W_2 = \exp[-U v_p^2/2]$, where U suppresses the vortices.

Being able to unambiguously define the vortices and control their fugacity is tremendously important for understanding the role played by vortices in the BKT transition in 2d [11–14, 52] and the spontaneous symmetry breaking (SSB) transition in 3d and above. Let us elaborate on this point. Once W_2 is non-trivial, i.e. non-constant, the "recovery of XY model" (7) can no longer happen exactly no matter what we choose for W_{XY} and W_1, because the m_l being summed over now appears not only in $W_1(\gamma_l)$ but also in those $W_2(v_p)$ of neighboring plaquettes. Summing out m_l will thus generate effective couplings of $e^{id\theta_l}$ between different neighboring links. One might worry that this may ruin the physics of the XY model, but we emphasize that in the renormalization sense, the detailed model is not of fundamental importance. For a finite range of choices of W_2, the physics at scales much larger than the lattice scale is unaffected. In fact, the introduction of the vortex fugacity weight in the model *helps* control the renormalization behavior—the physical intuition is that, as we coarse-grain the lattice, the effects of such weight is going to be effectively generated anyways, so having such a weight explicitly in the model helps us keep track of the renormalization of the associated effects (whether vortices becomes more or less important at large scales).

---

[5] On non-orientable ones such as a Klein bottle, it is easy to see the total vorticity is only well-defined mod 2, and thus can take any even number, and which even number to choose depends on some choice in the definition.

See [14, 52] for a detailed analysis in the BKT context. [6] [7] The same physical picture is understood in the more general cases in this paper.

If we want to completely forbid the vortices, we can use an S^1 Lagrange multiplier [15]

$$W_2^{forbid}(v_p) := \int_{-\pi}^{\pi} \frac{d\widetilde{\theta}_p}{2\pi}\, e^{i\widetilde{\theta}_p v_p} = \delta_{v_p,0}\ , \tag{13}$$

and this will hence prohibit the disordered phase. [8] This is something the traditional XY model cannot achieve. [9] The Lagrangian multiplier S^1 field $e^{i\widetilde{\theta}_p}$ can be thought of as living on $(d-2)$-dimensional cells on the dual lattice, and it has a $(d-2)$-form $U(1)$ global symmetry $e^{i\widetilde{\theta}_p} \to e^{i\widetilde{\theta}_p}e^{i\widetilde{\alpha}_p}$ with $e^{i\widetilde{\alpha}_p}$ satisfying $e^{id^{\star}\widetilde{\alpha}_l} = 1$, where $d^{\star}$ is like d but performed on the dual lattice. [10] [11] This vortex-forbiddening symmetry is the conservation of winding number, because a vortex in spacetime is a change of winding number in space. Using the Villain model on the lattice, we can easily see the celebrated mixed anomaly between the original 0-form $U(1)$ and this dual $(d-2)$-form $U(1)$ symmetry: We can try to introduce a background $U(1)$ gauge field for the original symmetry, and find the only way to make it appear consistently in W_2^{forbid} is to let it explicitly break the dual $U(1)$. [12] (In 1d, while W_2^{forbid} cannot be defined, one can define a topological theta term $e^{i\widetilde{\Theta}\sum_l \gamma_l/2\pi} = e^{i\widetilde{\Theta}w}$ in the

---

[6]In 1d, there is an even stronger result that the Villain model with Gaussian W_1 is the "perfect action" under renormalization [].

[7]One may note the generated effective coupling of $e^{id\theta_l}$ between different neighboring links is analogous to the idea of Symanzik improvement [53–57]. We will discuss the possible relation in the last section of the paper. This is another perspective that suggests the vortex fugacity weight helps renormalization.

[8]Unfortunately, an S^1 nlσm with vortices forbidden is often wrongly said to be "non-compact" in the literature, but the theory is really is still a compact S^1 theory, because: 1) the legitimate local boson number (or angular momentum) creation/annihilation operator is still integer quantized, $e^{in\theta_v}, n \in \mathbb{Z}$, and 2) there can be non-trivial windings around non-contractible loops. Mathematically, m being closed does not mean it is exact. By contrast, an actually non-compact $\mathbb{R}$ theory does not require $n \in \mathbb{Z}$, and moreover there is no winding number. However, traditionally this topological distinction was not well-appreciated, so that an S^1 nlσm with vortices forbidden has been called "non-compact", leading to confusions.

[9]In the XY model, the vortex fugacity cannot be controlled directly since vorticity is not well-defined, however one can anticipate to suppress vortices by suppressing large $d\theta_l$ mod 2π in the choice of W_{XY}, only that such control is indirect, not as explicitly meaningful as the W_2 fugacity in the Villain model. On the other hand, obviously W_2^{forbid} can only be defined in the Villain model but not in the XY model.

[10]One may think of v_p as a 2-cochain and $\widetilde{\theta}_p$ as a 2-chain (hence a $(d-2)$-cochain on the dual lattice), and $d^{\star}$ acting on cochains on the dual lattice is the same as the boundary ∂ acting on chains on the original lattice.

[11]This symmetry can be seen via the lattice version of integration by parts $\sum_p \widetilde{\theta}_p d\gamma_p = -\sum_p d^{\star}\widetilde{\theta}_l \gamma_l +$ (boundary terms). The boundary terms might or might not be 0 depending on the boundary condition, and hence the said symmetry might or might not be respected on the boundary.

[12]The introduction of $U(1)$ background gauge field is to replace $\gamma_l \to \gamma_l - A_l$ in W_1, where the background A_l is $U(1)$ in the sense that any local $2\pi\mathbb{Z}$ shift $A_l \to A_l + 2\pi N_l$ can be absorbed by the dynamical field $m_l \to m_l + N_l$. However, this changes the value of $v_p := d\gamma_p/2\pi = dm_p$ by dN_p in W_2. To remedy this, in W_2 we might replace v_p by $(d\gamma_p - dA_p)/2\pi$, which is no longer $\mathbb{Z}$-valued. For a generic W_2, there is no particular problem, but for $W_2 = W_2^{forbid}$, the $\widetilde{\theta}_p$ and hence $\widetilde{\alpha}_p$ will cease to be $U(1)$-valued but $\mathbb{R}$. Alternatively, we can replace v_p by $d\gamma_p/2\pi + S_p$ in W_2, where $S_p \in \mathbb{Z}$ is the Dirac string part for A_l such that the background flux $F_p := dA_p + 2\pi S_p$ (see Section 2.2) remains invariant under the N_l shift; $d\gamma_p/2\pi + S_p$ also remains invariant. But S_p can at most be required to be closed on the lattice (closedness is a requirement that can be imposed locally, while exactness is a non-local requirement; in a complimentary view, if S_p is required

16

path integral. One can discuss the notion of a dual "(-1)-form global symmetry" of $\widetilde{\Theta}$ and its mixed anomaly with the original $U(1)$ symmetry [58].)

An analytical convenience for choosing W_1 and W_2 to be Gaussian is the following. [13] In 2d, by performing Hubbard-Stratonovich transformations for both W_1 and W_2 and then summing out m, and viewing the result on the dual lattice, one can derive the exact *boson-vortex duality* between the lattice and the dual lattice [14] (here the terms are those on the exponent):

$$-\frac{1}{2T}\sum_l (d\theta + 2\pi m)_l^2 - \frac{U}{2}\sum_p dm_p^2$$

$$\Downarrow \quad \text{Hubbard-Stratonovich fields } \widetilde{\gamma}_l/2\pi \in \mathbb{R} \text{ and } \widetilde{\theta}_p + 2\pi\widetilde{\kappa}_p \in \mathbb{R}$$

$$-\frac{T}{2}\sum_l \frac{\widetilde{\gamma}_l^2}{(2\pi)^2} + i\sum_l \frac{\widetilde{\gamma}_l}{2\pi}(d\theta + 2\pi m) - \frac{1}{2U}\sum_p (\widetilde{\theta} + 2\pi\widetilde{\kappa})^2 + i\sum_p \widetilde{\theta} dm_p$$

$$\Downarrow \quad \text{sum out } m_l, \text{ enforcing } \widetilde{\gamma}_l - d^\star\widetilde{\theta}_l =: 2\pi\widetilde{m}_l \in 2\pi\mathbb{Z}$$

$$-\frac{T}{2(2\pi)^2}\sum_l (d^\star\widetilde{\theta} + 2\pi\widetilde{m})_l^2 - \frac{1}{2U}\sum (\widetilde{\theta} + 2\pi\widetilde{\kappa})_p^2 + i\sum_v \theta_v d^\star\widetilde{m}_v . \tag{14}$$

Note that the $\mathbb{R}/2\pi\mathbb{Z}$ part of the Hubbard-Stratonovich field for W_2 is nothing but the $\widetilde{\theta}$ in W_2^{forbid}. The $1/2U$ term explicitly breaks the dual $U(1)$ global symmetry of $\widetilde{\theta}$. As $U \to \infty$, the Hubbard-Stratonovich transformed W_2 reduces to W_2^{forbid} as expected, and the dual $U(1)$ symmetry emerges. In this limit, the boson-vortex duality becomes a self-duality (with $2\pi/\widetilde{T} = T/2\pi$), which is the lattice version of the T-duality in string theory [15]. In $d \geq 3$, the derivation for boson-vortex duality is exactly the same, and one can easily see that in $d = 3$ the resulting dual theory is a $U(1)$ gauge theory (see Section 2.2) coupled to a $U(1)$ nlσm Higgs field [16], while in more general dimensions it is a $(d-2)$-form $U(1)$ theory (see Section 2.3) coupled to a $(d-3)$-form $U(1)$ field; when $U \to 0$ the $(d-3)$-form field cease to exist and a $(d-2)$-form dual $U(1)$ global symmetry emerges.

All these discussions show that Villainization is a topological refinement to better connect the lattice theory to the continuum, and is tremendously useful in the analytical studies of important non-perturbative physics.

To prepare for our later discussions, however, we need some further understandings of the Villain model. We begin by reinterpreting Villainzation as gauging a $\mathbb{Z}$ global symmetry. While such kind of group theoretic interpretation will no longer be possible in the more general cases that we aim at (and this is a crucial point—we will only have category theoretic

---

to be exact, it is equivalent to A_l being $\mathbb{R}$ rather than $U(1)$), it is unlike $d\gamma_p/2\pi = dm_p$ which is exact by definition. Now that S_p might be non-exact in a Dirac quantized flux situation, when $W_2 = W_2^{forbid}$, it will explicitly break the dual $U(1)$ symmetry parametrized by $\widetilde{\alpha}$, demonstrating the said mixed anomaly.

[13]We emphasize that while the dualities below are exactly derived at the lattice level by choosing the weights to be Gaussian, if the weights are modified by not too much, the physics of the dualities should still hold in the IR. Hence this is a connivence, rather than something fundamental.

interpretation in general), it is helpful for bringing up some important points that we want to discuss.

Suppose we begin with an $\mathbb{R}$-valued theory, where $\theta_v \in \mathbb{R}$ instead of S^1. Each link has weight $W_1(d\theta_l)$, so the theory has a 0-form $\mathbb{R}$ global symmetry $\theta_v \to \theta_v + \alpha_v$, $\alpha_v \in \mathbb{R}$, $d\alpha_l = 0$. We want to reduce this $\mathbb{R}$ global symmetry to $U(1)$, and we can do so by gauging the $2\pi\mathbb{Z}$ subgroup of the global symmetry. Denoting the $2\pi\mathbb{Z}$-valued dynamical gauge field by $2\pi m_l$, the gauging process is to replace $d\theta_l$ by $d\theta_l + 2\pi m_l$ in W_1 and sum over m_l, and the gauge invariance is $\theta_v \to \theta_v + 2\pi k_v$, $m_l \to m_l - dk_l$ for any $k_v \in \mathbb{Z}$; moreover, the gauge flux dm_p can have its own dynamical weight, some $W_2(dm_p)$ on each plaquette p. Thus, we basically obtained the Villain model, except here $\theta_v \in \mathbb{R}$. But there is the $2\pi\mathbb{Z}$ gauge invariance k_v that we can exploit, to gauge fix each θ_v to $(-\pi, \pi]$. Thus, we obtained the Villain model by gauging the $2\pi\mathbb{Z}$ subgroup from an $\mathbb{R}$ theory and then fixing the $2\pi\mathbb{Z}$ on θ_v.

A first observation from this reinterpretation is that the Villain model relies on the fact that $S^1 = \mathbb{R}/2\pi\mathbb{Z}$. More exactly, it relies on finding the universal cover of S^1, which is $\mathbb{R}$:

$$
\begin{array}{c}
2\pi\mathbb{Z} \to \mathbb{R} \\
\downarrow \\
S^1
\end{array}
\qquad . \tag{15}
$$

While such a two row notation is standard for a fibre bundle in mathematics, in our context there is an extra meaning to have two rows—different rows are fields that live on lattice cells of different dimensions: the lowest row contains fields that live at the 0-dimensional vertices, $e^{i\theta_v} \in S^1$, while the row above are fields that live at the 1-dimensional links, $\gamma_l = d\theta_l + 2\pi m_l \in \mathbb{R}$ subjected to $e^{i\gamma_l} = e^{id\theta_l}$, and $m_l \in \mathbb{Z}$ that helps parametrize γ_l as $d\theta_l + 2\pi m_l$.

Why is finding the universal cover such a useful thing to do? This leads to the algebraic motivation behind Villainization, in complimentary to the geometrical motivation explained before. It is because Villainization leads to an isomorphism from $\pi_1(S^1)$ to $\pi_0(\mathbb{Z})$, through the universal cover $\mathbb{R}$ which is a non-trivial $\mathbb{Z}$ bundle over S^1. Generally, for a fibre bundle $F \to E \to B$, their homotopy groups satisfy the long exact sequence [14]

$$
\cdots \to \pi_n(F) \to \pi_n(E) \to \pi_n(B) \to \pi_{n-1}(F) \to \pi_{n-1}(E) \to \pi_{n-1}(B) \to \cdots . \tag{16}
$$

The reason to find the universal cover E of B is so that $\pi_1(E)$ is trivial and $\pi_0(E) = \pi_0(B)$ (which is trivial as well if B is connected), hence the long exact sequence leads to an isomorphism $\pi_1(B) \xrightarrow{\sim} \pi_0(F)$. In our case, $\pi_1(S^1)$ is what characterizes the winding of $e^{i\theta}$, which becomes ambiguous on the lattice due to the said discontinuity problem; by capturing this information into $\pi_0(\mathbb{Z})$, which is counted by m, the discontinuity problem is resolved because $\mathbb{Z}$ is discrete to begin with. Using the same idea, we can use the Villainization process to capture general π_1 topological information, see Section 2.3.

The $\mathbb{Z}$ gauge theory perspective also brings us to the front of an important conceptual question: In what sense things are continuous/smooth in the Villain model?

First of all, this is a question because apparently $\gamma_l = \theta_{v'} - \theta_v + 2\pi m_l$ is no longer continuous in the original S^1 variables $e^{i\theta_v}$, and therefore if we think of $e^{i\theta_v} \in S^1$ and

---

[14]Which means the image of each arrow is the kernel of the next arrow.

$m_l \in \mathbb{Z}$ as some kind of "fundamental local d.o.f.", the path integral weight $W_1(\gamma_l)$ appears discontinuous in $e^{i\theta_v}$.

This question arises because $e^{i\theta_v} \in S^1$ and $m_l \in \mathbb{Z}$ are not a good set of variables to simultaneously think about. We can either simultaneously think about $\theta_v \in (-\pi, \pi]$ and $m_l \in \mathbb{Z}$, or simultaneously think about $e^{i\theta_v} \in S^1$ and $\gamma_l \in \mathbb{R}$ subjected to the constraint $e^{i\gamma_l} = e^{id\theta_l}$. The path integral weight is smooth in either way.

The $\mathbb{Z}$ gauge theory perspective helps us understand this important conceptual point. In gauge theory, it is common to either describe the d.o.f. by gauge fixing, or by looking at gauge invariant combinations:

- Apparently, $\theta_v \in (-\pi, \pi]$ and $m_l \in \mathbb{Z}$ are the $\mathbb{Z}$ gauge fixed d.o.f.. The path integral weight is smooth in $\theta_v \in (-\pi, \pi]$, but the desired continuity from $\theta_v \gtrsim -\pi$ to $\theta_v \lesssim \pi$ in only recovered by absorbing the $2\pi\mathbb{Z}$ shift into the neighboring m_l's, or in other words, the path integral, smooth in $\theta_v \in (-\pi, \pi]$, becomes smooth in $e^{i\theta_v}$ only after summing over all the m_l's, see (7), but not before the sum.

- On the other hand, $e^{i\theta_v} \in S^1$ and $\gamma_l \in \mathbb{R}$ are $\mathbb{Z}$ gauge invariant. Physical observables must be built out of them. In terms of these $\mathbb{Z}$ gauge invariant variables, the path integral weight is smooth as expected. [15] The price paid is, the independent $\mathbb{Z}$ gauge invariant variables are not locally factorized, due to the link constraint $e^{i\gamma_l} = e^{id\theta_l}$, and this is a common feature of gauge theory. [16]

For instance, consider a lattice consisting of a single plaquette only, with vertices v_1, v_2, v_3, v_4 and links $l_{21}, l_{32}, l_{43}, l_{14}$. The locally factorized $\mathbb{Z}$ gauge fixed d.o.f., forming the space $(-\pi, \pi]^4 \times \mathbb{Z}^4$, are shown on the left, while an independent set of $\mathbb{Z}$-gauge invariant fields can be chosen as on the right, forming the actual configuration space $S^1 \times \mathbb{R}^3 \times \mathbb{Z}$, obtained from the restriction $(S^1)^4 \times \mathbb{R}^4|_{\text{link constraints } e^{i\gamma} = e^{id\theta}}$:
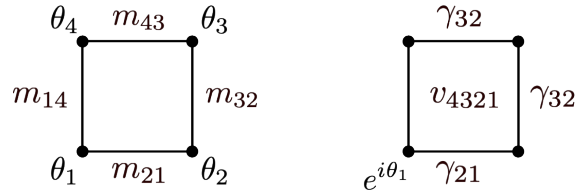
The $\mathbb{Z}$ gauge fixed space $(-\pi, \pi]^4 \times \mathbb{Z}^4$ is glued along suitable boundaries into the actual configuration space $S^1 \times \mathbb{R}^3 \times \mathbb{Z}$ (rather than into the naive $(S^1)^4 \times \mathbb{Z}^4$)—this is when we express the variables on the right in terms of those on the left. The path integral weight is smooth over the actual configuration space. Moreover, the actual configuration space $S^1 \times \mathbb{R}^3 \times \mathbb{Z}$ can be mapped to the $(S^1)^4$ of the XY model by exponentiating the γ_l. On a more general lattice, the configuration space has a topology of $(S^1)^{B_0} \times \mathbb{R}^{N_0 - B_0} \times \mathbb{Z}^{N_1 - (N_0 - B_0)}$

---

[15]The factors W_1, W_2 are smooth in γ_l, and do not otherwise depend on $e^{i\theta_v}$ due to the 0-form $U(1)$ global symmetry. If the $U(1)$ global symmetry is explicitly broken, there can be some vertex weight $W_0(e^{i\theta_v})$, which is still $\mathbb{Z}$ gauge invariant.

[16]For general gauge theories in the Hamiltonian formulation, it is familiar that the gauge invariant Hilbert space is not locally factorized. Although we are in a path integral rather than a Hamiltonian formulation, this aspect is similar.

obtained from the restriction $(S^1)^{N_0} \times \mathbb{R}^{N_1}\big|_{e^{i\gamma}=e^{id\theta}}$, rather than the naive $(S^1)^{N_0} \times \mathbb{Z}^{N_1}$ or the $\mathbb{Z}$ gauge fixed $(-\pi, \pi]^{N_0} \times \mathbb{Z}^{N_1}$, where N_0, N_1 are the numbers of vertices and links, and B_0 is the zeroth Betti number, i.e. the number of disconnected pieces of the lattice. The $(S^1)^{B_0}$ factor is where the $U(1)$ global symmetry acts on, the $\mathbb{Z}^{N_1-(N_0-B_0)}$ factor counts all possible winding and vortex configurations, and the $\mathbb{R}^{N_0-B_0}$ factor is space of the independent γ's given the winding and vorticity. The dependence on the topological number B_0 is a reflection that the configuration space is not a local factorization.

In summary, the apparent $\theta_v \in (-\pi, \pi]$ and $m_l \in \mathbb{Z}$ variables allow us to write the path integral measure in an explicitly locally factorized form, and they can be further glued into the actual configuration space which is not locally factorized; the path integral weight is smooth over the actual configuration space, effectively reproducing the traditional $e^{i\theta_v} \in S^1$ when we exponentiate the γ_l's. Alternatively, the space of the physical observables $e^{i\theta_v} \in S^1$ and $\gamma_l \in \mathbb{R}$ is also factorized, but there is an extra constraint $e^{i\gamma_l} = e^{id\theta_l}$ on every link, making the constrained actual configuration space not locally factorized. It seems a little verbose here to describe this trade-off between continuity and local factorizability, though fortunately the $\mathbb{Z}$ gauge theory perspective helps us understand this point, thanks to our familiarity with gauge theories. Later we will show the Villainization process can be recasted in the language of the Lie groupoid $S^1 \times \mathbb{R} \rightrightarrows S^1$. There, this continuity and locality issue becomes naturally understood in terms of functors from the lattice to this Lie groupoid. When we tackle our main problems of S^3 nlσm and $SU(N)$ Yang-Mills, the familiar gauge group approach becomes mathematically inadaquate, but these two alternative pictures of "apparently locally factorized d.o.f. glueing into a not locally factorized actual configuration space" and "apparently locally factorized d.o.f. being constrained down to a not locally factorized actual configuration space" remain valid, and is naturally understood from the category theory perspective.

2.2 Villainized $U(1)$ gauge theory: Dirac quantization, monopole, Chern-Simons and instanton

Soon after the Villainization method appeared in the S^1 nlσm context, it has been applied to $U(1)$ gauge theory as well [16–18]. In the recent years the Villainized $U(1)$ gauge theory (along with further generalizations) has attracted revived attention in the purview of (ordinary and higher form) symmetries and anomalies in topological terms [59] and topological orders [60], as well as the more exotic fractons [61]. The idea is extremely simple—just put those d.o.f. we have for Villainized S^1 nlσm onto lattice cells of one higher dimension. This leads to natural lattice descriptions for Dirac quantization in 2d, monopole in 3d or higher, abelian Chern-Simons (CS) term in 3d and abelian instanton in 4d, and so on.

In the traditional $U(1)$ lattice gauge theory [1], on each link there is a $U(1)$ variable e^{ia_l}, which can be thought of as a Wilson line across that link. The flux around a plaquette is also $U(1)$-valued, e^{ida_p}, which can be thought of as a Wilson loop around the plaquette; the path integral of the traditional $U(1)$ lattice gauge theory is weighted by a positive increasing function $W(e^{ida_p} + c.c.)$ on each plaquette. If the gauge theory is coupled to matter, such as

in lattice QED [17] or abelian Higgs model, e^{ia_l} appears in the hopping of the matter particles. For examples, when coupled to fermion ψ of charge $q_\psi \in \mathbb{Z}$, the hopping is $\bar{\psi}_{v'} e^{iq_\psi a_l} \psi_v$; when coupled to an XY model boson $e^{i\theta_v}$ of charge $q_\theta \in \mathbb{Z}$, the hopping is $e^{-id\theta_l + iq_\theta a_l}$. The charge must be integer due to the $U(1)$ nature of e^{ia_l}. The $U(1)$ gauge transformation is $e^{ia_l} \to e^{ia_l} e^{id\alpha_l}$, $\psi \to e^{iq_\psi \alpha_v} \psi$, $e^{i\theta_v} \to e^{iq_\theta \alpha_v} e^{i\theta_v}$ for arbitrary $e^{i\alpha_v} \in U(1)$.

The path integral of the gauge field is to integrate over $e^{ia_l} \in U(1)$ for all links l. As emphasized in the introduction, gauge redundancy is unimportant and does not require any treatment on the lattice. In the partition function, gauge redundancy is merely a $U(1)$ at each vertex, which is a locally finite size space and hence only leads to a product of *local* constant factor in the partition function [1]. And observables are not demanded to be gauge invariant, since any gauge non-invariant part will automatically vanish anyways, by Elitzur's theorem [24]. Therefore, gauge fixing or any other treatment about the gauge redundancy is not needed. This is a remarkable point, because in many cases in the continuum, gauge fixing involves solving (usually differential) equations over the spacetime manifold, generally leading to global issues, but these issues are artifacts from the choice of gauge fixing condition, rather than anything intrinsic to the gauge invariance itself. Any physical effect, local or global, must manifest on the lattice without any extra treatment about the gauge.

A pure $U(1)$ gauge theory has a 1-form $U(1)$ global symmetry $e^{ia_l} \to e^{ia_l} e^{i\beta_l}$, where $e^{i\beta_l}$ satisfies $e^{id\beta_p} = 1$, which does not change e^{ida_p} and hence the path integral weight. [18] This is *not* a $U(1)$ gauge transformation in general, because when the spacetime has non-contractible loops, the closedness condition $e^{id\beta_p} = 1$ does not imply exactness, i.e. there might be no choice of $e^{i\alpha_v}$ such that $e^{i\beta_l} = e^{id\alpha_l}$. Thus, when the $U(1)$ gauge field is coupled to matter, while the $U(1)$ gauge invariance must still be there, the 1-form $U(1)$ global symmetry is explicitly broken.

Similar to the winding and vortex configurations in XY model, configurations which look like having non-trivial Dirac quantized fluxes or non-trivial monopoles do appear in fluctuations in the traditional $U(1)$ lattice gauge theory, but there is no natural way to actually define these topological operators. Being able to define and hence forbid (or at least highly suppress) the monopole operator is particularly important for application to Maxwell theory in reality, in which monopoles have not been observed; monopole proliferation will lead to the confinement phase [1, 16, 52, 62] rather than the realistic Coulomb phase, i.e. the 1-form $U(1)$ SSB phase. [19]

---

[17]It is understood that QED is not "renormalizable" in the sense that if we reduce the lattice unit length in the UV, meanwhile changing the path integral weight in order to maintain the IR physics, then we expect, in analogy to the Landau pole, that the path integral weight will run into some singularity at some finite unit length, i.e. the unit length cannot be made arbitrarily small, unless new physics is introduced in the UV. But at any finite unit length before that happens, the lattice model is still well-defined and we can still discuss its IR physics.

[18]By "1-form global" here, it does not mean β is "constant". It means the $e^{i\beta}$ holonomy for any two loops (generally non-contractible) that can be deformed to each other must be the same. This is like, by "0-form global", it means $e^{i\alpha}$ for any two points can be connected by a path to each other must be the same, but not necessarily so for those that cannot.

[19]In the Coulomb vs the confinement phase, the Wilson loops' exponential suppression is proportional to the perimeter vs the (minimal) bounded area, generalizing the long vs short ranged correlation for order parameters in 0-form symmetry SSB. When coupled to matter field, both phases have perimeter law, but a

We have to Villainize the traditional theory to have natural definitions for the topological operators. That is, on each plaquette we now have the real-valued flux $f_p \in \mathbb{R}$ satisfying the constraint $e^{if_p} = e^{ida_p}$; if we fix the range $a_l \in (-\pi, \pi]$, then we can write $f_p = da_p + 2\pi s_p$, where $s_p \in \mathbb{Z}$ is to be thought of as the Dirac string variable (if viewed on the dual lattice) and summed over in the path integral. If we think of the plaquette as being embedded in the continuum, the lattice gauge flux $f_p \in \mathbb{R}$ can be thought of as the integral of the continuum field strength over the plaquette. Over a closed oriented 2d surface, we find the Dirac quantization condition

$$\oint_{2d} \frac{f}{2\pi} := \sum_p \frac{f_p}{2\pi} = \sum_p s_p \in \mathbb{Z} \tag{17}$$

(just like the winding number in the S^1 nlσm). On each lattice cube c (or 3d cell of other shapes), we can define the monopole number

$$m_c := \frac{df_c}{2\pi} = ds_c \in \mathbb{Z} , \tag{18}$$

(just like the vorticity in the S^1 nlσm) which satisfies

$$\oint_{3d} m := \sum_c m_c = 0, \quad dm_h = 0 \tag{19}$$

where h denotes a hypercube (or 4d cell of other shapes). So monopoles are $(d-3)$-dimensions defects without boundary, if viewed on the dual lattice. The Villainized $U(1)$ gauge theory reads

$$Z = \left[\prod_{l'} \int_{-\pi}^{\pi} \frac{da_{l'}}{2\pi} \right] \left[\prod_{p'} \sum_{s_{p'} \in \mathbb{Z}} \right] \prod_p W_2(f_p) \prod_c W_3(m_c) . \tag{20}$$

(If there are charged matter fields, Villainization makes no change to their coupling with the gauge field.) The usual Gaussian choices for the weights are $W_2(f_p) = \exp[-f_p^2/2e^2]$ (with e^2 the usual Maxwell coupling), $W_3(m_c) = \exp[-Um_c^2/2]$. Again, if we want to completely forbid the monopoles and hence prohibit the confinement phase—as it should for the Maxwell theory in reality—we can use the Lagrange multiplier [59, 60]

$$W_3^{forbid}(m_c) := \int_{-\pi}^{\pi} \frac{d\tilde{a}_c}{2\pi} \, e^{i\tilde{a}_c m_c} = \delta_{m_c,0} \tag{21}$$

[20] where $\tilde{a}_c$ can be thought of as living on $(d-3)$-dimensional cells on the dual lattice, and has a dual $(d-3)$-form $U(1)$ global symmetry $e^{i\tilde{a}_c} \to e^{i\tilde{a}_c} e^{i\tilde{\beta}_c}$ satisfying $e^{id^\star \tilde{\beta}_p} = 1$. Again,

---

closer inspection shows in the Coulomb phase the perimeter law can be realized as a zero law [63].

[20]Similar to the situation in footnote 8, a $U(1)$ gauge theory with monopoles forbidden has often been called "non-compact" in the literature, which is confusing, because it is in fact still a compact $U(1)$ gauge theory rather than a non-compact $\mathbb{R}$ gauge theory. The topological distinctions are whether the Wilson loop operators have to have quantized charges, and whether it is possible to have non-zero Dirac quantized fluxes over non-contractible 2d surfaces.

the original 1-form $U(1)$ global symmetry (exist only in a pure gauge theory) has a mixed anomaly with this dual $(d-3)$-form $U(1)$. (In $d=2$, while W_3^{forbid} cannot be defined, one can define the topological theta term $e^{i\widetilde{\Theta}\sum_p f_p/2\pi}$, and discuss the "$(-1)$-form global symmetry" of $\widetilde{\Theta}$ and its mixed anomaly with the 1-form $U(1)$ [58].) And again, dualities can be derived just like in the S^1 nlσm case; a remarkable case is the electromagnetic duality in 4d [16], which is self dual with $2\pi/\widetilde{e}^2 = e^2/2\pi$ when both charged matter particles and monopoles are forbidden (or both present).

The Villainized $U(1)$ gauge theory can be thought of as gauging a 1-form $\mathbb{Z}$ global symmetry from an $\mathbb{R}$ gauge theory, and then gauge fixing the 1-form $\mathbb{Z}$ by fixing the range of $a_l \in (-\pi, \pi]$. This uses the universal cover central extension

$$
\begin{array}{c}
2\pi\mathbb{Z} \to \mathbb{R} \\
\downarrow \\
U(1)
\end{array}
\tag{22}
$$

which is similar to the structure in S^1 nlσm, except everything is in one higher dimension, and thus the space S^1 becomes the group $U(1)$ because consecutive link variables can be naturally composed. We would like to reiterate the conceptual point made at the end of Section 2.1. The configuration space for Villainized $U(1)$ pure gauge theory is $U(1)^{B_1} \times \mathbb{R}^{N_1-B_1} \times \mathbb{Z}^{N_2-(N_1-B_1)}$ rather than the naive $U(1)^{N_1} \times \mathbb{Z}^{N_2}$ or the 1-form $\mathbb{Z}$ gauge fixed $(-\pi, \pi]^{N_1} \times \mathbb{Z}^{N_2}$, where N_2, N_1 are the numbers of plaquettes and links, and B_1 is the first Betti number. The $U(1)^{B_1}$ factor is the space on which the 1-form $U(1)$ global symmetry acts, while the $\mathbb{Z}^{N_2-(N_1-B_1)}$ factor counts all possible quantized flux and monopole configurations. The appearance of the topological number B_1 shows the configuration space is not locally factorized, but this is *not* due to the $U(1)$ gauge invariance (since the space for $U(1)$ gauge redundancy is just $U(1)^{N_0}$ which is local); again this comes from Villainization. Later, we will recast the Villainized $U(1)$ gauge theory in the language of the Lie 2-group $U(1) \times \mathbb{R} \rightrightarrows U(1) \rightrightarrows *$ [32, 33], which is the delooping of the Lie groupoid used for S^1 nlσm.

All the above are straightforward generalizations from the S^1 nlσm, by putting everything in one higher dimension. There are also aspects which do not have familiar counterparts in nlσm. They are the abelian CS term and abelian instanton.

When monopole is forbidden, the Dirac quantized real-valued flux f is a representative element for the first Chern class c_1 in the image of $H^2(|BU(1)|; \mathbb{Z}) \to H^2(|BU(1)|; \mathbb{R})$. Taking the cup product with itself will give an element in the image of $H^4(|BU(1)|; \mathbb{Z}) \to H^4(|BU(1)|; \mathbb{R})$, which is the abelian instanton number. [21]

More explicitly, on a hypercube h (or 4d cell of other shapes), we can use cup product to define the abelian instanton density over a hypercube [59, 60]

$$
\mathcal{I}_h := \left(\frac{f}{2\pi} \cup \frac{f}{2\pi} \right)_h .
\tag{23}
$$

---

[21] The classifying space of a group G is usually denoted as BG, but in this work we will reserve the notation BG for the category obtained by delooping G (see Section 5.1), while the classifying space will be denoted as $|BG|$, the geometric realization of the category BG (see Section 5.4).

[22] The cup product satisfies the Leibniz rule, so clearly in 5d and above, the instanton non-conservation defect, $d\mathcal{I}$, is proportional to the monopole defect $df/2\pi$. Moreover, in the below, suppose monopoles are forbidden with W_3^{forbid}, so that $ds = df/2\pi = 0$ after integrating out the Lagrange multiplier, then we have

$$\mathcal{I}_h = \frac{d\mathcal{C}_h}{2\pi} + (s \cup s)_h, \quad \mathcal{C}_c := \frac{1}{2\pi}(a \cup da + a \cup 2\pi s + 2\pi s \cup a). \tag{24}$$

Here $\mathcal{C}_c$ is the CS density which will be discussed soon. [23] This equation implies that the total abelian instanton number over a closed oriented 4d spacetime is quantized as expected:

$$I := \oint_{4d} \mathcal{I} = \sum_h \mathcal{I}_h = \sum_h (s \cup s)_h \in \mathbb{Z} . \tag{25}$$

Topological theta term in 4d can hence be defined. [24]

If the 4d spacetime is a spin manifold, it is well known [19] that the quantization is even stronger: $\sum_h (s \cup s)_h \in 2\mathbb{Z}$ for any s_p satisfying $ds_c = 0$. Therefore, in fermion-related contexts, there is another convention that calls $\mathcal{I}/2$ rather than $\mathcal{I}$ the abelian instanton density, and $I/2$ rather than I the total abelian instanton number.

The CS density $\mathcal{C}_c$ is only well-defined as $e^{i\mathcal{C}_c} \in U(1)$, because under the 1-form $2\pi\mathbb{Z}$ shift $a_l \to a_l + 2\pi n_l$ (which effectively restores the 2π periodicity of a_l) and $s_p \to s_p - dn_p$ that keeps the physical flux f_p invariant, $\mathcal{C}_c$ might shift by $2\pi\mathbb{Z}$. Now, on oriented 3d spacetime (or 3d submanifold embedded in higher dimensional spacetime), one may include another factor in the path integral, the CS phase (recall we supposed monopole are forbidden):

$$W_{CS}^k := e^{ik \sum_c \mathcal{C}_c} \tag{26}$$

with any CS level $k \in \mathbb{Z}$. Under $U(1)$ gauge transformation, the CS weight changes by a boundary factor, therefore if the 3d spacetime has a boundary, Dirichlet boundary condition is needed to avoid boundary gauge transformation. It is easy to check, using the expression (24), that a non-trivial CS phase breaks the 1-form $U(1)$ global symmetry of the $U(1)$ gauge field to a $\mathbb{Z}_{2k}$ subgroup, [25] and moreover this 1-form $\mathbb{Z}_{2k}$ global symmetry is anomalous. [26]

---

[22]On a hypercube, one choice of the cup product is the following. Suppose the hypercube has corner vertices given by coordinates $x, y, z, \tau \in \{0, 1\}$. There are a total of six pairs of plaquettes p and p_f on the hypercube, such that the center of p_f is shifted from the center of p by $\hat{x}/2 + \hat{y}/2 + \hat{z}/2 + \hat{\tau}/2$; for example one such pair is the xy-plaquette p centered at $x = y = 1/2, z = \tau = 0$ and the $z\tau$-plaquette p_f centered at $x = y = 1, z = \tau = 1/2$. Multiply $f_p f_{p_f}$ for each such pair, and then adding up the contributions from all six pairs, we get $(f \cup f)_h$. One can show the cup product satisfies the Leibniz rule under lattice exterior derivative. The choice of cup product is not unique, but any choice is required to satisfy the Leibniz rule. On more general 4d lattice, the choice of cup product is given by a branching structure.

[23]On a cubic lattice, a choice of cup product is defined using the shift $\hat{x}/2 + \hat{y}/2 + \hat{z}/2$, and this choice is compatible with the 4d choice made above when the 3d is embedded in 4d as the xyz hyperplane.

[24]One can also let the theta become local and dynamical, but then for consistency we will need to introduce the Villainization integer field for this theta (on the dual lattice), which will couple to the CS density. This is the lattice axion theory.

[25]To get the factor of 2, we need the property $\beta \cup s = s \cup \beta + d(\cdots)$ when β and s are both closed. This $(\cdots)$ is denoted by the cup-1 product $\beta \cup_1 s$, and this is how the notion of higher cup product is motivated.

[26]Which means if a 2-form $\mathbb{Z}_{2k}$ background is introduced, the CS phase will not be gauge invariant under

If the 3d spacetime is endowed with a spin structure, then level $k \in \mathbb{Z}/2$ is also possible. The $e^{i\pi}$ ambiguity in $e^{ik\sum_c \mathcal{C}_c}$ for half-integer k can be absorbed by an extra fermionic path integral $z_\chi[s] = \pm 1$ that depends on s_p mod 2 as well as a choice of the spin structure, so that the well-defined combination valid for any $k \in \mathbb{Z}/2$ is

$$W_{CS}^k := e^{ik\sum_c \mathcal{C}_c}(z_\chi[s])^{2k} . \tag{27}$$

The explicit construction of $z_\chi[s]$ can be found in [65] for simplicial complex and in [60] for cubic lattice along with an intuitive Berry phase interpretation. Because of this, in fermion-related contexts, there is another convention that calls $\mathrm{k} := 2k \in \mathbb{Z}$ rather than $k \in \mathbb{Z}/2$ the abelian CS level.

The 3d $U(1)$ Chern-Simons-Maxwell theory on lattice reads [66,67]

$$Z_{kCS} = \left[\prod_{l'} \int_{-\pi}^\pi \frac{da_{l'}}{2\pi} \right] \left[\prod_{p'} \sum_{s_{p'} \in \mathbb{Z}} \right] W_{CS}^k \prod_p W_2(f_p) \prod_c W_3^{forbid}(m_c) . \tag{28}$$

It is important to note that the theory becomes ill-defined when the Maxwell weight W_2 becomes trivial, $W_2 = 1$, i.e. when one attempts to define a purely topological CS theory on the lattice. This problem was originally analyzed in $\mathbb{R}$ gauge theory [68], but the problem stays the same in Villainized $U(1)$ gauge theory. [27] In fact, a purely topological CS theory is expected to be impossible, because the gapless chiral boundary mode must be non-topological. So it is natural to include a non-topological Maxwell term [66,67]. Even in the continuum, a Maxwell term with tiny $1/e^2$ is secretly understood in the regularization of the eta-invariant [70]; when we are on the lattice, the necessity to include a Maxwell term just gets better exposed.

the 1-form $\mathbb{Z}_{2k}$ gauge transformation. It is well-known and easy to check on the lattice [60,64] that gauging a $\mathbb{Z}_n$ subgroup of this $\mathbb{Z}_{2k}$ (which means introducing a 2-form $\mathbb{Z}_n$ background field and then promoting this background to dynamical—this will essentially make the $a \cup s + s \cup a$ terms in $\mathcal{C}$ rescale by $1/n$) is equivalent to (after rescaling a by $1/n$—which leads to some unimportant local constant in the path integral measure) dividing the CS level by n^2. So only those $\mathbb{Z}_n$ subgroups of this $\mathbb{Z}_{2k}$ where n^2 divides k will be non-anomalous. (And if n^2 divides $2k$ but not k, the theory can be made non-anomalous by introducing fermions; see below.)

[27]First consider the $\mathbb{R}$-valued CS term on the lattice, $\propto \oint_{3d} a \cup da$, $a_l \in \mathbb{R}$. We vary a to find the equation of motion, which means we want $\oint_{3d} (\delta a \cup da|_{\mathrm{EoM}} + da|_{\mathrm{EoM}} \cup \delta a) = 0$ for any δa. But the two terms are in general unequal, unlike the wedge product in the continuum. Therefore we cannot conclude $da|_{\mathrm{EoM}} = 0$, which means there are undesired zero modes that can be added to any given da configuration while leaving the action invariant. Moreover, unlike the gauge redundancy which occurs at each vertex locally, these extra zero modes have non-local profiles. Thus, they make the Gaussian path integral ill defined.

The problem is the same in Villainized $U(1)$ gauge theory with monopoles forbidden, because locally (though not globally) this theory looks the same as $\mathbb{R}$ gauge theory and hence inherits the same problem: Given any configuration of the gauge flux $f_p \in \mathbb{R}$, it is easy to see any shift Δf (due to shifts of a and s) that satisfies $d\Delta f_c = 0$ and $\oint_{3d} (\delta a \cup \Delta f + \Delta f \cup \delta a) = 0$ for any $\delta a_l \in \mathbb{R}$ will leave the CS term invariant. The partition function thus diverges, with one infinite factor from each of such non-local zero mode, and the number of such non-local zero modes depends non-locally all the details of the lattice and the cup product.

Imposing non-local constraints can directly forbid these zero modes [64,69]. However, in general we want a QFT to be local, and this can be achieved by having a non-topological but local Maxwell term that removes these non-local zero modes [66–68].

While the CS-Maxwell theory is non-topological, it is a free theory if the Maxwell weight W_2 is chosen to be Gaussian as usual. In this case, the CS-Maxwell theory can be solved. It reproduces all the interesting properties from a continuum $U(1)$ CS theory [71], but in an explicit, UV complete fashion. These include: the Wilson loop flux attachment, with the framing interpolating from point splitting [71] (determined by the cup product) at small $1/e^2$ to geometrical [72] (determined by the metric) at large $1/e^2$ [73]; the ground state degeneracy; the chiral boundary mode and, most non-trivially, the associated gravitational anomaly understood in a microscopic exposition. We will present these details in a separate work [66].

2.3 More general Villainizations, including $\mathbb{Z}_2$ vortex in $\mathbb{R}P^2$ non-linear sigma model, and $\mathbb{Z}_N$ monopole in $PSU(N)$ gauge theory

Villainization has many more applications. The most obvious is to work with multiple $U(1)$, which is useful in studying topological order and Hall conductivity [60, 74, 75]. Another obvious direction is to work with q-form $U(1)$ gauge fields, where $q = 0, 1$ reduce to the previous cases; by the same steps as before, we can derive boson-vortex-type dualities between q-form $U(1)$ gauge theory and $(d-q-2)$-form $U(1)$ gauge theory, and demonstrate the associated mixed anomaly between the q-form $U(1)$ and $(d-q-2)$-form dual $U(1)$ global symmetries. Interestingly, sometimes Villainization is even useful for dealing with discrete abelian gauge groups for more subtle purposes (compared to our main purpose of avoiding discontinuities): An important case is $2\mathbb{Z} \to \mathbb{Z} \to \mathbb{Z}_2$ for defining spin-c connection in footnotes 35 (also footnote 29); also, $n\mathbb{Z} \to \mathbb{Z} \to \mathbb{Z}_n$ helps manifest the Coulomb phase in $\mathbb{Z}_n$ gauge theory [76]; $n\mathbb{Z}_n \to \mathbb{Z}_{n^2} \to \mathbb{Z}_n$ facilitates a nice lattice implementation of Dijkgraaf-Witten twisted abelian topological order [75, 77]; and there are further applications in more exotic models [61].

It is common to develop the impression that Villainization is to deal with $U(1)$, or at most including other abelian groups built out of (or being a subgroup of) $U(1)$. This is not the case. Through our algebraic motivation discussed below (15), it should be clear that the real purpose of Villainization is to capture π_1, and has nothing to do with whether the symmetry or gauge group is abelian or not. This leads to many more applications.

We begin with nlσm, i.e. 0-form theory. Suppose the nlσm target space $\mathcal{T}$ has a non-trivial $\pi_1(\mathcal{T}) \cong \Gamma$, with Γ some discrete group, not necessarily abelian. To capture the Γ winding/vorticity, we can Villainize the traditional $\mathcal{T}$ lattice nlσm by the universal cover $\widetilde{\mathcal{T}}$

$$
\begin{array}{c}
\Gamma \to \widetilde{\mathcal{T}} \\
\downarrow \\
\mathcal{T}
\end{array}
. \tag{29}
$$

Note that $\widetilde{T}$ does not have to be a group, only Γ does. Let us take the $\mathbb{R}P^2$ nlσm for an example, which describes the physics of nematicity in systems like liquid crystals. The target space has $\pi_1(\mathbb{R}P^2) \cong \mathbb{Z}_2$, and hence there is $\mathbb{Z}_2$ winding in 1d and $\mathbb{Z}_2$ vorticity in higher dimensions; $\mathbb{R}P^2$ also has higher π_n's, but for now we ignore their physical effects and only

focus on the π_1 effects. The structure

$$\begin{array}{c} \mathbb{Z}_2 \to S^2 \\ \downarrow \\ \mathbb{R}P^2 \end{array} \tag{30}$$

can be implemented on the lattice as an S^2 nlσm with a $\mathbb{Z}_2$ global symmetry gauged. Thus, the Villanized partition function reads

$$Z = \left[\prod_{v''} \int_{S^2} \frac{d^2 \hat{n}_{v''}}{4\pi} \right] \left[\prod_{l'} \sum_{\sigma_{l'} = \pm 1} \right] \prod_{l = \langle v'v \rangle} W_1(\hat{n}_{v'} \cdot \sigma_l \hat{n}_v) \prod_p W_2(D\sigma_p) \tag{31}$$

where W_1, W_2 are some positive increasing function, and $D\sigma_p := \prod_{l \in \partial p} \sigma_l$ describes the $\mathbb{Z}_2$ vortex. (We can also introduce W_2^{forbid} that uses a $\mathbb{Z}_2$ Lagrange multiplier field to forbid the $\mathbb{Z}_2$ vortex. In 1d, while there is no W_2, we can have a topological $\mathbb{Z}_2$ theta term for the $\mathbb{Z}_2$ winding number.) The $\mathbb{Z}_2$ gauge invariance here is $\hat{n}_v \to s_v \hat{n}_v$, $\sigma_{l = \langle v'v \rangle} \to s_{v'} \sigma_l s_v^{-1}$. Note that we have not fix the $\mathbb{Z}_2$ gauge here, which is fine because it is merely a local, finite factor of 2 on each vertex; this is in contrast to the $\mathbb{Z}$ gauge invariance before, which is of infinite size and hence must be fixed. (If we do want to fix the $\mathbb{Z}_2$ gauge, we can, for instance, require every $\hat{n}_v$ to live on the upper hemisphere which is sufficient to specify a nematic variable living on $\mathbb{R}P^2$.) With this model, we can understand the point raised before, that in what sense a link variable takes value in $\widetilde{\mathcal{T}}$ which is not a group in general. Consider a nematic order parameter pointing along $\pm \hat{n}_v$ at vertex v, and focus on, say, its $+\hat{n}_v$ end. Moving along the link l, this end will gradually move and reach some other direction in S^2, denoted by $\sigma_l \hat{n}_{v'} \in S^2$; correspondingly, the $-\hat{n}_v$ end will move and reach $-\sigma_l \hat{n}_{v'}$.

Next we move to gauge theory, i.e. 1-form theory. The mathematical structure for Villainization is the central extension of a group G, not necessarily abelian, to its universal covering group:
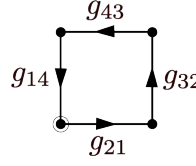
$$\begin{array}{c} \Gamma \to \widetilde{G} \\ \downarrow \\ G \end{array} \quad . \tag{32}$$

Here Γ has to be abelian because the Γ-valued field lives on plaquettes, and the composition of adjacent plaquettes has no specified order, unlike the links. An important example is $PSU(N)$ lattice gauge theory, which contains $\mathbb{Z}_N$ monopoles [31]. Recall $PSU(N) := SU(N)/Z(SU(N))$ where the the center of $SU(N)$ is $Z(SU(N)) = e^{i2\pi \mathbb{Z}_N/N} \mathbf{1}_{N \times N} \cong \mathbb{Z}_N$. Note that $PSU(N)$ has higher π_n's inherited from $SU(N)$ (the next non-trivial one being π_3 inherited from $SU(N)$, which is the main problem we will tackle in the work), but here we only focus on the π_1 effects, which arises from the mod-out of the center. The structure

$$\begin{array}{c} \mathbb{Z}_N \to SU(N) \\ \downarrow \\ PSU(N) \end{array} \tag{33}$$

can be implemented on the lattice as an $SU(N)$ gauge theory with the 1-form $Z(SU(N)) \cong \mathbb{Z}_N$ global symmetry gauged.

We first briefly review the traditional $SU(N)$ lattice gauge theory defined by Wilson [1–3]. The dynamical d.o.f. is $g_l \in SU(N)$ at each link l, thought of as a Wilson line along the link. The path integral is weighted by a plaquette weight which is a positive, increasing function of $(\mathbf{tr}Dg_p + c.c.)$, where the $SU(N)$ flux Dg_p is the Wilson loop around a plaquette, i.e. the ordered product of the g_l's around p starting from some chosen vertex:

$$Dg_p := g_{14}g_{43}g_{32}g_{21} \tag{34}$$

(for abelian group, $De^{ia}{}_p = e^{ida_p}$). Gauge transformation $g_l \to h_{v'}g_l h_v^{-1}$ (where $l = \langle v'v \rangle$) changes Dg_p by a conjugation, hence the weight remains invariant. [28] Similarly, choosing another starting vertex only changes Dg_p by a conjugation, which does not change the weight; the starting vertex can even be located away from the plaquette, as long as we conjugate the flux by a suitable Wilson line. The flux Dg_p satisfies $DDg_c = 1$ (lattice version of Bianchi identity) on any cube c, where the definition and why it equals 1 is illustrated by the picture
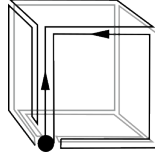
$$\tag{35}$$

A conceptual point, similar to that regarding the $\widetilde{\mathcal{T}}$-valued link variable before, is that now we have a $SU(N)$-valued plaquette variable Dg_p, which might seem problematic, because plaquette variables should be abelian as mentioned above. The solution is, this is not problematic because Dg_p is not an *independent* plaquette variable, it is defined via link variables starting from a chosen vertex, and composition between plaquettes can be defined accordingly using the conjugation of Wilson lines built out of link variables (for example see the pictorial definition of DDg).

The flux Dg_p respects the 1-form global symmetry $g_l \to g_l z_l$ for $z_l \in Z(SU(N)) \cong \mathbb{Z}_N$ satisfying $Dz_p = 1$. This is what we will gauge, in order to obtain the Villainized $PSU(N)$ gauge theory [31]:

$$Z = \left[\prod_{l'} \int_{g_{l'} \in SU(N)} \right] \left[\prod_{p'} \sum_{\sigma_{p'} \in Z(SU(N))} \right] \prod_p W_2(\mathbf{tr}(\sigma_p Dg_p) + c.c.) \prod_c W_3(D\sigma_c) . \tag{36}$$

Here W_2 and W_3 are some positive, increasing functions, and $D(\sigma Dg)_c = D\sigma_c := \prod_{p \in \partial_c} \sigma_p$, with the orientations of p here chosen to be consistent with that of ∂c, describes the $\mathbb{Z}_N$

---

[28]Therefore, the weight does not have to depend on Dg_p through the trace, but through any function of the eigenvalues.

monopole. (We can also use W_3^{forbid} by introducing a $\mathbb{Z}_N$ Lagrange multiplier to forbid the $\mathbb{Z}_N$ monopoles.) We can also define $\mathbb{Z}_N$ skyrmion and the associated $\mathbb{Z}_N$ topological theta term over a 2d surface [33]. For $N = 2$ the $\mathbb{Z}_N$ skyrmion configuration represents the second Stiefel-Whitney class of the $PSU(2) \cong SO(3)$ gauge field. [29]

Moving further up to higher form gauge theories, since the $q \geq 2$ form independent variables can only be abelian for reasons explained above, only abelian examples exist, which have already been discussed at the beginning of this subsection.

This concludes what Villainization in its general form can do. It captures the π_1 of nlσm target spaces or gauge groups by taking universal covers. An important step in furthering the understanding of Villainization appeared in [32,33], which turned out to be an important inspiration for our present work. The physical context there is to study the possible low energy phases of Yang-Mills theory. Under this context, the Villainized $PSU(N)$ gauge theory is interpreted in terms of the Lie 2-group $PSU(N) \times SU(N) \rightrightarrows PSU(N) \rightrightarrows *$. Importantly, [32,33] shows the low energy phases of Yang-Mills theory admit more possibilities, which are described by more general Lie 2-groups gauge theories [45], $G \ltimes H \rightrightarrows G \rightrightarrows *$, in which H might not fully cover G, leading to the exact sequence $* \to \ker(\tilde{t}) \to H \xrightarrow{\tilde{t}} G \to \mathrm{coker}(\tilde{t}) \to *$. We will review such structure in Sections 5.1 and 5.3.

2.4 Spinon-decomposed S^2 non-linear sigma model: Berry phase, skyrmion and hedgehog

The Villain model and all its variants serve to capture the π_1 of continuous-valued fields. There is another type of known examples, the $\mathbb{C}P^1$ representation, also known as the spinon decomposition, for S^2 nlσm, which captures $\pi_2(S^2) \cong \mathbb{Z}$; this can be generalized to capture π_2 of more general target spaces such as $\mathbb{C}P^N$. Beyond these examples, there is no more known example that captures higher π_n's of continuous-valued fields on the lattice, and we will explain why in Section 3. In this subsection we review how the $\mathbb{C}P^1$ representation works. It will bring out more discussions about the geometrical intuition as well as some technical points, which will be useful for our construction in Section 4 and beyond.

A traditional S^2 nlσm on lattice has a unit vector $\hat{n}_v \in S^2$ at each vertex v, and the link weight is a positive increasing function $W(\hat{n}_{v'} \cdot \hat{n}_v)$. Again there are topological configurations that cannot be naturally defined in the traditional lattice model: the skyrmion in 2d, and the hedgehog defect in 3d or above (that can be seen as the non-conservation of skyrmion number in the 2d space over time), which are characterized by $\pi_2(S^2) \cong \mathbb{Z}$. In fact, even around a 1d loop there is an important piece of physics that cannot be naturally defined, the Berry phase around the loop, whose 2π periodicity is due to the same topological information

---

[29]In fact, even if we have used W_3^{forbid} to forbid the $\mathbb{Z}_N$ monopoles, there remains a new piece of interesting topological effect in $d \geq 3$. We can Villainize $\sigma_p =: e^{i2\pi s_p/N}$ by introducing an $Nh_c \in N\mathbb{Z}$, forming $N\mathbb{Z} \to \mathbb{Z} \to \mathbb{Z}_N$, where the $\mathbb{Z}$ variable $ds_c - Nh_c$ is invariant under $s_p \to s_p + Nn_p, h_c \to h_c + dn_c$. Then W_3^{forbid} is enforcing that there exists some h_c such that $ds_c - Nh_c = 0 \in \mathbb{Z}$. While this implies h_c is closed, it might not be exact in $\mathbb{Z}$ because in general $s_p/N \notin \mathbb{Z}$. Therefore we can have a closed, non-exact h_c topological configuration; for $N = 2$ it represents the third integral Stiefel-Whitney class of the $PSU(2) \cong SO(3)$ gauge field, which will be important later in footnote 35.

$\pi_2(S^2) \cong \mathbb{Z}$. [30]

Viewing S^2 as $\mathbb{C}P^1 := \mathbb{C}^2/\mathbb{C}_* \cong SU(2)/U(1)$ solves this problem [78–81]. This $\mathbb{C}P^1$ representation was originally developed and much more well-known in the continuum context, but on the lattice it becomes more crucial for capturing topology. Algebraically, the idea is obvious: to cover S^2 by $U(1) \to SU(2) \to S^2$, and then Villainize the $U(1)$:

$$
\begin{array}{c}
2\pi\mathbb{Z} \to \mathbb{R} \\
\downarrow \\
U(1) \to SU(2) \\
\downarrow \\
S^2
\end{array}
\qquad (37)
$$

The sequence of fibre bundles leads to the sequence of isomorphisms $\pi_2(S^2) \xrightarrow{\sim} \pi_1(U(1)) \xrightarrow{\sim} \pi_0(\mathbb{Z})$. The Berry phase is captured at the $U(1)$ stage, while the skyrmion and hedgehog are captured at the last stage.

The implementation goes as the following. Across the link $l = \langle v'v \rangle$, we introduce an $SU(2)$ variable $\mathcal{V}_l \in SU(2)$ that rotates $\hat{n}_v$ to $\hat{n}_{v'}$, i.e. subjected to the constraint $R_{\mathcal{V}_l}\hat{n}_v = \hat{n}_{v'}$, where $R_{\mathcal{V}_l}$ is the rotation matrix by casting $\mathcal{V}_l$ in the spin-1 representation. Equivalently, we can say the constraint is $\mathcal{V}_l(\hat{n}_v \cdot \vec{\sigma})\mathcal{V}_l^{-1} = \hat{n}_{v'} \cdot \vec{\sigma}$. This is like the constraint $e^{i\gamma_l}e^{i\theta_v} = e^{i\theta_{v'}}$ in the Villain model. This constraint does not uniquely fix $\mathcal{V}_l$ but leaves a $U(1)$ d.o.f., because after a given rotation we can still make a rotation around $\hat{n}_{v'}$ without changing $\hat{n}_{v'}$. In the spin-1/2 representation, the constraint implies $\mathcal{V}_l u_{\hat{n}_v} = e^{-ia_l}u_{\hat{n}_{v'}}$, where $\hat{n}_v = u_{\hat{n}_v}^\dagger \vec{\sigma} u_{\hat{n}_v}$ (the $u_{\hat{n}_v}$ is called *spinon*, and therefore this $\mathbb{C}P^1$ representation is also called spinon decomposition), and e^{ia_l} is the said $U(1)$ d.o.f., with $2a_l$ being the rotation angle around $\hat{n}_{v'}$. This dynamical e^{ia_l} part is then viewed as a $U(1)$ gauge field, which we will Villainize as we did in Section 2.2. The partition function reads:

$$
Z = \left[\prod_{v'} \int_{S^2} \frac{d^2\hat{n}_{v'}}{4\pi}\right] \left[\prod_{l'} \int_{-\pi}^{\pi} \frac{da_{l'}}{2\pi}\right] \left[\prod_{p'}\sum_{s_{p'}\in\mathbb{Z}}\right] \prod_l W_1(\mathrm{tr}\mathcal{V}_l + c.c.) \prod_p W_2(f_p) \prod_c W_3(m_c)
$$

$$(38)$$

where W_1 is positive and increasing, W_2, W_3 are positive and decreasing with the absolute value of the arguments (or we can use W_3^{forbid}). (This is for $d \geq 3$. For $d = 2$ just ignore the W_3 part. For $d = 1$ ignore the s_p field and the W_2 and W_3 parts.) The skyrmion configuration and hedgehog defect of the S^2 d.o.f. are then defined as the Dirac quantized flux and monopole of the $U(1)$ gauge theory. In particle physics, this is the familiar situation of an $SU(2)$ gauge field being Higgsed by an S^2 vacua down to a residual $U(1)$ gauge field [82,83]; the constraint $R_{\mathcal{V}_l}\hat{n}_v = \hat{n}_{v'}$ means the massive gauge bosons are set to be infinitely massive.

More explicitly, for $\hat{n}_v$ given by the spherical coordinates (θ_v, ϕ_v), it is common to make

---

[30]In the previous S^1 nlσm, in 0d there is also a phase, the $e^{i\theta}$ itself, which is well-defined without Villainization. In $U(1)$ gauge theory, in 1d there is also a phase, the Wilson loop $\prod_l e^{ia_l}$, which is again well-defined without Villanization. So the S^2 nlσm is the first example where some physical phase requires suitable refinement of the traditional lattice theory to be well-defined.

a standard choice of $SU(2)$ matrix $\mathcal{U}_{\hat{n}_v}$ whose spin-1 representation would rotate $\hat{z}$ to $\hat{n}_v$:

$$\mathcal{U}_{\hat{n}_v} = e^{-i\sigma^z \phi_v/2} e^{-i\sigma^y \theta_v/2} e^{i\sigma^z \phi_v/2} = \begin{bmatrix} \cos(\theta_v/2) & -e^{-i\phi_v}\sin(\theta_v/2) \\ e^{i\phi_v}\sin(\theta_v/2) & \cos(\theta_v/2) \end{bmatrix} = \begin{bmatrix} u_{\hat{n}_v} & -i\sigma^y u_{\hat{n}_v}^* \end{bmatrix} .$$
(39)

This is like fixing $\theta_v \in (-\pi, \pi]$ in the Villain model. Then $\mathcal{V}_l$ can be parametrized as $\mathcal{V}_l = \mathcal{U}_{\hat{n}_{v'}} e^{-i\sigma^z a_l} \mathcal{U}_{\hat{n}_v}^{-1}$, where $e^{-ia_l} \in U(1)$ is a new dynamical variable not fixed by the constraint $R_{\mathcal{V}_l} \hat{n}_v = \hat{n}_{v'}$; indeed, it manifests in $\mathcal{V}_l u_{\hat{n}_v} = e^{-ia_l} u_{\hat{n}_{v'}}$. If we change the standard choice of $\mathcal{U}_{\hat{n}_v}$ by a $U(1)$ gauge transformation $\mathcal{U}_{\hat{n}_v} e^{i\psi_v \sigma^z}$, accompanying it by $a_l \to a_l + d\psi_l$ leaves $\mathcal{V}_l$ unchanged. The apparent singularity in $\mathcal{U}_{\hat{n}_v}$ at $\theta_v = \pi$ (due to the ambiguity of ϕ_v) is like the apparent but not physically harmful discontinuity between $\theta_v = \pm\pi$ in the Villain model—$\mathcal{V}_l$ has nothing singular, just like γ_l has nothing discontinuous in the Villain model. [31]

Now we want to show that e^{ia_l} should be naturally interpreted as the Berry connection across the link, so that the Berry phase $e^{i\Phi}$ around a loop is given by

$$e^{i\Phi} := e^{i\oint_{1d} a} = \prod_l e^{ia_l} \quad \text{or equivalently} \quad e^{-i\vec{\sigma}\cdot\hat{n}_v \Phi} := \prod_l \mathcal{V}_l \quad \text{(path ordered staring from } v) ,$$
(40)

and the Berry curvature is the Berry phase around a single plaquette, $e^{if_p} := e^{ida_p}$, i.e. $e^{-i\vec{\sigma}\cdot\hat{n}_v f_p} = D\mathcal{V}_p$. Thus, a 1d worldloop weighted by the Berry phase reads

$$Z_{qBerry} = \left[\prod_{v'} \int_{S^2} \frac{d^2\hat{n}_{v'}}{4\pi} \right] \left[\prod_{l'} \int_{-\pi}^{\pi} \frac{da_{l'}}{2\pi} \right] e^{iq\Phi} \prod_l W_1(\mathbf{tr}\mathcal{V}_l + c.c.) .$$

for any $q \in \mathbb{Z}$. This is actually the simplest non-trivial case of putting a 1d *coadjoint orbit theory* [71] onto the lattice. We will discuss more about coadjoint orbit theories on lattice in subsequent works. For odd q, the $SO(3)$ global symmetry becomes anomalous and is extended to $SU(2)$, and the interpretation is familiar: the total Berry phase over the sphere being $2\pi q$ means the spin is $q/2$. [32]

To check this Berry connection interpretation, consider the spinon decomposed link weight (ignoring the W_2, W_3 dependence on a_l for now)

$$W(\hat{n}_{v'} \cdot \hat{n}_v) \approx \int_{-\pi}^{\pi} \frac{da_l}{2\pi} W_1(\mathbf{tr}\mathcal{V}_l + c.c.)$$
(41)

---

[31]If we still want to remove this unharmful singularity, we can simply leave the $U(1)$ gauge unfixed, as it only contributes a finite factor at each vertex. Thus $\mathcal{U}_{\hat{n}_v}$ is any $SU(2)$ element that rotates $\hat{z}$ to $\hat{n}_v$, which corresponds to multiplying an arbitrary $e^{i\sigma^z \psi_v/2}$ to the right of our standard choice (39). This is the common practice in the numerical implementation of [78, 79]. On the other hand, in most implementations, the subsequent Villainization step is not adapted, despite the existing proposals [80, 81].

[32]To manifestly see the anomaly on lattice, we can introduce a background gauge field $V_l \in SU(2)$, which appears in W_1 as $\mathbf{tr}\mathcal{V}_l \to \mathbf{tr}(\mathcal{V}_l V_l^\dagger)$. If the background is $SO(3) \cong PSU(2)$, then V_l and $-V_l$ must be equivalent, and this is realized by an $e^{i\pi}$ shift of e^{ia_l}, which leaves Z_{qBerry} invariant only for even q.

where both W and W_1 are positive increasing functions. It is technically useful to express $\mathbf{tr}\mathcal{V}_l$ in terms of the spinons (and the result turns out to be the hopping of spinons):

$$\mathbf{tr}\mathcal{V}_l = \mathbf{tr}\mathcal{V}_l^\dagger = e^{ia_l}u_{\hat{n}_{v'}}^\dagger u_{\hat{n}_v} + c.c. \ . \tag{42}$$

The dominating contribution to W_1 comes from $e^{ia_l} \approx u_{\hat{n}_v}^\dagger u_{\hat{n}_{v'}}/|u_{\hat{n}_v}^\dagger u_{\hat{n}_{v'}}|$. When $\hat{n}_v$ and $\hat{n}_{v'}$ are close to each other, we have $a \approx -iu^\dagger du$, recovering the familiar expression for Berry connection in continuum. On the other hand, when $\hat{n}_v$ and $\hat{n}_{v'}$ are nearly opposite to each other, $u_{\hat{n}_{v'}}^\dagger u_{\hat{n}_v} \to 0$ and W_1 becomes insensitive to e^{ia_l}. [33]

It is better to understand the general situations in geometrical terms.
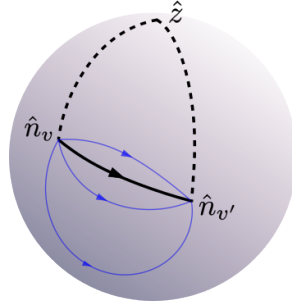
$$\tag{43}$$

Clearly $\mathbf{tr}\mathcal{V}_l$ is maximized when the rotation angle in $R_{\mathcal{V}_l}$ is minimized, and this is when it is a rotation around the $\hat{n}_v \times \hat{n}_{v'}$ axis by an angle of $\arccos|\hat{n}_v \cdot \hat{n}_{v'}|$. When the rotation angle gradually increases from 0 to $\arccos|\hat{n}_v \cdot \hat{n}_{v'}|$, the $\hat{n}$ is brought along the black solid curve, which is the shortest geodesic on the sphere. When $\mathbf{tr}\mathcal{V}_l$ is not maximized, $\mathcal{V}_l$ will bring the $\hat{n}$ along some other rotational path with a larger rotation angle, such as the blue ones. [34] The two black dashed paths are associated with our choice of $\mathcal{U}_{\hat{n}}$, i.e. when the rotation angle in $\mathcal{U}_{\hat{n}}$ gradually increases from 0 to θ, the $\hat{n}$ is brought along a dashed path; the path that corresponds to our standard choice of $\mathcal{U}_{\hat{n}}$ turns out to be the shortest geodesic. A generic (blue) rotational path associated with $\mathcal{V}_l$, together with the two dashed paths associated with $\mathcal{U}_{\hat{n}_v}$ and $\mathcal{U}_{\hat{n}_{v'}}$, bound a solid angle whose value turns out equal to twice the rotation angle of $\mathcal{U}_{\hat{n}_v}^{-1}\mathcal{V}_l^{-1}\mathcal{U}_{\hat{n}_{v'}}$, which is $2a_l$ by definition. Thus, the value of e^{ia_l} that maximizes $\mathbf{tr}\mathcal{V}_l$ and hence the weight, given by $u_{\hat{n}_v}^\dagger u_{\hat{n}_{v'}}/|u_{\hat{n}_v}^\dagger u_{\hat{n}_{v'}}|$, is geometrically given by half of the solid angle of the area bounded by the black solid geodesic and the two dashed geodesics. The sensitivity of the weight to changes in a_l, say characterized by the second derivative, is proportional to $|u_{\hat{n}_v}^\dagger u_{\hat{n}_{v'}}| = \sqrt{(\hat{n}_v \cdot \hat{n}_{v'} + 1)/2}$.

One can notice that the value of e^{ia_l} maximizing $\mathbf{tr}\mathcal{V}_l$ sees two kinds of singularities, one is an artifact and the other is meaningful. The first kind of singularity occurs when either of $\hat{n}_v, \hat{n}_{v'}$ is in the vicinity of $-\hat{z}$. In this case, the associated dashed curve and hence the

---

[33]This is not the minimum of W_1, because $\mathbf{tr}\mathcal{V}_l + c.c.$ can take negative values for "bad choices" of e^{ia_l}.

[34]The rotational paths are allowed to have a rotation angle larger than π, and this is desired. In $SO(3)$, a rotation around some axis $\hat{m}$ by some amount $\alpha \in [0, \pi]$ is the same as a rotation around $-\hat{m}$ by $2\pi - \alpha$, but they correspond to opposite elements in $SU(2)$ (differing by $2a_l \to 2a_l + 2\pi$), and this is manifested by the fact that the associated curves are different and can join into a circle of 2π rotation. Thus, the $U(1)$ d.o.f. in $\mathcal{V}_l$ is geometrically seen by the fact that the mid point of the rotational path can be anywhere on the great circle about which $\hat{n}_v$ and $\hat{n}_{v'}$ are symmetric.

solid angle changes rapidly. But this is an artifact due to our standard gauge choice for $\mathcal{U}_{\hat{n}}$; in the gauge invariant Berry phase, which only depends on the actual rotational paths associated with $\mathcal{V}_l$, the dependence on the dashed paths will cancel out anyways. Such kind of artifact is unavoidable if we use the ($U(1)$ gauge fixed) $\mathcal{U}_{\hat{n}_v}$ and e^{ia_l} to parametrize $\mathcal{V}_l$, because $SU(2)$ is a non-trivial $U(1)$ bundle over S^2 (but this is crucial, making it possible to capture $\pi_2(S^2) \xrightarrow{\sim} \pi_1(U(1))$). This is like, in the Villain model, the most probably choice of m_l will jump when either of θ_v, θ'_v moves across $\pm\pi$ when the $\mathbb{Z}$ gauge is fixed to $\theta_v \in (-\pi, \pi]$, but γ_l does not jump. We would also like to remark that the Berry phase (40) can be defined from $\mathcal{V}_l$ directly without referring to the $\mathcal{U}_{\hat{n}_v}$ and e^{ia_l} parametrization.

The second kind of singularity occurs when $\hat{n}_v$ and $\hat{n}_{v'}$ are nearly opposite. In this case, the black solid geodesic between them changes rapidly; as the two points become exactly opposite, there is no unique choice of the shortest geodesic, hence no unique choice for the most probable e^{ia_l}. Such singularity occurring in the most probable choice of e^{ia_l} does not mean any physical observable becomes singular. Rather, it simply means all choices of rotational paths, parametrized by e^{ia_l}, become equally probable, as we would intuitively expect when $\hat{n}_v$ and $\hat{n}_{v'}$ become opposite. Indeed, $\mathbf{tr}\mathcal{V}_l$ and hence the weight W_1 becomes insensitive to e^{ia_l} as $|u_{\hat{n}_{v'}}^\dagger u_{\hat{n}_v}| \to 0$. This is like, in the Villain model, when $e^{i\theta_v}$ and $e^{i\theta_{v'}}$ become opposite, γ_l taking $\pm\pi$ become equally probable.

Having understood the Berry phase, the skyrmion and hedgehog become easy to understand. The skyrmion configuration is when the configuration of $\hat{n}$ over a 2d space wraps around the S^2; the Berry curvature thus accumulates to 2π, a Dirac quantized flux of Berry curvature. From this we can identity the Berry curvature $f_p := da_p + 2\pi s_p$ as the skyrmion density, and define a topological theta term in 2d. The hedgehog defect is a skyrmion around a single cube, and hence counted by the Berry curvature monopole $m_c = df_c/2\pi = ds_c$. If we use W_3^{forbid} to forbid the hedgehogs, there will be a dual $(d-3)$-form $U(1)$ global symmetry, and we can explicitly see on the lattice that it has the celebrated mixed anomaly with the 0-form $SO(3)$ global symmetry of the S^2. [35]

---

[35]This anomaly can be seen by introducing an $SO(3)$ background gauge field and finding that any consistent modification to the definition of the hedgehog breaks the dual $U(1)$. Alternatively, it can be seen by introducing a dual $U(1)$ background and finding that along its background Dirac string there is a $q = 1$ Berry phase integral (40), extending the $SO(3)$ global symmetry to $SU(2)$ according to footnote 32. Below we focus on the first route.

The $SO(3)$ background gauge field appears in W_1 as $\mathbf{tr}\mathcal{V}_l \to \mathbf{tr}(\mathcal{V}_l V_l^\dagger)$, where the background field $V_l \in SU(2)$. But since the background should really be $SO(3) \cong PSU(2)$ rather than $SU(2)$, somehow V_l and $-V_l$ must be equivalent. In W_1 we can absorb this sign ambiguity into e^{ia_l}. But then in W_2, the flux f_p is ambiguous by π. The solution is to introduce a 2-form $\mathbb{Z}_2$ background $S_p \in \mathbb{Z}$ mod 2 that absorbs this ambiguity, so that the modified flux $f_p - \pi S_p$ is unambiguous (note that the $2\mathbb{Z}$ ambiguity in S_p needs to be absorbed by s_p) and can be used in W_2. What has happened is that the 2-form $\mathbb{Z}_2$ background S_p effectively reduces the 1-form $SU(2)$ background V_l to $PSU(2) \cong SO(3)$, just like in (33) for dynamical gauge fields.

Interestingly, the skyrmion number $\oint_{2d}(f - \pi S)/2\pi$ becomes half-quantized. This is true even if we have demanded S_p to be $\mathbb{Z}_2$ closed, i.e. $dS_c = 0$ mod 2, because the $\oint_{2d}$ can be around a non-contractible 2d surface. In fact, $\oint_{2d} S/2 := \sum_p S_p/2$ mod 1 characterizes the second Stiefel-Whitney class of the $SO(3)$ background. The flux $f_p - \pi S_p$ is no longer a $U(1)$ gauge flux, but a spin-c gauge flux associated with the $SO(3)$ background. Therefore in 2d we can have a non-trivial $\mathbb{Z}_2$ topological theta term coupled to this half-quantized spin-c flux, realizing the celebrated Haldane quantum spin chain phase [84, 85].

In W_3, $df_c/2\pi = ds_c$ is no longer a good hedgehog, since a $2\mathbb{Z}$ ambiguity in S_p must be absorbed by

It is straightforward to generalize this construction to $\mathbb{C}P^N$ nlσm. The spinon u_v will become an $(N+1)$-component complex unit vector taking value in $S^{2(N+1)-1}$, and in a $S^{2(N+1)-1}$ nlσm the link is weighted by $u_{v'}^\dagger u_v + c.c.$. Gauging the diagonal $U(1)$ phase global symmetry leads to the spinon-decomposed $\mathbb{C}P^N$ nlσm, and the $U(1)$ gauge field is then Villainized. Generalizations can also be made to capture the π_2 of spaces beyond $\mathbb{C}P^N$.

A more interesting direction of generalization is to consider $\mathbb{R}P^2 \cong S^2/\mathbb{Z}_2$. In Section 2.3 we have capture its π_1, and now we can capture its π_2 inherited from the S^2. The spinon decomposition becomes $\mathbb{Z}_2 \ltimes U(1) \to SU(2) \to \mathbb{R}P^2$, where the gauge group is no longer abelian. Upon further Villainizing the $U(1)$ part, the $\pi_1 \cong \mathbb{Z}_2$ will act on the $\pi_2 \cong \mathbb{Z}$ by flipping the sign, leading to a non-trivial 2-group structure [86]. We will discuss such interplay between different π_n's in subsequent works.


For now, we want to return to and think more closely about the geometrical intuition behind the $\mathbb{C}P^1$ representation, and in particular its relation to continuum nlσm. Understanding this will turn out important for motivating our main constructions in this work.

Recall in the Villain model in Section 2.1, we motivated the introduction of $\gamma_l \in \mathbb{R}$ by thinking about the link l as being embedded in the continuum and the γ_l representing the length a path in S^1 emanating from θ_v and gradually reaching some $\theta_{v'}$. Now, the paths in S^2 emanating from $\hat{n}_v$ can take all kinds of shapes and form an infinite dimensional space, and our task is to understand why it is sufficient for our purpose to "truncate away the unimportant details of how a generic path wiggles" and only consider the rotational paths parametrized by a finite dimensional space $SU(2)$.

The intuitive explanation is the following. As we have seen above, what really matters for the defining the topological operators is the Berry phase around a closed loop, given by half of the solid angle bounded; the skyrmion density and hedgehog defect are defined from the Berry curvature. Thus, the wiggling details of a path in S^2 are indeed unimportant, and we can represent many different wiggling paths by a same representative rotational path, as long as the solid angle bounded between the original wiggling path and the representative

---

s_p. The unambiguous hedgehog defect should become $m_c := d(f - \pi S)_c/2\pi = ds_c - dS_c/2$; this is still an integer if we have demanded $dS_c = 0 \bmod 2$. In fact, it is better to describe the condition $dS_c = 0 \bmod 2$ along the lines of footnote 29, i.e. to introduce a $2H_c \in 2\mathbb{Z}$ background to form $2\mathbb{Z} \to \mathbb{Z} \to \mathbb{Z}_2$, such that the combination $dS_c - 2H_c \in \mathbb{Z}$ is unambiguous and is enforced to be 0 everywhere. Then the good hedgehog is $m_c := ds_c - H_c$. While $H_c = dS_c/2$ shows H_c is closed, it might not be exact in $\mathbb{Z}$ since in general $S_p/2 \notin \mathbb{Z}$; this characterizes the third integral Stiefel-Whitney class of the $SO(3)$ background. Note that demanding the third integral Stiefel-Whitney class to vanish is a non-local demand, in contrast to the previous closedness condition $dS_c = 0 \bmod 2$ which is local. When the third integral Stiefel-Whitney class is indeed non-trivial, if we still use W_3^{forbid} for W_3, the dual $U(1)$ global symmetry is explicitly broken, leading to a vanishing partition function, because there is no solution of s_p that can possibly make $m_c = ds_c - H_c$ vanish everywhere. This is the familiar fact that a spin-c gauge field cannot be free from monopole if the associated third integral Stiefel-Whitney class is non-trivial.

rotational path is 0, for instance
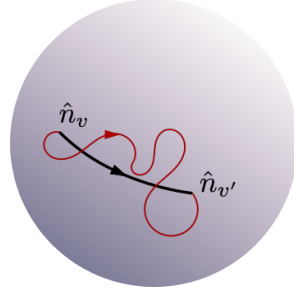
$$\tag{44}$$

Topologically, the space of all paths interpolating from $\hat{n}_v$ to $\hat{n}_{v'}$, though infinite dimensional, can be easily seen to have a $\pi_1 \cong \mathbb{Z}$, if we think of a path as a rubber band and wrap it around the sphere. When effectively reduced to the rotational paths from $\hat{n}_v$ to $\hat{n}_{v'}$, the space becomes $U(1)$, which keeps the π_1 topological information.

This intuitive explanation can be casted into mathematical terms, allowing us to relate the algebraic motivation (37) and the geometrical intuition. Let $\mathcal{P}_* X$ and $\Omega_* X$ be the pointed path space and pointed loop space of X respectively, i.e. the spaces of (parameterized) paths and loops starting from a given point. The space of all paths emanating from a given starting point is by definition $\mathcal{P}_* X$, while the space of all paths given both the starting and ending point is homeomorphic to $\Omega_* X$, where the loop is formed by returning from the ending point to the starting point via some standard choice of path. Then obviously, for the S^1 Villain model,

$$\mathcal{P}_* S^1 \xrightarrow{length} \mathbb{R} \ni \gamma_l, \qquad \Omega_* S^1 \xrightarrow{length} 2\pi\mathbb{Z} \ni 2\pi m_l \tag{45}$$

by taking the (signed) length of the image of the path. Hence the Villainization fibre bundle $2\pi\mathbb{Z} \to \mathbb{R} \to S^1$ can be naturally recognized as $\Omega_* S^1 \to \mathcal{P}_* S^1 \to S^1$ after taking the length, or say, ignoring the parametrization.

For S^2, we not only need to think about the continuum paths traced out by the links, but also the continuum surfaces—i.e. paths between paths—swept out by the plaquettes. The intuitive discussion above suggests that the important information on a continuum surface swept out by a plaquette is its solid angle, i.e. the integral of the continuum Berry curvature over the surface. This means the continuum field is reduced to the lattice field via

$$
\begin{array}{ccc}
\Omega_*^2 S^2 \to \mathcal{P}_* \Omega_* S^2 & \xrightarrow{\int_{2d} \text{Berry}} & 2\pi\mathbb{Z} \to \mathbb{R} \\
\downarrow & & \downarrow \\
\Omega_* S^2 \to \mathcal{P}_* S^2 & & U(1) \to SU(2) \\
\downarrow & & \downarrow \\
S^2 & & S^2
\end{array}
\tag{46}
$$

The space of (topologically trivial) 2d surfaces emanating from a given path between two given points is homeomorphic to $\mathcal{P}_* \Omega_* S^2$; an element of it, geometrically a 2d surface ((topologically trivial and with open boundary) on S^2, is thought of as being swept out by a plaquette embedded in the continuum. Integrating such a surface with the continuum Berry

curvature results in a value in $\mathbb{R}$; in particular $\Omega^2 S^2$, the space of closed 2d surface on S^2, indeed maps to $2\pi\mathbb{Z}$, the quantized total Berry phase determined by the winding number around the sphere. This explains the map at the top row, which are fields on the plaquettes. Induced from that, a loop on S^2, i.e. an element of $\Omega_* S^2$, formed by an arbitrary path and a standard choice of path connecting two given points, is then mapped to the $U(1)$ Berry phase bounded by these two paths (the $2\pi\mathbb{Z}$ part is not determined because which side is the bounding surface is not chosen); the dependence on the standard choice of path leads to the gauge transformation of the Berry connection. Consequently, an arbitrary path from a given starting point has an induced map to an $SU(2)$ rotational path in the way pictured in (44); we can say $SU(2) \cong \mathcal{P}_* S^2 \times U(1)/Berry$, where the latter space means two elements of $\mathcal{P}_* S^2 \times U(1)$ are considered equivalent if the the two paths in $\mathcal{P}_* S^2$ share the same starting and ending points, and moreover they bound a Berry phase that is equal to the difference in the two $U(1)$ phases.

In our main problem later, we will no long have the familiar language of Lie groups and fibre bundles, but such a picture of "truncating away the unimportant details" from the infinite dimensional space of continuum fields is what we need in order to understand how to think about and work with the more flexible yet unfamiliar categorical structures.

3 Difficulty beyond the Known Examples

The examples reviewed in Section 2 have all been worked out before. Yet what we uncovered through our review is the *relation* between the examples. They are not scattered; rather, they are organized by the same rationale, capable of capturing the π_1 and π_2 of target spaces (or gauge groups), and moreover the rationale makes connection to the continuum.

With this, we can now understand why similar efforts trying to capture $\pi_{n\geq 3}$ (such as skyrmion in pion nlσm and instanton in Yang-Mills) onto the lattice have not been successful. It is not because of bad luck. It will become clear in this section that it is mathematically impossible to achieve this goal within the familiar languages of Lie groups and fibre bundles. More flexible mathematical structures become necessary. These structures are not so easy to come up with by regular attempts; or even if one comes up with them by good physical intuition, the structures might seem "not mathematically nice enough" to be convincing. In fact, more systematic mathematical considerations will naturally lead to these structures, which will end up being physically intuitive.

Let us now think about S^3 nlσm, which can describe the pion vacua. The skyrmion configuration is now over the 3d space, characterized by $\pi_3(S^3) \cong \mathbb{Z}$, and represents the baryons over the pion vacua [87]. The hedgehog defect in 4d represents the non-conservation of baryons, which we might want to be able to forbid on the lattice. Over a 2d space, we can also define a $U(1)$ phase, the Wess-Zumino-Witten (WZW) term, just like the Berry phase in S^2 nlσm. Of course, S^3 also has higher π_n's (e.g. the 4d WZW term is due to π_5), but in this work we will only focus on the physics due to π_3, the smallest non-trivial π_n.

In the continuum, for a field $g(x) \in SU(N)$ (with $|SU(2)| \cong S^3$, here $|G|$ means the manifold of a Lie group G), the WZW curvature, a 3-form analogue of the Berry curvature, is defined as $\mathbf{tr}[(g^{-1}dg)^3]/6(2\pi)^2$, which integrates to an integer—the skyrmion number—

over a closed 3d manifold. The WZW curvature can be written as the exterior derivative of the WZW curving, a 2-form analogue of the Berry connection, which will not be globally well-defined if the skyrmion number is non-zero. Integrating the WZW curving over a closed 2d manifold yields the WZW term. We will review some technical details at the beginning of Section 4.

We show below that it is mathematically impossible to naturally define these π_3 related topological operators in S^3 nlσm on the lattice, if we use the usual Lie group or fibre bundle approaches. Of course, our original motivating problem is $SU(N)$ lattice Yang-Mills theory, not $|SU(N)|$ lattice nlσm (with $|SU(2)| \cong S^3$ the pion effective theory). The relation between the two is like the $U(1)$ gauge theory in Section 2.2 versus the S^1 nlσm in Section 2.1. They are, roughly speaking, related by "putting everything in one higher dimension". Thus, if we have demonstrated the said impossibility for S^3 lattice nlσm, the same must also be true for $SU(N)$ lattice Yang-Mills.

Before our full analysis of the problem, let us first discuss the role played by global symmetry. We are bringing this up because in the S^1 nlσm, the Villainization involved elevating S^1 to $\mathbb{R}$, the universal cover of the $U(1)$ global symmetry, and in S^2 nlσm, the spinon decomposition involved elevating S^2 to $SU(2)$, the universal cover of the $SO(3)$ global symmetry. This might generate a misleading impression that looking at (the universal cover of) the global symmetry is the key. But this is not true. For $|SU(N)|$ nlσm (with $|SU(2)| \cong S^3$), denoting a field by g, the continuous part of the global symmetry is $SU(N)_L \times SU(N)_R / Z(SU(N)) \cong PSU(N)_C \times SU(N)'_R$, manifested as

$$g \to h_L g h_R^{-1} = h_C g h_C^{-1} h_R'^{-1}, \tag{47}$$

$$(h_L, h_R) \sim (h_L z, h_R z), \text{ i.e. } (h_C, h_R') \sim (h_C z, h_R') \text{ for any } g \in SU(N), z \in Z(SU(N)) ,$$

and the universal cover of it is $SU(N)_L \times SU(N)_R$. But now we see that $SU(N) \to SU(N) \times SU(N) \to SU(N)$ is a trivial bundle, unlike in the examples of S^1 and S^2 before ((15) and (37)). It does not serve the desired purpose of "transmitting the desired π_3 to the π_2 in the layer above".

Now we are ready to see the problem in full. Based on the rationale of how we captured π_1 and π_2 before, it seems in order to capture $\pi_3 \cong \mathbb{Z}$ we naively need some sequence of fibre bundles of the form

$$
\begin{array}{c}
2\pi\mathbb{Z} \to \mathbb{R} \\
\downarrow \\
U(1) \to \text{???} \\
\downarrow \\
\text{??} \to \text{?} \\
\downarrow \\
S^3
\end{array}
\tag{48}
$$

Moreover, we can even have an interpretation of what the top layers represent: The $2\pi\mathbb{Z}$ on the cubes sum over to the skyrmion number, the $\mathbb{R}$ on a cube represent the WZW curvature on lattice, and the $U(1)$ on a plaquette the WZW curving on the lattice; these are all desired. It seems all we need is to fill out the question marks. But this is impossible. Look at the

"??" slot. Topologically what we want is $\pi_3(S^3) \xrightarrow{\sim} \pi_2(``??") \xrightarrow{\sim} \pi_1(U(1)) \xrightarrow{\sim} \pi_0(\mathbb{Z})$. So in particular we want $\pi_2(``??") \cong \mathbb{Z}$. On the other hand, it is a link variable, so we traditionally want it to be a group-valued variable, so that the variable can be composed when we compose consecutive links. The contradiction is, finite dimensional Lie groups always have trivial π_2, so this rationale fails.

What if we relax the requirement that the "??" slot should be a group, and hope that we somehow can still make sense of it as a link variable? The familiar examples of finite dimensional fibre bundles in physics are mostly principal or associated bundles, i.e. the transition functions between the fibres are described by Lie group actions, so we still encounter the same failure. In fact, after knowing our final solution in Section 5.5 and looking back, it can be shown [37] at full generality that any finite dimensional fibre cannot serve the purpose of transmitting the topological information from the layer below to above, $\pi_3(S^3) \xrightarrow{\sim} \pi_2(``??") \xrightarrow{\sim} \pi_1(U(1))$.

Obviously the same failure occurs if we want to use this rationale to capture on the lattice any non-trivial $\pi_{n \geq 3}$ of general spaces.

Now, if we still want to solve our problem, we are left with two possibilities:

1. To work with infinite dimensional spaces.

2. To work with more flexible, finite dimensional structures beyond groups and fibre bundles.

Our very reason to be interested in lattice theories is the finite dimensionality of the local d.o.f. in the path integral, so of course our final solution will take the second route. However, it is important make connection to the first route, because the first route just points to the continuum theory itself.

Indeed, if we think of the lattice as being embedded in the continuum, the continuum field over the vertices, links, plaquettes and cubes organize into a fibre bundle sequence structure similar to that on the left panel of (46):

$$
\begin{array}{c}
\Omega_*^3 S^3 \to \mathcal{P}_* \Omega_*^2 S^3 \\
\downarrow \\
\Omega_*^2 S^3 \to \mathcal{P}_* \Omega_* S^3 \\
\downarrow \\
\Omega_* S^3 \to \mathcal{P}_* S^3 \\
\downarrow \\
S^3
\end{array}
\tag{49}
$$

where every layer except for the bottom is infinite dimensional, as is expected for a continuum theory. [36] At the top layer, similar to (46), we indeed can map a 3d volume in S^3 (the image of the continuum field over the region of a lattice cube) to $\mathbb{R}$ by integrating over the continuum

---

[36] Along two consecutive links, the two paths in S^3 compose by concatenation in the obvious way—some reparametrization of the new path is needed but that does not affect anything to be discussed below. For more systematically treatment, see Section 5.

WZW curvature, leading to

$$\Omega_*^3 S^3 \to \mathcal{P}_*\Omega_*^2 S^3 \xrightarrow{\int_{3d} \text{WZW}} 2\pi\mathbb{Z} \to \mathbb{R}$$
$$\downarrow \qquad\qquad\qquad\qquad\qquad\qquad \downarrow$$
$$\Omega_*^2 S^3 \qquad\qquad\qquad\qquad\qquad U(1) \tag{50}$$

where the right-hand-side reproduces the desired structure in (48). The problem is, unlike in (46), this is not sufficient to reduce the $\mathcal{P}_*\Omega_* S^3$ slot to anything finite dimensional, because this slot is expected to become a $U(1)$ bundle over $\Omega_* S^3$, but the base $\Omega_* S^3$ is still infinite dimensional. So more has to be done to truncate away the unimportant details there in order to obtain something finite dimensional. More exactly, after the previous integral with continuum WZW, the remaining fibre bundle sequence structure in the lower layers is

$$U(1) \to \frac{\mathcal{P}_*\Omega_* S^3 \times U(1)}{WZW}$$
$$\downarrow$$
$$\Omega_* S^3 \to \mathcal{P}_* S^3 \tag{51}$$
$$\downarrow$$
$$S^3$$

where $\mathcal{P}_*\Omega_* S^3 \times U(1)/WZW$ means, two elements in $\mathcal{P}_*\Omega_* S^3 \times U(1)$ are considered equivalent, if the two surfaces in $\mathcal{P}_*\Omega_* S^3$ share the same boundary, and moreover they together bound a volume whose WZW phase is equal to the difference between the two $U(1)$ phases [37, 38]. Our task is to recast this structure into a perspective that is more general than groups and fibre bundles—the perspective of category theory, and find a topologically equivalent but finite dimensional representative.

There is another idea, less geometrical and more algebraical, on what kind of infinite dimensional spaces we may want to use. In (15), $\mathbb{R}$ is the universal (i.e. 1-connected) cover of S^1, and in (37), $SU(2)$ is the 2-connected cover of S^2. [37] Then in (48) we might want the "?" slot to be the 3-connected cover over S^3. But 3-connected covers are in general infinite dimensional. Then the task would be to find finite dimensional structure that effectively plays the role of a 3-connected cover.

Naturally, the geometrical/continuum idea and the algebraic idea come to confluence. In fact, the structure (51) already plays the effective role of a 3-connected cover [37, 38] in the category theory sense. So no matter which idea we take, we are led to the task of finding a finite dimensional equivalence of this structure. Thus, the task has now become a well-posed mathematical problem—and whose answer turns out to be already known [35] in terms of multiplicative bundle gerbe [36]. The task of finding more general topological operators for more general continuous-valued lattice fields can be turned into well-posed mathematical problems in the same manner, and such relevance to physics provides a good motivation to study these more general mathematical problems.

---

[37] m-connected cover means a space whose $\pi_{m>n}$ are the same as the original given space while $\pi_{m\leq n}$ vanish. m-connected covers over a given space form the Whitehead tower.

4 Main Construction

In this section we will introduce the construction that allows us to define the 2d WZW term (not the 4d one) and 3d skyrmion in S^3 lattice nlσm, as well as the 3d CS term and 4d instanton in $SU(N)$ lattice Yang-Mills—which all originate from $\pi_3 \cong \mathbb{Z}$. The derivation process and the resulting structure lies in higher category theory, as mentioned before. However, to explicitly present the resulting structure, no knowledge of category theory is required—in the end, structures are described by a set of rules; the familiar Lie groups are also described by a set of rules, except the "rules of the game" we need now are more flexible than those for a group; anyways, these rules are all that is needed for a computer to carry out Monte-Carlo numerics. Therefore, in this section, we will first state these rules and explain the physical intuition behind, while the derivation and the systematic understanding in terms of higher category theory will be deferred to Sections 5 and 6.

We have explained in Section 3 that any fibre bundle covering S^3 or more generally $|SU(N)|$ cannot fulfill our goal. So let us now motivate what kind of covering, if not fibre bundle, we might need. Continuum theory provides a good hint. Consider an S^3 or more generally $|SU(N)|$ nlσm in the continuum, parametrized by $g(x) \in SU(N)$. How do we show the continuum integral of the WZW curvature $\oint_{3d} \mathbf{tr}[(-ig^{-1}dg)^3]/6(2\pi)^2$ is an integer, which can be interpreted as the skyrmion number? We can first diagonalize $g = \mathcal{U}e^{i\lambda}\mathcal{U}^{-1}$, and find

$$\frac{1}{6}\mathbf{tr}[(g^{-1}dg)^3] = d\left(\mathbf{tr}[\lambda(\mathcal{U}^{-1}d\mathcal{U})^2] - \frac{1}{2}\mathbf{tr}[e^{i\lambda}(\mathcal{U}^{-1}d\mathcal{U})e^{-i\lambda}(\mathcal{U}^{-1}d\mathcal{U})]\right)$$

$$= d\left(\mathbf{tr}[d\lambda(\mathcal{U}^{-1}d\mathcal{U})] - \frac{1}{2}\mathbf{tr}[e^{i\lambda}(\mathcal{U}^{-1}d\mathcal{U})e^{-i\lambda}(\mathcal{U}^{-1}d\mathcal{U})]\right) . \quad (52)$$

The parenthesis is the WZW curving (whose integral over a closed 2d surface gives the WZW term), and the two lines correspond to two different gauge choices. Note that neither λ nor $\mathcal{U}$ is uniquely defined, since g is invariant under $\lambda \to \lambda + 2\pi\kappa$ for any $\mathbb{Z}$-valued diagonal matrix κ, and under $\mathcal{U} \to \mathcal{U}\mathcal{V}$ for any $\mathcal{V}$ that commutes with $e^{i\lambda}$, which means $\mathcal{V}$ must be diagonal unless g has eigenvalue degeneracy. [38] The WZW curving is in general not everywhere continuous, just like the Berry connection. If we cut the closed 3d space into many patches labeled by $\alpha, \beta \ldots$ that intersect along 2d common boundaries (this is known as a polyhedron decomposition of the space), at the 2d boundary between two patches α and β, the transformations above are allowed, constituting the transition functions $\kappa_{\alpha\beta}$ and $\mathcal{V}_{\alpha\beta}$ for the WZW curving. Substituting into the two gauge choices of the WZW curving above, we have respectively

$$\oint_{3d} \frac{i}{6(2\pi)^2}\mathbf{tr}[(g^{-1}dg)^3] = \sum_{\text{patches } \alpha<\beta} \int_{2d \text{ between } \alpha,\beta} \mathbf{tr}\left[\kappa_{\alpha\beta}\frac{i(\mathcal{U}_\beta^{-1}d\mathcal{U}_\beta)^2}{2\pi}\right]$$

$$= \sum_{\text{patches } \alpha<\beta} \int_{2d \text{ between } \alpha,\beta} \mathbf{tr}\left[\frac{d\lambda_\alpha}{2\pi}\frac{i\mathcal{V}_{\alpha\beta}^{-1}d\mathcal{V}_{\alpha\beta}}{2\pi}\right] . \quad (53)$$

---

[38] g is also invariant under $e^{i\lambda} \to \sigma^{-1}e^{i\lambda}\sigma$, $\mathcal{U} \to \mathcal{U}\sigma$ where $\sigma \in S_N$ permutes the eigenvalues (the Weyl group). This will not come up in the calculation here.

From either expression we can see the result is an integer: In the first gauge choice, recall κ is a diagonal integer matrix, so after projecting $i(\mathcal{U}^{-1}d\mathcal{U})^2$ to the diagonal elements by κ, the integrand is some linear sum of 2d Berry curvatures with integer coefficients, hence integrating to an integer; in the second gauge choice, each diagonal component of $d\lambda/2\pi$ picks up some winding number (recall λ will well-defined mod 2π) upon integration, and so does each diagonal component of $i\mathcal{V}^{-1}d\mathcal{V}/2\pi$, hence also leading to an integer. More explicitly, further using Stokes' theorem, either form above reduces to

$$\oint_{3d} \frac{i}{6(2\pi)^2}\mathbf{tr}[(g^{-1}dg)^3] = \sum_{\text{patches } \alpha<\beta<\gamma} \int_{1d \text{ between } \alpha,\beta,\gamma} \mathbf{tr}\left[\kappa_{\alpha\beta}\ \frac{i\mathcal{V}_{\beta\gamma}^{-1}d\mathcal{V}_{\beta\gamma}}{2\pi}\right] \tag{54}$$

$$= \sum_{\text{patches } \alpha<\beta<\gamma<\delta} \mathbf{tr}\left[\kappa_{\alpha\beta}\, n_{\beta\gamma\delta}\right]_{0d \text{ between } \alpha,\beta,\gamma,\delta} \in \mathbb{Z} \tag{55}$$

where $n_{\beta\gamma\delta} := i(\ln\mathcal{V}_{\beta\gamma}^{\text{diag}} - \ln\mathcal{V}_{\beta\delta}^{\text{diag}} + \mathcal{V}_{\gamma\delta}^{\text{diag}})/2\pi$ is an integer diagonal matrix once we fix the logarithm branch cut convention. [39] [40] In the same manner, we can also show that for a continuum Yang-Mills theory, the integral $\oint_{4d}\mathbf{tr}f^2/2(2\pi)^2$ gives an integer. The integrand is the exterior derivative of the CS 3-form, and at the 3d patch boundaries the CS 3-forms differ by the WZW curvature (plus some extra term, see e.g. [8]), and then the computation essentially reduces to that in the above.

Through this computation in the continuum, we can spot the appearance of some covering that is *not* a fibre bundle. The diagonalization of g that we performed in order to find a useful explicit presentation of the WZW curving corresponds to the Weyl map

$$T \times SU(N)/T \to SU(N) \tag{56}$$

where $T \cong (S^1)^{N-1}$ is the maximal torus parameterized by $e^{i\lambda}$, and $SU(N)/T$ is parameterized by $\mathcal{U}$ with the diagonal $\mathcal{V}$ action mod out. But the Weyl map is not a fibre bundle over $SU(N)$, because when two eigenvalues in $e^{i\lambda}$ happen to be degenerate, the space of $\mathcal{V}$ that commutes with $e^{i\lambda}$ is enlarged to include non-diagonal matrices, but only the space of diagonal ones is being mod out. Thus the Weyl map violates the local triviality condition for a fibre bundle. When we express the WZW curving (52) by λ and $\mathcal{U}$, we are further extending the covering into $\mathbb{R}^{N-1} \times SU(N) \to T \times SU(N)/T \to SU(N)$; since the diagonalization Weyl map is already not a fibre bundle, nor is it after this further extension.

---

[39]In this derivation we have been consecutively using Stokes' theorem to reduce quantities onto the intersections between more and more patches. Mathematically, such a structure is known as a Deligne-Beilinson double cochain in the context of Deligne-Beilinson double cohomology (see e.g. [36]), where one direction of the cohomology is the (de Rham) d, and the other direction is the (Čech) transition between patches. For $U(1)$ gauge theory such a description is presented in details in e.g. [60]. Here, instead of $U(1)$ gauge connection 1-form, in $|SU(N)|$ nlσm we have $U(1)$-valued WZW curving 2-form, and later in $SU(N)$ Yang-Mills theory we have $U(1)$-valued non-abelian CS 3-form, but the idea is similar.

One might also note the resemblance between Deligne-Beilinson double cohomology and BRST double cohomology (in particular, the structure we have described resembles the BRST descent equations). Their correspondence is via the notion of ananatural isomorphism to be introduced in Section 5.2.

[40]Just like the Berry connection is widely used in studying the topological and geometrical effects in the momentum space / Brillouin zone (as opposed to the real space), the WZW curving is also useful—though less well-known—in the same context, and is especially necessary when the system is interacting [88].

While diagonalizing $SU(N)$ does not give rise to a fibre bundle, diagonalization is familiar enough to make sense of and work with. This is indeed how we will construct our non-fibre-bundle finite dimensional structure to solve our problem on the lattice. [41] We will first present the construction for S^3 lattice nlσm, which can be generalized to $|SU(N)|$ nlσm. Next, similar to how we went from S^1 nlσm in Section 2.1 to $U(1)$ gauge theory in Section 2.2, roughly speaking "putting everything in one higher dimension" will lead to the construction for $SU(N)$ lattice Yang-Mills. How to actually carry out this step is not as obvious as in the $U(1)$ case, and interestingly, if we carry out this step in the "literal" way, a troublesome issue that requires solving some generalized version of Yang-Baxter equation will come up. [42] Fortunately, if we carry out this step in a way that better appeals to the relation with the continuum theory [10]—which will involve some techniques similar to the traditional work [8] but under a shifted mindset—then the generalized Yang-Baxter equation issue will be automatically resolved. In fact, our construction recovers [8] if we assume the gauge field strength is weak (which [8] requires) and take the saddle point approximation.

4.1 S^3 non-linear sigma model: Wess-Zumino-Witten, skyrmion and hedgehog

In the traditional S^3 nlσm, the dynamical S^3 d.o.f. at each vertex is parametrized by $g_v \in SU(2) \cong S^3$. Across each link there is a link weight $W(\mathbf{tr}Dg_l + c.c.)$ where $Dg_{l=\langle v'v\rangle} := g_{v'}g_v^{-1} \in SU(2)$, and W is a positive, increasing function. Note that $\mathbf{tr}Dg_l$ is indeed invariant under the $SO(4) \cong SU(2)_L \times SU(2)_R/\mathbb{Z}_2$ global symmetry (47) as $(Dh_L)_l = \mathbf{1} = (Dh_R)_l$. Now we want to topologically refine the traditional theory, so that we can naturally define the topological operators such as WZW term and skyrmion. The result is (65). We will now step by step introduce the d.o.f. and the desired properties of the path integral weights, i.e. "the rules of the game", using geometrical intuitions, leaving the formal mathematics to Sections 5 and 6. [43]

The hint from continuum, which we discussed above, suggests that we should perform diagonalization in order to find some useful cover over $SU(2)$. In the continuum, e.g. in getting (52), we diagonalized $g(x)$ itself. On the lattice, it turns out more natural to diagonalize $Dg_l \in SU(2)$ instead of $g_v \in S^3$. It is desired that whatever we do should manifest the $SO(4) \cong SU(2)_L \times SU(2)_R/\mathbb{Z}_2$ global symmetry (47). Under this transformation, $Dg_l \to h_L Dg_l h_L^{-1}$, the eigenvalues remain unchanged. This strongly suggests it is good to consider the diagonalization of Dg_l rather than the diagonalization of g_v; in particular,

---

[41] As we cut the continuum spacetime into fine enough patches, the resulting Čech nerve can be viewed as a lattice, where each patch is seen as a lattice vertex. So the fact that diagonalization is useful on continuum patches indeed suggests that it is useful on the lattice as well.

[42] It might seem surprising that some kind of Yang-Baxter equation is involved in Yang-Mills theory. In Sections 5.3 and 5.5 we will explain why the appearance of Yang-Baxter equation is completely natural when we pass from a nlσm to a gauge theory.

[43] In mathematical terms, what we will describe is a concrete construction of a bundle gerbe over $SU(N)$, following [89] but with some necessary technical modifications (see footnote 47), then turned multiplicative [36] using some geometrical intuition which gives a concrete implementation of the procedure in [90] yet again with some necessary extra technical modifications (see footnote 49). See Section 5.5 for the formal discussions.

diagonalizing g_v and $g_{v'}$ does not lead to a diagonalization of $Dg_l = g_{v'}g_v^{-1}$. [44]

On each lattice link l, we will construct a non-fibre-bundle cover over $SU(2) \ni Dg_l$. First, let us consider covering $SU(2)$ by two open patches, $SU(2)\backslash\{-\mathbf{1}\}$ and $SU(2)\backslash\{+\mathbf{1}\}$, although this is slightly different from what we will use in the end. The disjoint union of the two patches, $SU(2)\backslash\{-\mathbf{1}\} \sqcup SU(2)\backslash\{+\mathbf{1}\}$, which covers $SU(2)$, is indeed not a fibre bundle over $SU(2)$ since $\pm\mathbf{1}$ are special points. To understand why we choose patches in such way, let us diagonalize $Dg_l =: \mathcal{U}_l e^{i\lambda_l \sigma^z} \mathcal{U}_l^{-1}$. (Note the difference with the convention in (52): there we are diagonalizating g while here Dg_l, moreover there λ is a diagonal matrix while here a number, the coefficient of σ^z.) The first patch contains those Dg_l elements such that $\lambda_l \in [0, \pi)$, while the second patch contains those Dg_l elements such that $\lambda_l \in (0, \pi]$. [45] Thus:

- The patches are defined using only the eigenvalues of Dg_l, ensuring the patches to remain invariant under the $SO(4) \cong SU(2)_L \times SU(2)_R/\mathbb{Z}_2$ transformation (47) which transforms Dg_l by conjugation. In other words, if a patch contains some group element, the patch must contain the entire conjugacy class of that group element.

- The special points $\pm\mathbf{1}$ are where the eigenvalues of Dg_l become degenerate. These are indeed special loci in the diagonalization, because the ambiguity $\mathcal{U}_l \to \mathcal{U}_l \mathcal{V}_l$ enhances from $U(1)$ to $SU(2)$ at these loci. It is anticipated that these special loci require special treatments in what we will do later.

In our actual construction, the link d.o.f. will take value in a non-fibre-bundle cover over $SU(2) \ni Dg_l$ given by

$$Y := \left(SU(2)\backslash\{-\mathbf{1}\}\right) \sqcup \left(SU(2)\backslash\{+\mathbf{1}\} \times S^2\right) \tag{57}$$

and what this extra S^2 does on top of the second patch will be explained at (59). Let us denote an element $y_l \in Y$ by $y_l = (Dg_l, m_l, \hat{n}_l)$, where $m_l = +$ means y_l belongs to the $SU(2)\backslash\{-\mathbf{1}\}$ patch (so $m_l = +$ implies $Dg_l \neq -\mathbf{1}$), $m_l = -$ means y_l belongs to the $SU(2)\backslash\{+\mathbf{1}\}$ patch (so $m_l = -$ implies $Dg_l \neq +\mathbf{1}$), and $\hat{n}_l \in S^2$ is only going to be meaningful when $m_l = -$ (i.e. when $m_l = +$, $\hat{n}_l$ will not appear anywhere in the theory and can be ignored).

Note that while m_l is a two-valued label, it by no means forms a $\mathbb{Z}_2$ group, as there is no sensible group composition; nor is there is $\mathbb{Z}_2$ symmetry acting on m_l. And of course, the whole space Y itself is not a group and cannot compose, either. We will see why this is not a problem.

---

[44]That fact that $g_v \in S^3$ is not naturally a group element while $Dg_l \in SU(2)$ is naturally a group element already suggests it is better to diagonalize Dg_l. In the continuum, there is no nice counterpart of Dg_l, since $g^{-1}dg$ is only a Lie algebra element, not a Lie group element.

In a very recent work [47], while the d.o.f. are still the traditional vertex variables g_v, bundle gerbe techniques have nonetheless been employed to compute the lattice skyrmion number. (By constrast, in our work, we have new d.o.f., which, together with the traditional g_v, form a bundle gerbe type structure.) There, the diagonalization is indeed performed on g_v rather Dg_l.

[45]$\lambda \in (-\pi, 0)$ is equivalent to $\lambda \in (0, \pi)$ upon exchanging the two eigenvalues.

In the lattice path integral, we will replace the traditional link weight $W(\mathbf{tr}Dg_l + c.c.)$ (note $\mathbf{tr}Dg_l + c.c. = 4\cos\lambda_l$) by some link weight $W_1(\lambda_l, m_l)$ over Y, with $m_l = \pm$ summed over (pretending there are no other weights that depend on m_l for now):

$$W(\mathbf{tr}Dg_l + c.c.) \approx \sum_{m_l=\pm} W_1(\lambda_l, m_l) . \tag{58}$$
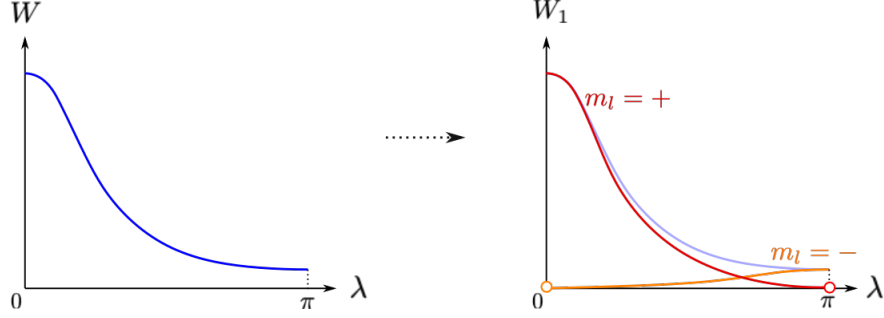
(At this point there is no dependence on $\hat{n}_l$, so $\int d^2\hat{n}_l/4\pi$ yields a trivial factor 1.) This is similar in idea to (7) and (41), but now there is a new aspect that should be emphasized: Each patch of Y does not cover the entire $SU(2) \ni Dg_l$, and we require W_1 to smoothly vanish towards the boundary of the image of each patch of Y, indicated by the hollow circles above, so to ensure the smoothness in λ_l after summing over m_l.

We shall develop some physical intuition of what $y_l = (Dg_l, m_l, \hat{n}_l)$ means. If we think of the lattice link as a path embedded in the continuum, then along it the continuum field $g(x)$ traces out a path in S^3 interpolating from g_v to $g_{v'}$. The infinite dimensional details of how the path wiggles are unimportant, while the useful homotopy information is to be kept in y_l. Clearly the $Dg_l = g_{v'}g_v^{-1}$ part of y_l indicates the relative position of the starting and the ending point. As long as $Dg_l \neq -\mathbf{1}$, there is a unique shortest geodesic from g_v to $g_{v'}$, given by $\{\mathcal{U}_l e^{i\lambda'\sigma^z}\mathcal{U}_l^{-1}g_v | 0 \leq \lambda' \leq \lambda_l\}$. Then $m_l = +$ represents the contributions from all those continuum paths that are "close enough" to the geodesic. On the other hand $m_l = -$ represents the contributions from all other continuum paths. Schematically:
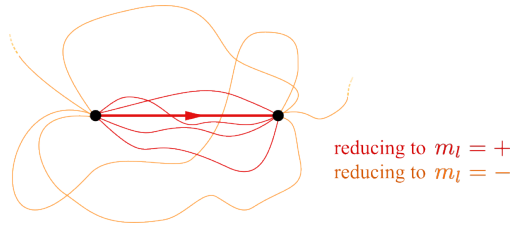
How to define "close enough" in detail is also unimportant, but when $Dg_l \to -\mathbf{1}$, fewer and fewer paths are considered "close enough" till none is (indeed, when $Dg_l = -\mathbf{1}$ there is no unique shortest geodesic), and when $Dg_l \to +\mathbf{1}$, more and more paths are considered "close enough" till all paths are. This explains the qualitative behavior of $W_1(\lambda_l, m_l)$ illustrated in (58).

It is helpful for both intuitive and practical purposes to pick a representative path for a given $y_l \in Y$; in particular we will use the representative to construct the μ function in the

44

plaquette weight (60) later. Clearly, for the $m_l = +$ patch, the most natural choice of the representative path for $y_l = (Dg_l, +)$ is the shortest geodesic, $\{\mathcal{U}_l e^{i\lambda'\sigma^z}\mathcal{U}_l^{-1}g_v|0 \leq \lambda' \leq \lambda_l\}$. On the other hand, a good choice of the representative path for the $m_l = -$ patch is less obvious, and that is why we will need the $\hat{n}_l \in S^2$ d.o.f.: For $y_l = (Dg_l, -, \hat{n}_l)$, the choice of representative path is to first go from g_v to $-g_v$ via $\{e^{i\lambda''\hat{n}_l\cdot\vec{\sigma}}g_v|0 \leq \lambda'' \leq \pi\}$, and then go from $-g_v$ to $g_{v'}$ via the shortest geodesic $\{\mathcal{U}_l e^{i\lambda'\sigma^z}\mathcal{U}_l^{-1}g_v|\pi \geq \lambda' \geq \lambda_l\}$. [46] We illustrate the representative paths (one with $m_l = +$, one with $m_l = -$ and some choice of $\hat{n}_l$) by picturing $SU(2) \ni Dg_l$ as a 3d ball centered at $\mathbf{1}$ and with radial coordinate λ, so that whole the surface at $\lambda = \pi$ is identified to a single point $-\mathbf{1}$:

$$m_l = + \qquad\qquad m_l = - \qquad\qquad\qquad\qquad (59)$$
$$(g_{v'}g_v^{-1} \neq -\mathbf{1}) \qquad\qquad (g_{v'}g_v^{-1} \neq +\mathbf{1})$$

From this interpretation of $\hat{n}_l$, we can see that under the $SO(4)$ global symmetry transformation (47), not only does $\mathcal{U}_l \to h_L\mathcal{U}_l$, but also $\hat{n}_l \to R_{h_L}\hat{n}_l$, i.e. $\hat{n}_l \cdot \vec{\sigma} \to h_L(\hat{n}_l \cdot \vec{\sigma})h_L^{-1}$, so that the representative path transforms covariantly. [47]

After introducing the link variable $y_l \in Y$ and the link weight W_1, we now move on to the plaquette. In the known examples in Section 2, the link variables always form a group, whose group composition is useful on the plaquette. Now we want to emphasize it is not necessary for the link variables to be composable—indeed, for our construction now Y is not composable (even if we have chosen some representative paths, the space of these paths is not closed under concatenation). We want to show the plaquette variable can still be well-defined as long as we have specified the link variables around, without being able to compose them. From the discussions in Section 3, it is clear that the new d.o.f. on the plaquette should be $U(1)$-valued, effectively representing the WZW curving over the plaquette.

In the continuum, the WZW curving (52) is in general not continuously defined everywhere (just like the Berry connection), and correspondingly, on the lattice, the WZW curving

---

[46]Upon reversing the orientation of the link (i.e. exchanging g_v and $g_{v'}$), the representative path for $m_l = +$ only reverses the running direction but the trajectory of the path remains the same. On the other hand, the trajectory of the representative path for $m_l = -$ changes. This is not a major issue: we can either fix some ordering of the vertices and hence the orientations of the links, or introduce an extra two-valued random variable to decide which orientation of the link is to be used when choosing the representative path.

[47]Our introduction of $\hat{n}_l \in S^2$ is the technical deviation from the bundle gerbe introduced in [89], which used the simpler $Y = (SU(2)\backslash\{-\mathbf{1}\}) \sqcup (SU(2)\backslash\{+\mathbf{1}\})$. This is because relating an element in Y to a continuum path is not an issue under consideration in [89]. This issue is under consideration (though not very transparently) in the later work [90], but there it suffices to fix, say, $\hat{n}_l = \hat{x}$. Here, we do not want to fix any choice that breaks the $SO(4)$ global symmetry, so we let $\hat{n}_l$ take all possible directions. (Whether this extra $S^2 \ni \hat{n}_l$ can be viewed as some reminiscent of the more general bundle gerbe construction introduced in [91, 92] should be further investigated.)

$U(1)$ d.o.f. on the plaquette forms a non-trivial $U(1)$ bundle over the space of the vertex and link variables around the plaquette (just like in Section 2.4, the Berry connection $U(1)$ d.o.f. on the link forms a non-trivial $U(1)$ bundle over the space of the vertex d.o.f. at the ends of the link). We will first have an abstract description of the topology of this non-trivial $U(1)$ bundle, which might appear intractable; then we will describe the properties that the plaquette weight W_2 (60) should have, in order to realize such non-trivial $U(1)$ bundle in a manageable and intuitive manner, and to prescribe the desired dynamical properties.

Consider a plaquette with vertex and link variables labelled as the following.
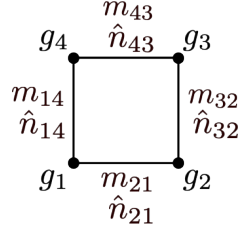
The WZW curving $U(1)$ d.o.f. takes value in a $U(1)$ bundle over the space of $(y_{21}, y_{32}, y_{43}, y_{14})$ (note this space is not Y^4, because the Dg_l parts are not independent, as they have to satisfy $DDg_p = \mathbf{1}$). Since the m_l's are discrete, this can be equivalently stated as that, for each given combination of the m_l's, the WZW curving d.o.f. takes value in a $U(1)$ bundle over the space of the allowed g_v's (and $\hat{n}_l$'s, if some $m_l = -$).

- First, consider the case where all four m_l's take $+$ (so that the $\hat{n}_l$'s are ignored). Then the allowed g_v's do not form $(S^3)^4$, but a space with $\pi_2 \cong \mathbb{Z}$ formed by carving out some parts from $(S^3)^4$: First, g_1 is chosen freely from S^3 and nothing should depend on this first choice due to the $SO(4)$ global symmetry—if we want we can freely set g_1 to $\mathbf{1}$ using the symmetry. Next, g_2 is chosen from $S^3\backslash\{-g_1\} \cong D_3$ since $m_{21} = +$. Likewise g_3 is chosen from $S^3\backslash\{-g_2\} \cong D_3$ since $m_{32} = +$. Finally, and most non-trivially, g_4 is to be chose from $S^3\backslash\{-g_1, -g_3\}$ since $m_{43} = + = m_{14}$. Generically $g_3 \neq g_1$, and in such generic case the space $S^3\backslash\{-g_1, -g_3\} \ni g_4$ is homotopic to S^2, which has $\pi_2 \cong \mathbb{Z}$, and supports a non-trivial $U(1)$ bundle homotopic to $U(1) \to S^3 \to S^2$ for the WZW curving (note the S^3 here is *not* the original S^3 in which g lives). Actually, the presence of the $g_3 = g_1$ spot will not alter the fact that the space formed by the allowed g_v's has $\pi_2 \cong \mathbb{Z}$, and the $U(1)$ bundle can be extended to this spot. [48]

- Now, say we change from $m_{14} = +$ to $m_{14} = -$, keeping other $m_l = +$. Then the carved-out part has changed, because the previous condition $g_4 \in S^3\backslash\{-g_1, -g_3\}$ now becomes $g_4 \in S^3\backslash\{+g_1, -g_3\}$. This space of allowed g_4, though changed, again has $\pi_2 \cong \mathbb{Z}$. Moreover, there is now the $\hat{n}_{14} \in S^2$ d.o.f. which also has $\pi_2 \cong \mathbb{Z}$. The WZW curving $U(1)$ bundle is determined by the following conditions (the reasoning

---

[48]The space for g_3 where $g_3 \neq g_1$ is homotopic to S^2, because $g_3 \neq -g_2$ due to $m_{32} = +$, and $-g_2 \neq g_1$ due to $m_{21} = +$. The space for (g_3, g_4) under the additional assumption that $g_3 \neq g_1$ is thus homotopic to $S^2 \times S^2$. Now we include the spot $g_3 = g_1$. At this spot, the space for g_4 is homotopic to a point. Thus, the total space for (g_3, g_4) is homotopic to such a space: start with $S^2 \times S^2$ (which represents $g_3 \neq g_1$), drag/collapse $S^2 \times \{\text{north pole}\} \subset S^2 \times S^2$ to a single point (which represents $g_3 = g_1$).

behind these conditions will become intuitive when we explain the plaquette weight (60)): Generically $g_3 \neq -g_1$, and in such generic case, for any fixed $\hat{n}_{14}$, the space $S^3 \backslash \{+g_1, -g_3\} \ni g_4$ is homotopic to S^2, over which the WZW curving $U(1)$ d.o.f. forms a non-trivial $U(1)$ bundle homotopic to $U(1) \to S^3 \to S^2$; on the other hand, for any fixed g_4, we have the space $S^2 \ni \hat{n}_{14}$, over which the WZW curving $U(1)$ d.o.f. also forms a non-trivial $U(1)$ bundle $U(1) \to S^3 \to S^2$. The bundle can be extended to the $g_3 = -g_1$ spot.

- Similarly for other combinations of the m_l's.

(Apparently, the same idea applies when the plaquette is not a square but has other numbers of links around.)

It seems these $U(1)$ bundles over such bizarre base spaces are extremely difficult to parametrize, let alone to prescribe a reasonable weight W_2 over them. But the intuitive relation to the continuum makes the task much more manageable than it might seem. It is useful to borrow the ideas from the discussions below (41) and (42). We denote the WZW curving d.o.f. by $e^{i\mathcal{W}_p} \in U(1)$. We let the plaquette weight take the form

$$W_2 \big(e^{i\mathcal{W}_p} \mu^*_{g_v \in \partial p, m_{l \in \partial p}, \hat{n}_{l \in \partial p}} + c.c. \big) \tag{60}$$

where W_2 is positive and increasing. Here the function μ^* plays the role of $u^\dagger_{\hat{n}_{v'}} u_{\hat{n}_v}$ in (42). Similar to the discussion there, we can *simultaneously* take care of the topology of the $U(1)$ bundle and the desired dynamical properties of the weight by requiring suitable properties for the complex function μ. The WZW curving maximizes the weight W_2 when $e^{i\mathcal{W}_p} = \mu/|\mu|$, which is well-defined when $|\mu| \neq 0$. On the other hand, as $|\mu|$ approaches 0, the weight W_2 becomes less and less sensitive to the value of the WZW curving $e^{i\mathcal{W}_p}$. By thinking the plaquette as being embedded in the continuum, it is easy to picture the following desired properties for μ:

- The phase $\mu/|\mu|$ is given by the continuum WZW curvature integrated over such a pyramid: The four base corners are at the g_v's and the tip is at $\mathbf{1}$; the neighboring base corners are connected to each other by the aforementioned representative paths (59) for the given m_l's and $\hat{n}_l$'s, while the tip is connected to each base corner by the shortest geodesic; the four triangles on the side and the quadrangle at the base are then wrapped with some standard choice of interpolating surfaces (discussed below), forming a pyramid. Which side is called the "inside" of the pyramid does not matter, since the two choices only differ by a 2π phase. An illustration of the pyramid in

$S^3 \ni g_v$, assuming each $m_l = +$, looks like

$$\text{when each } m_l = +$$
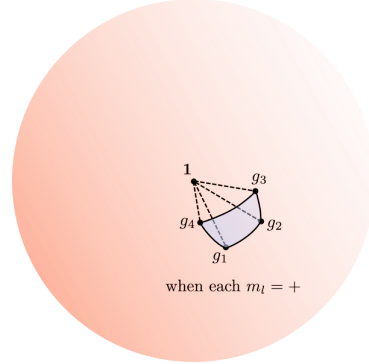
(61)

[49] Similar to the case of Berry connection in (43), here only the base quadrangle surface is physical, while the tip and the four triangles on the side are just some gauge choice; a gauge change can be absorbed by a redefinition of $e^{i\mathcal{W}_p}$. When six plaquettes piece up to a cube, we can compute the lattice WZW curvature over the cube, $e^{id\mathcal{W}_c}$, in which the dependence on the gauge choices (the tip and the triangles on the sides) cancel out.

Fluctuations of $e^{i\mathcal{W}_p}$ away from $\mu/|\mu|$ can be thought of as capturing the fluctuation of the interpolating surface at the base quadrangle away from the standard choice (recall the equivalence relation explained below (51)), just like the fluctuation of the Berry connection in the case of S^2 nlσm (recall (42), (43) and (44)).

- How to choose the standard interpolating surfaces in detail is not so important as long as the choice is "reasonable", approaching some notion of minimal surface when the g_v's are close to each other and all $m_l = +$. For concreteness we will discuss one choice for the interpolating surface later. For now we explain the general idea of how topology is taken into account. It is easy to see that the choice of interpolating surface cannot be made continuously everywhere over the space of variables $g_{v\in\partial p}, m_{l\in\partial p}, \hat{n}_{l\in\partial p}$ in μ, and singularities will be developed. Topology is taken into account by treating the singularities appropriately:

  - When the choice of the interpolating surface for any of the side triangles of the pyramid becomes singular, the phase $\mu/|\mu|$ also becomes singular. But this does not matter because the side triangles are gauge choices anyways, just like the singularity in the case of Berry connection (when either of $\hat{n}_v$ and $\hat{n}_{v'}$ approaches $-\hat{z}$ in (42)). Such singularity is unavoidable, indeed because we want the WZW

---

[49]Our use of geometrical concatenation of the representative paths and surfaces is a technical deviation from [90], which uses pointwise multiplication and Mickelsson product. This is because strictly speaking, what we are constructing as the target category is not a multiplicative bundle gerbe—only if we ignore $g_v \in S^3$ but keep $Dg_l \in SU(2)$ subjected to $DDg_p = \mathbf{1}$ will the structure reduce to a multiplicative bundle gerbe (see Section 5.5 for details, in particular (123) versus (128)). Therefore, for us it is natural to consider paths in S^3 with arbitrary starting points, and natural to consider their concatenation. On the other hand, [90] considered paths in $SU(2)$ with the starting point fixed at the origin, so the Mickelsson product is used.

curving to take value in a non-trivial $U(1)$ bundle over the space of the vertex and link variables.

- On the other hand, when the choice of the interpolating surface for the base quadrangle becomes singular, we require $|\mu| \to 0$, so that W_2 becomes insensitive to the value of the WZW curving e^{iW_p}, and this agrees with our intuition, just like in the case of Berry connection (when $\hat{n}_v = -\hat{n}_{v'}$ in (42)). More generally, we want $|\mu| = 1$ when all $m_l = +$ and all g_v equal, and $|\mu|$ decreases as the base quadrangle loop becomes larger and larger, until $|\mu| = 0$ when the choice of interpolating surface becomes singular.

- For concreteness we will discuss two reasonable choices for the standard interpolating surfaces. The choices are of course non-unique, and what choice is the best for numerical purpose can only be determined via future numerical investigations.

Choice 1: First we further cut the base quadrangle into two triangles by connecting g_2 and g_4 via the shortest geodesic (we will see that the special case of $g_2 = -g_4$ will naturally have $|\mu| = 0$), so that we have six triangles, four on the sides of the pyramid, and two on the base.

- First consider a triangular loop such that all three edges are given by the shortest geodesics, that is, when all m_l involved in this triangle takes $+$. Just like two points in S^2 determine a great circle as long as the two points are not opposite, three points in S^3 (the three vertices of the triangle) determine a great sphere as long as the three points are not on a same geodesic circle. [50] The great sphere is cut into two pieces by the edges of the triangular loop, and one piece is always smaller than the other given that the three points are not on a geodesic circle, and we pick the smaller piece to be the interpolating surface.
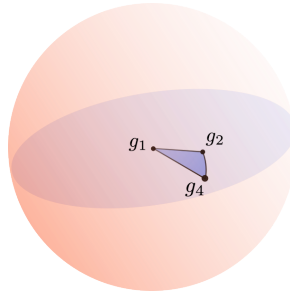
Here we drew one of the triangles at the base of the pyramid, and we placed g_1 at the origin to make it easier to illustrate what a great sphere means.

---

[50]To see this, denote the three points by $p_1, p_2, p_3 \in SU(2)$. The below would be most easily pictured by setting $p_1 = \mathbf{1}$ though we will keep it general. Two points p_1, p_2 determine a geodesic circle ℓ_{21} (which is generated by diagonalizing Dp_{21}, letting the eigenvalue take any value between 0 and 2π, and then multiplying the matrix back on p_1). Similarly p_1, p_3 determine a geodesic circle ℓ_{31}, which is distinct from ℓ_{21} assuming the three points are not on a same geodesic circle. Now ℓ_{21} can be rotated to ℓ_{31} by an $SO(2)$ rotation living in the $SO(3) \subset SO(4)$ that keeps p_1 unchanged. Letting this $SO(2)$ rotation take angles from 0 to 2π generates the desired great sphere.

In the special cases where the three points lie on a same geodesic circle, either the triangular loop is degenerate (i.e. one point lies on the shortest geodesic between the other two points) or the triangular loop itself is the geodesic circle. Obviously, for the former kind, we will take the interpolating surface to be trivial. On the other hand, for the latter kind, the choice of the interpolating surface will become singular, and this is the topological issues we discussed before—if the triangle loop is a side triangle of the pyramid, it is fine that the interpolating surface becomes singular since it is merely a gauge choice; while if the triangular loop is on the base, we will let $|\mu| = 0$.

– Next consider a triangular loop such that one edge is flipped from $m_l = +$ to $m_l = -$. It seems it is a consistent, though perhaps crude, approximation to just set $|\mu| = 0$ whenever any $m_{l \in p} = -$. In that case, the description of the interpolating surface below will not be needed, and the $\hat{n}_l \in S^2$ variable can be ignored, so that the theory will be simplified. Whether this crude approximation is good enough to describe the physics of the nlσm is subjected to numerical investigation. For now we suppose we do not simply set $|\mu| = 0$ when some $m_{l \in p} = -$.

Since the representative path for $m_l = -$ is in general not a geodesic but two segments of geodesics, such "triangular loop" really looks like a quadrangular loop. The choice of the interpolating surface is illustrated as
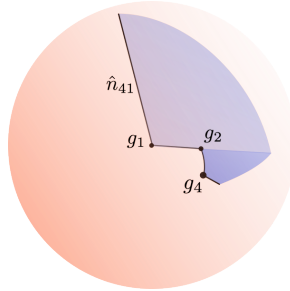
which is the union of two interpolating surfaces: one for the triangular formed by connecting $g_1, g_2, e^{i(\pi-0^+)\hat{n}_{14}\cdot\vec{\sigma}}g_1$ with the shortest geodesics, and another for the triangular loop formed by connecting $g_2, g_4, -g_1$ with the shortest geodesics. The idea is that, when $m_{14} = -$ and $g_4 \to -g_1$, the interpolating surface would approach that of when $m_{14} = +$ and $g_4 \to e^{i(\pi-0^+)\hat{n}_{14}\cdot\vec{\sigma}}g_1 \to -g_1$ from the $\hat{n}_{14}$ direction.

The treatments when the choice of interpolating surface becomes singular is the same as before.

When more $m_l = -$, the idea is the same.

Choice 2: In the subsequent work [10], an interpolation procedure is introduced for gauge theory (which we will discuss in the next subsection). One can use the similar idea to construct the interpolation surfaces in nlσm.

This describes the crucial topological and dynamical properties of the plaquette weight W_2 that probabilistically weighs the WZW curving d.o.f. $e^{i\mathcal{W}_p}$.

In 2d, we can readily define the WZW phase at level $k \in \mathbb{Z}$ as

$$W_{WZW}^k := e^{ik \oint_{2d} \mathcal{W}} := e^{ik \sum_p \mathcal{W}_p} \ . \tag{62}$$

Notably, a $k \neq 0$ WZW phase makes the $SO(4)$ global symmetry anomalous, although it does not directly break the symmetry. That is, if a non-trivial $SO(4)$ background gauge field is introduced, the definition of the WZW phase will become ambiguous. This can be seen on the lattice explicitly. However, to explain the details, we need to discuss how a topologically refined nlσm is coupled to a topologically refined non-abelian gauge field, and we will leave the detailed discussion to future works. [51]

Beyond 2d, the last step, of course, is to Villainize the lattice WZW curvature $e^{id\mathcal{W}_c} \in U(1)$ to the skyrmion density

$$\mathcal{S}_c := d\mathcal{W}_c/2\pi + s_c \in \mathbb{R} \tag{63}$$

by introducing an $s_c \in \mathbb{Z}$ dynamical variable on each cube. We have a cube weight $W_3(\mathcal{S}_c)$ that is positive and decreases with $|\mathcal{S}_c|$. The total skyrmion number over a 3d surface is then defined as $\oint_{3d} \mathcal{S} = \sum_c s_c \in \mathbb{Z}$. A topological theta term can hence be defined. In 4d or above, we can define the hedgehog like defect $d\mathcal{S}_h = ds_h$ (where h labels hypercubes), which represents the non-conservation of baryon number in the context of pion vacua effective theory in 4d spacetime. Again we can introduce a fugacity weight $W_4(d\mathcal{S}_h)$ for these defects, or forbid them using

$$W_4^{forbid}(d\mathcal{S}) = \int_{-\pi}^{\pi} \frac{d\widetilde{\phi}_h}{2\pi} \ e^{i\widetilde{\phi}_h d\mathcal{S}_h} \ . \tag{64}$$

[52] If we indeed use W_4^{forbid}, then there is the $(d-4)$-form dual $U(1)$ global symmetry $e^{i\widetilde{\phi}_h} \to e^{i\widetilde{\phi}_h} e^{i\widetilde{\alpha}_h}$, $e^{d^* \widetilde{\alpha}_c} = 1$, which in $d = 4$ is interpreted as the baryon conservation $U(1)$. Again there is a mixed anomaly between the original $SO(4)$ global symmetry and this dual $U(1)$ global symmetry. Just like the anomaly mentioned below (62), we will leave the detailed discussion of this anomaly to future works. [53]

---

[51]Briefly speaking, the main task is to generalize the definition of the μ function to situations where the global symmetry background is non-trivial, and this is done using some technique to be introduced in Section 4.2, in relation to non-abelian CS phase factor. After doing so, we will find that under local gauge transformation of the background gauge field, the phase of this generalized μ function will transform. In W_2, we can absorb this phase transformation of μ into $e^{i\mathcal{W}_p}$, but then the WZW phase factor would not remain invariant unless $k = 0$.

[52]Very recently, [93] also discussed defining and forbidding defects in lattice nlσm beyond the previously known examples (Villain and spinon-decomposition), by discretizing the target space (e.g. the S^3 here). Here what we showed is that the same can be done without discretizing the target space—the vertex d.o.f. still takes value in S^3 itself rather than some discrete points on S^3, and the $SO(4)$ global symmetry is still manifest. See footnote 143 for more discussions.

[53]One picture to describe the mixed anomaly is that the instanton of the $SO(4)$ background gauge field is charged under the dual $U(1)$. As we sketched in footnote 51, under gauge transformation of the $SO(4)$ background, the local WZW curving variable $e^{i\mathcal{W}_p}$ must transform accordingly to keep W_2 invariant. Then the remaining situation essentially becomes that of the gauge transformation of a Villainized 2-form $U(1)$

Piecing up the discussions above, we have our first main result: The lattice S^3 nlσm refined to include skyrmion reads

$$Z = \left[\prod_{v'} \int_{SU(2)} dg_{v'} \right] \left[\prod_{l'} \sum_{m_{l'}=\pm} \int \frac{d^2 \hat{n}_{l'}}{4\pi} \right] \left[\prod_{p'} \int_{-\pi}^{\pi} \frac{d\mathcal{W}_{p'}}{2\pi} \right] \left[\prod_{c'} \sum_{s_{c'} \in \mathbb{Z}} \right]$$
$$\prod_l W_1(\lambda_l, m_l) \prod_p W_2(e^{i\mathcal{W}_p} \mu^*_{g_{v \in \partial p}, m_{l \in \partial p}, \hat{n}_{l \in \partial p}} + c.c.) \prod_c W_3(\mathcal{S}_c) \prod_h W_4(d\mathcal{S}_h) \qquad (65)$$

for $d \geq 4$. The d.o.f. together form a mathematical structure that counts the π_3 of the lattice S^3 nlσm, to be explained with (128). [54] For $d = 3$, there is no W_4, but we can additionally consider a topological theta term

$$Z_\Theta = \left[\prod_{v'} \int_{SU(2)} dg_{v'} \right] \left[\prod_{l'} \sum_{m_{l'}=\pm} \int \frac{d^2 \hat{n}_{l'}}{4\pi} \right] \left[\prod_{p'} \int_{-\pi}^{\pi} \frac{d\mathcal{W}_{p'}}{2\pi} \right] \left[\prod_{c'} \sum_{s_{c'} \in \mathbb{Z}} \right]$$
$$e^{i\Theta \sum_c \mathcal{S}_c} \prod_l W_1(\lambda_l, m_l) \prod_p W_2(e^{i\mathcal{W}_p} \mu^*_{g_{v \in \partial p}, m_{l \in \partial p}, \hat{n}_{l \in \partial p}} + c.c.) \prod_c W_3(\mathcal{S}_c) \qquad (66)$$

for any $\Theta \in U(1)$. For $d = 2$, there is no s_c and W_3, but we can additionally consider a WZW term

$$Z_{kWZW} = \left[\prod_{v'} \int_{SU(2)} dg_{v'} \right] \left[\prod_{l'} \sum_{m_{l'}=\pm} \int \frac{d^2 \hat{n}_{l'}}{4\pi} \right] \left[\prod_{p'} \int_{-\pi}^{\pi} \frac{d\mathcal{W}_{p'}}{2\pi} \right]$$
$$e^{ik \sum_p \mathcal{W}_p} \prod_l W_1(\lambda_l, m_l) \prod_p W_2(e^{i\mathcal{W}_p} \mu^*_{g_{v \in \partial p}, m_{l \in \partial p}, \hat{n}_{l \in \partial p}} + c.c.) \qquad (67)$$

for any $k \in \mathbb{Z}$.

---

gauge field, constituting of $e^{i\mathcal{W}_p} \in U(1)$ and $s_c \in \mathbb{Z}$. A Villainized 3-form $U(1)$ background will be introduced as the refinement of the $SO(4)$ global symmetry background (similar to Section 4.2, but here the fields are not dynamical). This consists of $e^{iC_c} \in U(1)$, interpreted as the CS d.o.f. of the lattice $SO(4)$ background gauge field, and $I_h \in \mathbb{Z}$, such that $dC_h/2\pi + I_h$ is the background instanton density. In W_3, $d\mathcal{W}_c/2\pi + s_c \to d\mathcal{W}_c/2\pi + s_c - C_c/2\pi$ where C_c absorbs the aforementioned 2-form $U(1)$ gauge transformation of $d\mathcal{W}_c$, and in W_4, $ds_c \to ds_h - dC_h/2\pi - I_h$ which is no long exact, hence violating the dual $U(1)$ global symmetry.

An alternative picture is, if we introduce a background gauge field for the dual $U(1)$ global symmetry, it is easy to see the Dirac string part of this $U(1)$ background will couple to $e^{i\mathcal{W}_p}$, generating a WZW phase whose level is the Dirac string charge. (This is similar to the second perspective mentioned at the beginning of footnote 35.) Then by footnote 51, this makes the $SO(4)$ global symmetry anomalous.

[54] We emphasize again that in this paper we are only concerned with the topological physics due to $\pi_3 \cong \mathbb{Z}$. In the physical $d = 4$ spacetime, the actual $S^3 \cong |SU(2)|$ pion vacua effective theory contains a non-trivial 4d WZW term due to $\pi_5(S^3) \cong \mathbb{Z}_2$, and there are also topological effects due to $\pi_4(S^3) \cong \mathbb{Z}_2$; for $SU(N > 2)$, the pion-kaon vacua effective theory contains a non-trivial 4d WZW term due to $\pi_5(|SU(N > 2)|) \cong \mathbb{Z}$ [94, 95]. Hopefully our general framework to be sketched in Section 6 will lead to natural lattice definitions of these terms in future works.

We have described the crucial properties that the weight factors (and most particularly the μ function in W_2) should have, but how to optimize the weight factors in detail for best numerical performance is subjected to numerical investigation, and is indeed beyond the scope of the present work. Since some d.o.f. can no longer be group elements, the W_2 weight no longer has a simple analytic description in terms of the trace of some group element or so, and might need to be stored as a somewhat complicated function. [55] In practical implementation, if the phase $\mu/|\mu|$ slightly deviates from the value we described, there should be no crucial problem. Moreover, as a crude approximation, it is even consistent to set $\mu = 0$ when any of the $m_{l \in \partial p}$ involved is $-$; if this indeed works well numerically, then the implementation will be greatly simplified (and the $\hat{n}_l \in S^2$ can be entirely ignored).

It worths to reiterate the relation between the refined lattice nlσm and the traditional lattice nlσm. If we ignore the plaquette and cube d.o.f. and weights in (65), we can recover the traditional model via (58). But once the plaquette and cube d.o.f. and (reasonably chosen) weights are taken into account, there can be no exact recovery. This situation is similar to the inclusion of vortex fugacity weight in the Villain model, (12). As we emphasized there, instead of being a problem, this is expected (and analytically established in the Villain case [14]) to be a feature that *helps* control the renormalization, because in the renormalization process the effect of such d.o.f. and weights will be generated anyways, so having them in the model is expected to help keep track of the renormalization flow. Numerical computations will be needed to determine the renormalization behavior of the plaquette and cube weights.

Finally, we briefly explain how to generalize to nlσm with target space $\mathcal{T} = |SU(N)|$ beyond $N = 2$. Again we diagonalize $Dg_l = \mathcal{U}_l e^{i\lambda_l^a \tau_a} \mathcal{U}_l^{-1}$, where τ_a ($a = 1, \cdots, N-1$) is a set of generators for the root lattice, and the space of eigenvalues is parametrized by λ_l^a in the Weyl alcove, which for $SU(N)$ is an $(N-1)$-dimensional simplex (this generalizes $\lambda_l \in [0, \pi]$ for $N = 2$). Each of the co-dimension 1 faces of the simplex—there are N of them—corresponds to a pair of adjacent eigenvalues becoming degenerate (this generalizes $\lambda_l = 0$ and $\lambda_l = \pi$ for $N = 2$). We cover $SU(N)$ by N patches, each removing one pair of eigenvalue degeneracy, i.e. each removing one face of the Weyl alcove. Each patch can be labeled by the corner of the Weyl alcove that is opposite to the face being removed, which for $SU(N)$ turns out to be an element of the $\mathbb{Z}_N$ center (this generalizes the labels $m_l = \pm = \pm \mathbf{1}$ for the patches $\lambda_l \in [0, \pi)$ and $\lambda_l \in (0, \pi]$ respectively for $N = 2$), though this does not mean the m_l labels are going to be able to compose like a $\mathbb{Z}_n$ group. For the $m_l = \mathbf{1}$ patch, the representative path is given by connecting a straight line in the Weyl alcove from the origin to the point that represents λ_l^a, and then conjugating this path by $\mathcal{U}_l$ before multiplying by g_v on the right. For any other m_l patch, the representative path has two segments, one segment is given by connecting a straight line in the Weyl alcove from the corner that labels m_l to the point that represents λ_l^a, and then conjugating this path by $\mathcal{U}_l$ before multiplying by g_v on the right; the other segment is given by the edge of the Weyl alcove connecting the origin to the corner that labels m_l, and then conjugating this

---

[55]Some automated optimization program might be useful, for instance W_2 might be implemented as the output of some machine learning task.

edge by an arbitrary element in $SU(N)/U(N-1)$ where $U(N-1)$ is the subgroup that commutes with this edge (this generalizes the $\hat{n}_l \in SU(2)/U(1) \cong S^2$ for $N = 2$), before multiplying by g_v on the right. Thus we have described the non-fibre-bundle cover Y over $SU(N)$ and the representative paths. The rest is essentially the same as $N = 2$. We weigh the lattice WZW curving $e^{i\mathcal{W}_p}$ by constructing a suitable function μ, which involves suitably choosing interpolating 2d surfaces and integrating the continuum WZW curvature over the resulting pyramid—the closeness of the continuum WZW curvature is crucial here because that makes the integral independent of the choice of the interpolating 3d volume out of the $N(N-1)/2$-dimensional space of $SU(N)$. Finally we Villainize the lattice WZW curvature.

## 4.2 $SU(N)$ lattice gauge theory: Chern-Simons, instanton and Yang monopole

From the experience with Villainized S^1 lattice nlσm and Villainized $U(1)$ lattice gauge theory introduced in Sections 2.1 and 2.2, it is intuitive to expect that, now that we have topologically refined the $|SU(N)|$ lattice nlσm, the topologically refined $SU(N)$ lattice gauge theory can be obtained by "putting the d.o.f. on cells of one higher dimension". What this really means is the following: Traditionally, the lattice instanton is defined by interpolating the lattice gauge field to a continuum gauge field [8], and the problem is the interpolation choice will run into singularities or discontinuities as we vary the lattice gauge field (and the treatment in [8] is to disallow strongly fluctuating gauge fields); now what we have learned from the topological refinement of $|SU(N)|$ nlσm is how to consider different possibilities of the interpolation of the $|SU(N)|$ matter field at each level of lattice cell, and we are going to apply the idea to the different possibilities of interpolating the $SU(N)$ gauge field at each level of lattice cell, which is one dimension higher compared to the counterpart in nlσm.

Again we will focus on $SU(2)$ in the below, since the generalization to $SU(N)$ using the Weyl alcove is straightforward.

Traditionally, we have a lattice gauge connection $g_l \in SU(2)$ on each lattice link, and $Dg_p \in SU(2)$ is the gauge flux around the plaquette p. We first describe how to refine the gauge flux on the plaquette. Recall in the case of nlσm, on the link we refined $Dg_l \in SU(2)$ to $y_l \in Y$, and the patches (57) were chosen to be invariant under conjugation to manifest the $SO(4)$ global symmetry. Now, in gauge theory, on the plaquette we also refine $Dg_p \in SU(2)$ to $y_p \in Y$, and the patches being invariant under conjugation is desired because lattice gauge flux transforms by conjugation under gauge transformation and under changing the choice of the starting point. The plaquette weight W_2 for gauge theory has the same qualitative properties as the link weight (58) for nlσm.

In nlσm, an element $y_l = (Dg_l, m_l, \hat{n}_l)$ has been pictured as choosing an interpolating path from g_v to $g_{v'}$, that will be used in designing the plaquette weight. Now in gauge theory, an element $y_p = (Dg_p, m_p, \hat{n}_p)$ can be interpreted as choosing a way of interpolating the gauge field over the plaquette, that will be useful later in the cube weight. When $m_p = +$ (which requires $Dg_p \neq -\mathbf{1}$), the interpolation is the following. Consider the holonomy around a

portion of the plaquette, indicated by the shaded area:

$$(68)$$
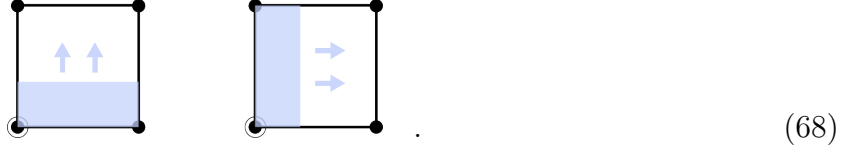
As the portion increases its size in either direction (indicated by the arrows), the holonomy around it interpolates along the shortest geodesic from $\mathbf{1}$ to Dg_p. [56] This essentially agrees with how the gauge field on the link is interpolated into the plaquette in [8]. When $m_p = -$ and some $\hat{n}_p$ is given, as the size of the portion increases, the holonomy interpolates in the alternative way as explained by (59) in the case of nlσm.

On the cube, there is a $U(1)$ dynamical field $e^{i\mathcal{C}_c}$ which is interpreted as the lattice version of the CS 3-form. Similar to the WZW curving d.o.f. in the case of nlσm, here the CS d.o.f. forms a non-trivial $U(1)$ bundle over the g_l and y_p on the links and plaquettes around the cube, and is weighed by some $W_3(e^{i\mathcal{C}_c}\nu^*_{g_l\in\partial c, m_p\in\partial c, \hat{n}_p\in\partial c} + c.c.)$ in analogy to (60). Just like the μ function in (60), here the ν function has the following properties: Its phase $\nu/|\nu|$ is given by interpolating the gauge fields on the plaquettes around the cube into the inside of the cube via some standardized procedure (as mentioned above, when all $m_p = +$, the plaquette interpolation is the same as that in [8], then we can also use the cube interpolation in [8]; when some $m_p = -$, some other interpolation into the cube will be used, similar to the nlσm case) [57] and then taking the continuum CS integral over the cube. Since the continuum CS term is gauge dependent, the phase $\nu/|\nu|$ will also be gauge dependent, but under gauge transformation it only changes by a lattice exterior derivative. Fluctuations of $e^{i\mathcal{C}_c}$ away from $\nu/|\nu|$ is interpreted as effectively capturing the fluctuations of the gauge field inside the cube away from the standard interpolation. Singularity in the phase $\nu/|\nu|$ due to singularity in the gauge dependence of the continuum CS term does not matter since such singularity will always drop out, while singularity in the choice of the standard interpolation of gauge field into the cube should occur at where $|\nu|$ decreases to 0. [58]

The above are the key requirements for the ν function in the cube weight. For concreteness, we will present one particular way to construct the standard interpolations and the ν function in a separate work [10]—because the procedure is highly technical and takes some length to describe. Some highly technical aspects are borrowed from [8], but used in a conceptually different way; moreover, at the technical level we also improved some expressions from [8] so that the construction can be described in terms of Wilson loops instead of more

---

[56]Note this description of the plaquette interpolation is independent of the gauge choice except at the starting point of the loop (as indicated at the lower left corner of the plaquette), and gauge transformation at the starting point acts by conjugation on the holonomy around the shaded area of any size.

[57]An important requirement of the procedure is that, just like in the previous footnote, the description of the standard interpolation into the cube must be stated in terms of holonomies, so that the description is gauge independent except at the starting point of the loop.

[58]In the model of [96], there is also a dynamical lattice CS $U(1)$ field weighted with some saddle. However, [96] does not include the dynamical variables on the plaquettes and the hypercubes (to be introduced below), and moreover there the counterpart of $|\nu|$ is a constant rather than a function which can, crucially, vanish under certain conditions. Hence the problem of discontinuity persists.

general Wilson lines, making the construction manifestly gauge invariant. Of course, in actual practice, what detailed design works the best is subjected to numerical investigations. And just like in the case of nlσm, we guess it might be a consistent approximation to just let $|\nu| = 0$ whenever any $m_{p \in \partial c} = -$, and this would largely simplify the implementation. How good this approximation is in capturing the physics is, again, subjected to numerical investigations.

In 3d, we can define the non-abelian CS phase with level $k \in \mathbb{Z}$ on lattice as

$$W_{CS}^k := e^{ik \oint_{3d} \mathcal{C}} := e^{ik \sum_c \mathcal{C}_c} . \tag{69}$$

If the 3d space has boundary, Dirichlet boundary condition is required to avoid gauge dependence on the boundary. Unlike the abelian CS (24) which depends on the traditional link gauge field explicitly, here the non-abelian CS only depends on the link gauge field probabilistically through the weights W_2, W_3, constituting a CS-Yang-Mills theory. (Even in the continuum CS theory, a Yang-Mills term with tiny coefficient is secretly understood in the regularization of the eta-invariant [70].)

In 4d, on the hypercube, we Villainize $e^{id\mathcal{C}_h}$ by introducing an integer d.o.f. $\iota_h \in \mathbb{Z}$, so that the non-abelian instanton density is defined as

$$\mathcal{I}_h := \frac{d\mathcal{C}_h}{2\pi} + \iota_h \tag{70}$$

(cf. (24)). [59] There is a hypercube weight $W_4(\mathcal{I}_h)$ that is positive and decreasing with $|\mathcal{I}_h|$. The total instanton number

$$I := \oint_{4d} \mathcal{I} = \sum_h \mathcal{I}_h = \sum_h \iota_h \in \mathbb{Z} . \tag{71}$$

---

[59]Our definition of instanton density reduces to that in [8], if we consider weak enough field strength, and use the saddle point approximation on the new local weights W_4, W_3, W_2, i.e. always choose those new d.o.f. $\iota_h, e^{i\mathcal{C}_c}, m_p(= +)$ that maximize W_4, W_3, W_2. We will discuss this comparison to [8] in greater details in a separate work [10]; for now let us briefly explain the idea.

In [8], the gauge field (assumed weak field strength) on the links is interpolated via a standard procedure into the plaquettes and the cubes. This corresponds to choosing $m_p = +$ (which maximizes W_2 when the field strength is weak) and choosing $e^{i\mathcal{C}_c} = \nu/|\nu|$ (which maximizes W_3). Let us explain the latter point. In [8] there was no explicit mention of CS, but the instanton density has been expressed as a sum of terms on the cubes around the hypercube, and these terms are effectively playing the role of $\mathcal{C}_c$. A practical difference is that, in [8], the gauge fields in different hypercubes are under different gauge choices (referred to as the "complete axial gauge" in each hypercube), and hence the CS on a cube c has two gauge choices—one each hypercube on the two sides of that cube (let us thus denote the CS value as $\mathcal{C}_c^{(\text{gauge of } h)}$ where $c \in \partial h$); while here, the CS on a cube uses only one gauge, which is good for defining the CS phase W_{CS}^k in 3d (which is not part of the consideration in [8]). Note that $e^{i2\pi\mathcal{I}_h} = e^{id\mathcal{C}_h}$ is gauge independent. However, if the logarithm of it is defined as $2\pi\mathcal{I}_h = d\mathcal{C}_h^{(\text{gauge of } h)}$ following [8], then gauge choice on h determines the $\mathbb{Z}$ part of $\mathcal{I}_h$; this allows for a non-zero value of $\oint_{4d} \mathcal{I}$ in [8]. On the other hand, we defined $2\pi\mathcal{I}_h = d\mathcal{C}_h + 2\pi\iota_h$ with some new dynamical integer ι_h. For weak enough field strength, if we take the value of ι_h that minimizes $\mathcal{I}_h$ (hence maximizes W_4), then the value of $\mathcal{I}_h$ will agree with that defined by [8].

A topological theta term can hence be defined in $d = 4$. [60] In $d \geq 5$, the instanton non-conservation defect $d\mathcal{I} \in \mathbb{Z}$ is the Yang monopole, which can be suppressed by some W_5, or forbidden by W_5^{forbid} which contains a $(d-5)$-form $U(1)$ Lagrange multiplier, manifesting a $(d-5)$-form dual $U(1)$ global symmetry.

Piecing up the discussions above, we have our second main result, the $SU(2)$ lattice gauge theory refined to include instanton and topological theta term reads

$$Z_\Theta = \left[\prod_{l'} \int_{SU(2)} dg_{l'}\right] \left[\prod_{p'} \sum_{m_{p'}=\pm} \int \frac{d^2\hat{n}_p}{4\pi}\right] \left[\prod_{c'} \int_{-\pi}^{\pi} \frac{d\mathcal{C}_{c'}}{2\pi}\right] \left[\prod_{h'} \sum_{\iota_{h'}\in\mathbb{Z}}\right]$$
$$e^{i\Theta I} \prod_p W_2(\lambda_p, m_p) \prod_c W_3(e^{i\mathcal{C}_c}\nu^*_{g_{l\in\partial c}, m_{p\in\partial c}, \hat{n}_{p\in\partial c}} + c.c.) \prod_h W_4(\mathcal{I}_h) \qquad (72)$$

for $d = 4$; see [10] for the highly technical details of the ν function. The d.o.f. together form a mathematical structure that implements the second Chern class of the lattice $SU(2)$ Yang-Mills theory, to be explained with (130). For $d \geq 5$ there is no topological theta term, but there can be a weight W_5 or W_5^{forbid} for the Yang monopole $d\mathcal{I}$. For $d = 3$, there is no ι_h and W_4, but we can additionally consider a CS term

$$Z_{kCS} = \left[\prod_{l'} \int_{SU(2)} dg_{l'}\right] \left[\prod_{p'} \sum_{m_{p'}=\pm} \int \frac{d^2\hat{n}_p}{4\pi}\right] \left[\prod_{c'} \int_{-\pi}^{\pi} \frac{d\mathcal{C}_{c'}}{2\pi}\right]$$
$$W_{CS}^k \prod_p W_2(\lambda_p, m_p) \prod_c W_3(e^{i\mathcal{C}_c}\nu^*_{g_{l\in\partial c}, m_{p\in\partial c}, \hat{n}_{p\in\partial c}} + c.c.) \qquad (73)$$

for any $k \in \mathbb{Z}$. The generalization from $SU(2)$ to $SU(N)$ is straightforward using the Weyl alcove parameterization introduced at the end of Section 4.1.

An important aspect of $SU(N)$ Yang-Mills theory is the 1-form $Z(SU(N)) \cong \mathbb{Z}_N$ global symmetry $g_l \to g_l e^{i\beta_l}$ for $\beta_l \in (2\pi/N)\mathbb{Z}_N$ such that $e^{id\beta_p} = 1$ (which we gauged in (33) to obtain the Villainized $PSU(N)$ gauge theory). Since this transformation leaves Dg_p invariant, it would not interfere with y_p and W_2. Moreover, since in the continuum the CS integral also respects this 1-form global symmetry, the ν function—which is conceptually defined using the continuum CS integral—respects this symmetry on the lattice, hence so does the lattice CS and the lattice instanton density.

Upon on introducing a 2-form background for this 1-form global symmetry, there should be a self-anomaly in the presence of a non-trivial CS weight in 3d, a breaking of the 2π periodicity of the topological Θ angle in 4d, [61] and a mixed anomaly with the dual $(d-5)$-form $U(1)$ that forbids the Yang monopole in $d \geq 5$. We will leave to future works to investigate how to see these anomalies explicitly on the lattice. [62]

---

[60] Just like in the abelian case (footnote 24), one can also let the theta become local and dynamical, but then for consistency we will need to introduce the Villainization integer field for this theta, that lives on the dual lattice link and couples to the CS density. We then obtain the lattice axion theory.

[61] See [97] for a lattice demonstration of this under the traditional notion [8] of lattice instanton.

[62] Let us explain what we should anticipate. To see these anomalies, we need to introduce the 2-form $\mathbb{Z}_N$

Now that we have explained the topologically refined $SU(N)$ lattice gauge theory, let us look back and discuss some important conceptual issue regarding the relation between the topological refinement of the $|SU(N)|$ nlσm and that of the $SU(N)$ gauge theory, which is obviously more involved than the relation between Villainized S^1 nlσm and Villainized $U(1)$ gauge theory.

Recall the link variable in nlσm is geometrically interpreted as sampling some representative path in $SU(N)$; when two links are joined together on the lattice, their associated representative paths also concatenate in the obvious manner. But such kind of concatenation interpretation becomes subtle in gauge theory. In gauge theory, the $SU(N)$ is the space of holonomy around some loop. In our refinement at the plaquette level, we deform the loop with one parameter, so that the loop increases its size to wipe over the plaquette, and throughout the process the holonomy indeed traces out a representative path in $SU(N)$. Now consider two plaquettes p, p' joined at a shared vertex. Each plaquette has been associated with a path in $SU(N)$, one path connecting $\mathbf{1}$ and Dg_p and the other connecting $\mathbf{1}$ and $Dg_{p'}$:

$$(74)$$
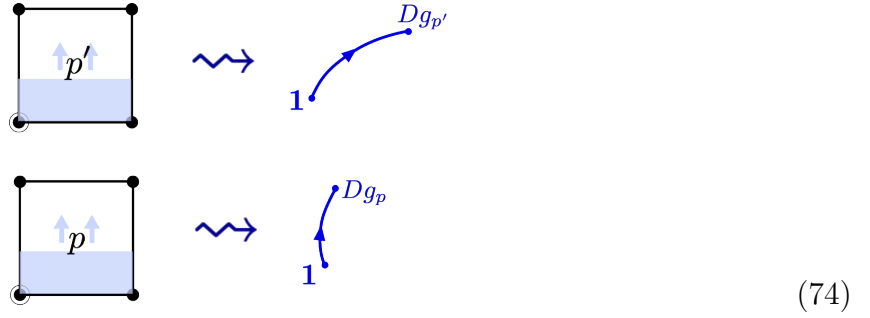
Now that the two plaquettes p, p' are joined together, do their associated paths somehow get joined together, too? There are two ways to increase the shaded area to fill up the two plaquettes, one filling up p first and then p', the other filling up p' first and then p. Suppose we choose the lower left corner of p (see picture below) as the starting point of the loop, then the holonomy around p' (or around any portion of p') needs to be conjugated by suitable Wilson lines. The two ways of filling lead to two different paths in $SU(N)$, though they

---

background gauge field. What we anticipate (and should further verify) is that there should be no way to define a ν function whose phase is invariant under the 1-form $\mathbb{Z}_N$ gauge transformation when the associated 2-form background is non-trivial. Suppose this is indeed so, then the remaining is straightforward: In $d = 3$, without a CS weight, this transformation of $\nu/|\nu|$ can be absorbed by a transformation of $e^{i\mathcal{C}_c}$, leaving the theory invariant; but when the CS level is non-trivial, the theory will not be left invariant, manifesting the said self-anomaly. Related to this, in $d \geq 5$ with W_5^{forbid}, if we introduce a non-trivial Villainized background for the dual $(d-5)$-form $U(1)$ on the dual lattice, its background Dirac string field (which is a $(d-3)$-form integer field on the dual lattice) will couple to $e^{i\mathcal{C}_c}$, i.e. a CS weight—which makes the 1-form $\mathbb{Z}_N$ anomalous—is attached on the Dirac string, manifesting the said mixed anomaly.

share the same starting and ending points:

$$(75)$$
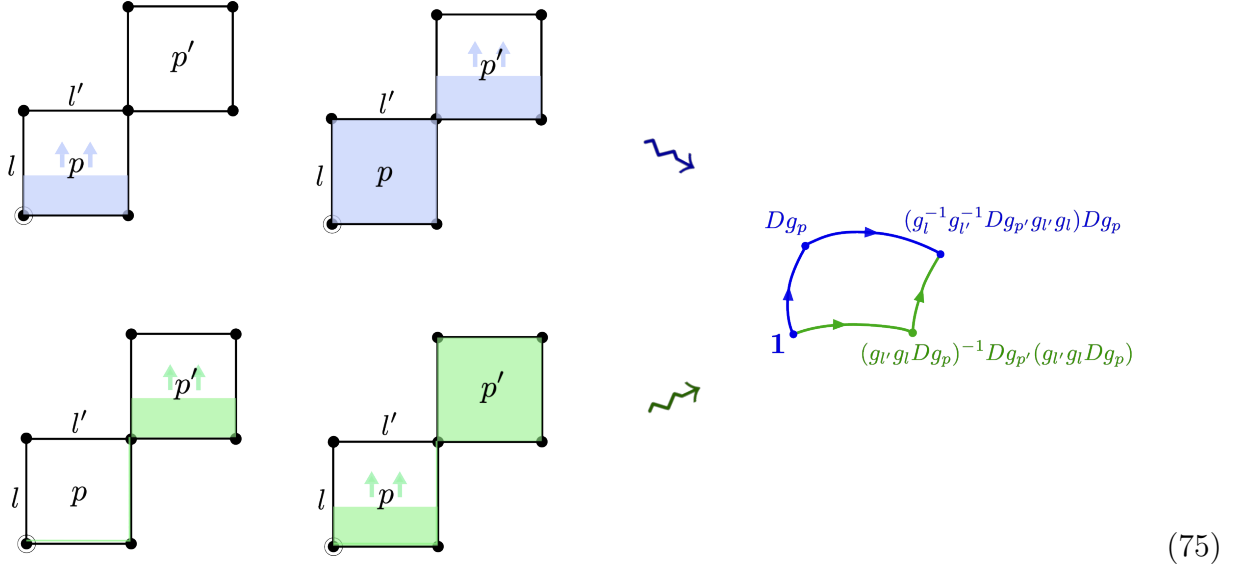
Unlike joining two links in nlσm where there is a natural ordering of which link comes first, when joining two plaquettes there is no natural choice of ordering, [63] so there is no way to determine which of the two ways of composing the representative paths is "the better choice". [64] (Similar issue happens when the two plaquettes are joined together at a shared edge instead of a shared vertex, or even more generally, joined together by a finite length Wilson line. The case of shared vertex pictured above is what will be relevant below.)

Why this issue worths any discussion? Because this is the underlying reason why passing from the topologically refined lattice $|SU(N)|$ nlσm to $SU(N)$ gauge theory is not as simple as passing from Villainized S^1 nlσm to Villainized $U(1)$ gauge theory. And the root of this lies in some important subject in category theory—delooping and Yang-Baxter equation.

Recall in $|SU(N)|$ nlσm, the WZW curving d.o.f. on the plaquette is interpreted as sampling some 2d surface in $SU(N)$ (and two surfaces are consider equivalent if they bound a volume over which the WZW integral vanishes—recall the discussion below (51)), [65] and the skyrmion density over the cube is interpreted as (the WZW integral over) some 3d volume in $SU(N)$. However, in gauge theory, it would not be very useful to think of the CS d.o.f. on the cube as some kind of 2d surface in $SU(N)$ and think of the instanton density over a hypercube as some kind of 3d volume in $SU(N)$.
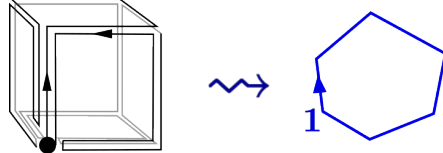
To understand this point, let us suppose we do think in this way and see what difficulties we run into. First consider the six plaquettes around a cube, which are joined together and leave no 1d boundary behind. Moreover, we can choose some ordering of filling up the plaquettes (such as the ordering we used in (35)), and once this ordering is fixed, the paths

---

[63]We have encountered this when discussing higher form symmetries / degrees of freedom in Section 2.3.

[64]Using Mickelsson product (as in [38] and [90]) instead of geometrical concatenation does not help with this problem.

[65]In particular, some standard choice of surface (for the base of the pyramid in (61)) is used in defining $\mu/|\mu|$, and the deviation of WZW curving d.o.f. $e^{i\mathcal{W}_p}$ away from $\mu/|\mu|$ captures the fluctuation of the surface away from the standard choice (up to the said equivalence relation).

associated with each plaquette should join together (after conjugations by suitable Wilson lines) unambiguously and form some hexagonal loop in $SU(N)$:

$$\tag{76}$$

[66] Suppose we interpret the CS d.o.f. on the cube as sampling some surface bounded by this hexagon (trying to mimic what happens for the WZW curving d.o.f. in nlσm). So far there is no problem. Next let us consider the eight cubes around a hypercube, we expect the eight associated hexagonal surfaces to glue up (again after conjugations by suitable Wilson lines) into some closed surface, which will bound some volume whose WZW integral is interpreted as the lattice instanton density. But a careful inspection shows the eight hexagonal surfaces do not glue up to a closed surface, but a truncated octahedron

hypercube $\quad\rightsquigarrow$

$$\tag{77}$$

on which the quadrangle loops are left unfilled with surfaces. These quadrangle loops precisely come from (75). While we have fixed the ordering of filling up the plaquettes around each cube, in a hypercube there are still pairs of plaquettes which do not belong to a same cube but nonetheless join at a shared vertex, [67] and for each such pair, both orderings of filling up the two plaquettes will come up when we try to join the cubes around a hypercube, leading to the open quadrangle loops on the truncated octahedron. [68]

Can we fix some standard choice of surfaces to fill up such unfilled quadrangle loops, so that the truncated octahedron that the hypercube associates with becomes a closed surface? [69] We can do so, but a further constraint must be satisfied in the standard choices that we make. Since the WZW integral over the volume bounded by the truncated octahedron is to

---

[66]If we have used a simplicial complex as the lattice, then a each tetrahedron will give rise to a quadrangle loop in $SU(N)$.

[67]These pairs of plaquettes are those that pair up in defining the cup product in footnote 22.

[68]If we have used a simplicial complex as the lattice, then the five tetrahedra in each 4-dimensional simplex will give rise to five filled faces on a cube in $SU(N)$, with the last face of the cube remaining unfilled due to the issue (75)—and the two plaquettes involved are, indeed, those that pair up in the cup product.

[69]Between two choices of surfaces to fill up the unfilled loop (75), their difference is, again, truncated to a $U(1)$ value, the WZW integral of the volume bounded between the two choice of surfaces.

Importantly, a close inspection shows this $U(1)$ now forms a trivial bundle over the space of choices of the paths around. Briefly speaking, this is because the four paths around are really only determined by two paths, one associated with p and the other associated with p', as shown in (75). In Section 4.1, there is a non-trivial constraint on the space of the link variables (see footnote 48), that gives rise to a non-trivial π_2 for the space of links variables, which then leads to the non-trivial WZW curving $U(1)$ bundle; by contrast, here, the corresponding constraint will be automatically satisfied, so that the space of the possible choices of paths in (75) has trivial π_2, and therefore any $U(1)$ bundle on it is necessarily trivial.

be interpreted as the lattice instanton density over the hypercube, we need to ensure that $d\mathcal{I}$ as well as $\oint_{4d} \mathcal{I}$ result in an integer. This requirement is equivalent to stating that when ten hypercubes piece up to a 5d-hypercube, we want the ten associated truncated octahedrons in $SU(N)$ to piece up to a closed 3d volume without any 2d surface leftover. This is a highly non-trivial constraint imposed on the standard choice of quadrangle surface. In fact, this constraint is a generalized version of Yang-Baxter equation, and specifying a standard choice for the quadrangle surface is an example of braiding data in category theory, as we will discuss in Sections 5.3 and afterwards.

It is in general a very difficult task to find non-trivial solutions to a Yang-Baxter equation. This is why, in our construction for lattice gauge theory, we *do not* take the perspective described here. That is, we do not think of the lattice CS d.o.f. on the cubes as sampling some surface in $SU(N)$ and the instanton density over the hypercube as some volume in $SU(N)$. This means we do not literally take the geometrical interpretation of those fields in the nlσm and "put them on lattice cells of one higher dimension" to get the geometrical interpretation of the fields in the gauge theory. Instead, in our construction of gauge theory, the geometrical interpretation is about how to interpolate the gauge fields on the links into the plaquettes and the cubes, much like in the previous work [8], except we consider different possibilities of interpolations in a manner guided by category theory. This interpretation makes better intuitive connection to the continuum gauge theory, and the Yang-Baxter equation issue never explicitly comes up—we can view it as being automatically solved. The reason behind this will be explored in Section 6.2.

5   Category Theory Foundation

In this section we first explain how to cast the previously known examples in Section 2 into the language of category theory, and then we will see how our construction in Section 4 can be naturally motivated from there.

The involvement of category theory in the lattice model construction is, however, more than just being motivational. We have been saying "we want a model that captures the π_3 topology of the nlσm or Yang-Mills theory on lattice", but so far we only have some intuitive idea what this "capture" is intended to mean. After some initial build-ups, in Section 5.5, we will turn this intuitive goal into a precise mathematical statement, i.e. making it clear what the mathematical requirements are for a lattice QFT to "capture the π_3 topology", and why the construction in Section 4 indeed serves this purpose.

We will begin by introducing some basics of category theory so that we can setup our notations and eventually lead the discussion towards the concepts that we will need for our construction. However, this section is not intended as a piece of comprehensive and/or rigorous introductory material to the subject of category theory itself. A gentle introduction to category theory containing some physics oriented perspectives can be found in [40]. For more comprehensive and rigorous treatment, one may consult textbooks and review articles of different levels and with different emphases. The online wiki nLab is a very useful source of knowledge on this subject.

5.1 Strict categories, and the known examples

We begin with *strict* higher categories. Being "strict" implies they are straightforward to define and easy to understand, but not as powerful as the more general higher categories in being descriptive. It is not surprising that the previously known examples introduced in Section 2 are all described in terms of strict higher categories.

A 0-category C is just a set C_0. [70] The elements in it are often called "objects" in the context of category theory. Often times C_0 can be endowed with extra structures, for examples it can be a group, a topological space or smooth manifold, etc.

A 1-category, which is what a "category" usually refers to, has two sets: the set C_0 of objects, and the set C_1 of all "morphisms", or relations, between objects. Of course the two sets should not be independent. There are some maps between them:

- Intuitively, C_1 should have a "source map" s and a "target map" t to C_0, so that $\mathsf{s}(f) = a, \mathsf{t}(f) = b$ means f is a morphism (relation) from object a to object b, which we can denote as $b \xleftarrow{f} a$. [71] Because of these two maps, we will often denote a category C as $C_1 \rightrightarrows C_0$. We use $C_1|_{b,a}$ to denote the subset of C_1 where the source and the target are restricted to a and b respectively. [72]

- Morphisms (relations) should be able to compose, i.e. there is a map $\circ$ from $C_1 \times_{C_0}^{\mathsf{s},\mathsf{t}} C_1$ to C_1 (where the fiber product notation $X \times_Z^{u,v} Y := \{(x,y) \in X \times Y | u(x) = v(y) \in Z\}$), or say from $C_1|_{c,b} \times C_1|_{b,a}$ to $C_1|_{c,a}$ for every $a, c \in C_0$. This specifies how $c \xleftarrow{g} b$ and $b \xleftarrow{f} a$ are composed to some $c \xleftarrow{g \circ f} a$. Moreover, when composing three morphisms, the composition should be associative. (Sometimes we might omit the "$\circ$" in composition.)

- There is a map i from C_0 to C_1, which for each $a \in C_0$ specifies an "identity morphism" $\mathbf{1}_a := \mathsf{i}(a) \in C_1|_{a,a} \subseteq C_1$, such that under composition, $f \circ \mathbf{1}_a = f$ for any f with $\mathsf{s}(f) = a$, and $\mathbf{1}_a \circ g = g$ for any g with $\mathsf{t}(g) = a$.

If C_0 and C_1 are endowed with some extra structure, then it is natural to require these maps to respect the extra structure. Particularly, if C_0 and C_1 are both manifolds, then it is natural to require these maps to be smooth—apparently this will be an important point in our application, and this point will be systematically formulated in terms of *internalization* in Section 5.2.

---

[70]Rigorously speaking, there are large versus small categories, where large categories can involve collections such as proper classes which are logically "larger" than any possible set (e.g. the collection of all sets is a proper class, which is not a set in the sense that it is not allowed to be taken as an element of any set or proper class, hence resolving the Russell paradox). However, such issue does not seem very important in physics. The categories that are directly involved in our detailed construction are all small categories.

[71]Usually the arrow is drawn from left to right. In this paper we will often use the convention from right to left, because when we compose functions or operations this is the conventional order of action.

[72]Usually the notation is $\mathrm{Hom}(a,b)$ or $\mathrm{Hom}_C(a,b)$. Here we emphasize that we prefer to primarily view the morphisms altogether as a whole set C_1, rather than to primarily view the morphisms between each given pair of objects, $\mathrm{Hom}(a,b)$, as a set. Of course these two views are equivalent for now, but the former view will be more suitable for generalization via *internalization*, which will be important for our work, while the later view is more suitable for generalization via *enrichment*, which we will not focus on (though also important in general).

A morphism $b \xleftarrow{f} a$ is invertible if there exists an $a \xleftarrow{f^{-1}} b$ such that $f^{-1} \circ f = \mathbf{1}_a$, $f \circ f^{-1} = \mathbf{1}_b$. [73] Now, we are ready to see that in category theory, a group can at least be perceived in two ways: as a 0-category (set) G endowed with the some extra structure that makes it a group, or as a 1-category BG, where BG_0 has only a single object, and every morphism in BG_1 is invertible, $BG := (G \rightrightarrows *)$. Such relation between G and BG is a simple example of *delooping* (from G to BG) and *looping* (from BG to G), a concept that will be important in our application when elevating the construction for nlσm to the construction for gauge field. One might note that the notation BG is also conventionally used to denote the classifying space of G. This is not a coincidence. The classifying space, which we will denote as $|BG|$ and is usually infinite dimensional, can be obtained from the category BG via the procedure of *geometric realization* which we will introduce in Section 5.4.)

More generally, a 1-category where every morphism in C_1 is invertible, but not necessarily with only a single object in C_0, is called a *groupoid*. (It is therefore said that the notion of groupoid is the "horizontal categorification" of the notion of group. More generally, a 1-category with a single object—but not necessarily with every morphism invertible—can be viewed as the delooping of a monoid, and thus the notion of 1-category is the "horizontal categorification" of the notion of monoid.) An intuitive example of a groupoid is an *action groupoid* $X \times G \rightrightarrows X$, where there is a set X and a group G acting on X, so that a morphism $(x, g) \in X \times G$ is depicted as $gx \xleftarrow{(x,g)} x$ or more simply $gx \xleftarrow{g} x$.

Groupoids are very common in our application. For some examples:

1. Given a continuum manifold $\mathcal{M}$ we can define its free path space $\mathcal{P}\mathcal{M}$. [74] Now we consider $\bar{\mathcal{P}}\mathcal{M}$, which is $\mathcal{P}\mathcal{M}$ with equivalence up to "thin homotopy", i.e. up to reparametrization and identifications like
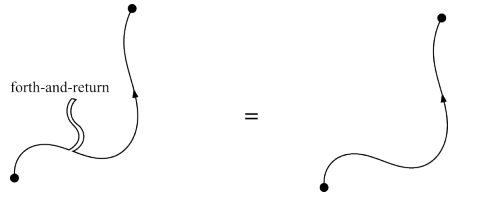
   so that the concatenation of paths is associative, and has identity path for each point and inverse path for each path. [75] We thus defined the *path groupoid* $\bar{P}_1\mathcal{M} := (\bar{\mathcal{P}}\mathcal{M} \rightrightarrows$

---

[73]It is also possible that only one of these two conditions can be satisfied, or both conditions can be satisfied but with two different "f^{-1}"s. Therefore in general we should define the notions of left inverse f_L^{-1} and right inverse f_R^{-1} of f.

[74]Compared to the pointed path space $\mathcal{P}_*\mathcal{M}$ we introduced before, the free path space $\mathcal{P}\mathcal{M}$ does not fix a starting point for the paths. They are related by the fibre bundle $\mathcal{P}_*\mathcal{M} \to \mathcal{P}\mathcal{M} \to \mathcal{M}$.

[75]More exactly, a path is a smooth function $\gamma(\tau) \in \mathcal{M}, \tau \in [0,1]$. Composition, i.e. concatenation, is defined by $(\gamma' \circ \gamma)(\tau)$ equals $\gamma(2\tau)$ if $0 \leq \tau \leq 1/2$ and $\gamma'(2\tau - 1)$ if $1/2 \leq \tau \leq 1$. To ensure the smoothness around the concatenation point, we need a "sitting instant" condition that $\gamma(\tau)$ stays constant for $|\tau - 0| < \epsilon$, $|\tau - 1| < \epsilon$ for some small ϵ.

The "thin homotopy" identification is that, two paths γ_1 and γ_2 are considered identified if they are

$\mathcal{M}$). [76]

A closely related concept is the *fundamental groupoid* $\Pi_1\mathcal{M}$ (we will explain this name in the next subsection), which is like the path groupoid except the identification of paths is not only made under thin homotopy, but any homotopy (any interpolation). Clearly, if $\mathcal{M}$ is 1d, then $\Pi_1\mathcal{M} = \bar{P}_1\mathcal{M}$, because any homotopy between paths is necessarily thin.

2. A lattice keeping only the vertices and the links but ignoring plaquettes and higher cells forms a groupoid $\bar{\mathcal{L}}_1 \rightrightarrows \bar{\mathcal{L}}_0$, where $\bar{\mathcal{L}}_0$ is the set of all vertices, and $\bar{\mathcal{L}}_1$ is the set of all lattice paths obtained by joining links, and each path indeed has an inverse path.

3. For $\mathcal{M} = S^1$, the path groupoid is $S^1 \times \mathbb{R} \rightrightarrows S^1$, which is at the same time an example of action groupoid. Apparently this structure will be related to the d.o.f. used in the Villain model.

4. Another example of action groupoid is $S^2 \times SU(2) \rightrightarrows S^2$, which will apparently be related to the spinon decomposition.

Having introduced 0- and 1-category, it is not hard to envision that higher categories involve more layers of higher morphisms equipped with suitable maps in-between. But now there arise definitions of different levels of strictness. The more strict ones are easier to define, but the less strict ones are more flexible and thus have higher descriptive power. Here we first introduce the strict higher categories, which are sufficient to describe the known examples in Section 2; in Section 5.3 we briefly introduce more flexible higher categories which are commonly used in describing topological phases and generalized symmetries; for our construction in Section 4, we will need an even more flexible version of higher categories to be introduced in Section 5.4.

A *strict* 2-category has, first of all, a 1-category $C_1 \rightrightarrows C_0$, but in addition, there is a set C_2 of "2-morphisms" between pairs of 1-morphisms which share the same source and target objects. Pictorially a 2-morphism φ takes a globular shape

$$b \overset{f}{\underset{g}{\Longleftarrow}} \Downarrow \varphi \quad a \qquad . \tag{78}$$

Thus C_2 has a source and a target map to C_1, such that when further taking the source or target map to C_0, we require $\mathsf{ss}(\varphi) = \mathsf{st}(\varphi)$, $\mathsf{ts}(\varphi) = \mathsf{tt}(\varphi)$. There are maps $\circ_{\mathsf{v}}$ from $C_2 \times_{C_1}^{\mathsf{s,t}} C_2$

_____

related by a "thin homotopy", i.e. there is a 2-parameter interpolation $\widetilde{\gamma}(\tau, \lambda)$ from $\gamma_1(\tau)$ to $\gamma_2(\tau)$, such that everywhere the differentials $(\partial_\tau\widetilde{\gamma}, \partial_\lambda\widetilde{\gamma})$ fails to be full rank, i.e. spans a less than 2 dimensional vector space tangent to $\mathcal{M}$ (so the image of $\widetilde{\gamma}(\tau, \lambda)$ in $\mathcal{M}$ has zero area, hence "thin"); moreover, $\widetilde{\gamma}$ itself satisfies the sitting instant condition in both τ and λ. See e.g. [46].

[76]We would like the spaces $\mathcal{P}\mathcal{M}, \bar{\mathcal{P}}\mathcal{M}$ (as well as any maps involved) to be "smooth". But $\mathcal{P}\mathcal{M}, \bar{\mathcal{P}}\mathcal{M}$ are in general infinite dimensional, and a suitable generalization of "smooth" to the infinite dimensional cases is known as "diffeological". We will mention this in Section 5.2.

to C_2 and $\circ_h$ from $C_2 \times^{\mathsf{ss,tt}}_{C_0} C_2$ that define the *vertical and horizontal compositions*

$$(79)$$

which are required to satisfy vertical and horizontal associativity, as well as interchangeability, i.e. in each kind of these situations,

$$(80)$$

the composition result is unique regardless of the order of composition, similar to the associativity requirement for 1-morphisms. Finally there is a map i from C_1 to C_2 that specifies, for each 1-morphism f, the identity 2-morphism $\mathsf{i}(f) = \mathbf{1}_f$ under vertical composition. On the other hand, in horizontal composition, when $g' \overset{\varphi'}{\Longleftarrow} f'$ is given by $f' \overset{\mathbf{1}_{f'}}{\Longleftarrow} f'$, it is conventional to collapse its globular shape in the pictorial notation:

$$(81)$$

and such a horizontal composition is called a left *whiskering*. Similar for right whiskering. A strict 2-groupoid is a strict 2-category in which each 1-morphism is invertible as in a 1-groupoid and moreover each 2-morphism is also invertible.

The generalization to strict n-categories should be obvious. Now we emphasize two very useful and illuminating points of view:

- A strict m-category can be naturally seen as a special kind of strict n-category for arbitrary $n \geq m$, such that all k-morphisms for $k > m$ are identity morphisms, i.e. $C_k = \{\mathbf{1}_u | u \in C_{k-1}\} \cong C_{k-1}$ for $k > m$. We may as well take n towards infinity. This point of view will be very useful for understanding many discussions below.

- While in a 1-category, the 1-morphisms between two given objects forms a set $C_1|_{b,a}$, in a strict n-category, the q-morphisms between two objects for $1 \leq q \leq n$ form a strict $(n-1)$-category, sometimes called the *hom-category* between a and b, which we will denote as $C|_{b,a}$, with $(C|_{b,a})_0 = C_1|_{b,a}$. [77]

[77]This can be phrased in terms of *enrichment*, which roughly speaking means the hom-set $C_1|_{b,a}$ in a 1-category is replaced by some structure richer than merely a set. Thus, a strict n-category is a 1-category enriched by strict $(n-1)$-category. In this paper we will not have much emphasis on the enrichment perspective, though it is generally important in category theory.

Some examples of strict higher categories—more particularly, strict higher groupoids—that will appear in our lattice theory application include:

1. Given a d-dimensional continuum manifold $\mathcal{M}$ we can define a *strict path d-groupoid* $\bar{P}_d\mathcal{M} := (\bar{\mathcal{P}}_d\mathcal{M} \rightrightarrows \cdots \rightrightarrows \bar{\mathcal{P}}\mathcal{M} \rightrightarrows \mathcal{M})$, where $\mathcal{P}_k\mathcal{M}$ is the space of "k-paths", the interpolation between two elements of $\mathcal{P}_{k-1}\mathcal{M}$ that share the same source and target in $\mathcal{P}_{k-2}\mathcal{M}$, starting with $\mathcal{P}_0\mathcal{M} = \mathcal{M}$ and $\mathcal{P}_1\mathcal{M} = \mathcal{P}\mathcal{M}$, and $\bar{\mathcal{P}}_k$ is $\mathcal{P}_k$ with identification under thin homotopy, in order for this d-groupoid to be strict. [78] Geometrically, if $k \leq d$, a generic element in $\bar{\mathcal{P}}_k\mathcal{M}$ wipes over a k-dimensional surface (topologically a k-disc) in $\mathcal{M}$.

   Being strict makes this category easy to think of, but then not so powerful in capturing the full homotopy information. [79] However, it is still sufficient for many physical application purposes. This perspective of continuum manifold is useful for relating lattice QFT to continuum QFT.

2. A d-dimensional lattice gives rise to a strict d-groupoid $\bar{\mathcal{L}} = (\bar{\mathcal{L}}_d \rightrightarrows \cdots \rightrightarrows \bar{\mathcal{L}}_1 \rightrightarrows \bar{\mathcal{L}}_0)$. The $\bar{\mathcal{L}}_1 \rightrightarrows \bar{\mathcal{L}}_0$ part has been introduced before, while $\bar{\mathcal{L}}_i$ for $i \geq 2$ is roughly speaking i-dimensional surfaces (including degenerate ones) on the lattice, but the source and target have to be specified. In $\bar{\mathcal{L}}_2$, two elements that wipe over the same plaquette(s) can still be different, but related by whiskering, e.g.
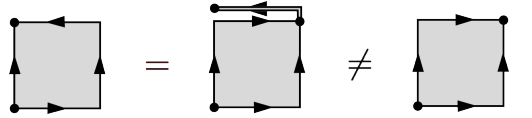
   Likewise for $\bar{\mathcal{L}}_i$, $i > 2$. We will denote by $\bar{\mathcal{L}}_{\leq m}$ (where $m \leq d$) the m-category obtained from $\bar{\mathcal{L}}$ by keeping up to $\bar{\mathcal{L}}_m$ and ignoring the higher morphisms (or equivalently, keeping only the identity higher morphisms).

   Just like the strict path n-groupoid for a continuum manifold, the strict d-groupoid $\bar{\mathcal{L}}$ does not capture the full homotopy information of the manifold that the lattice is discretizing, but it is sufficient for many physical applications.

---

[78]Similar to footnote 75, there are the higher dimensional versions of the sitting instant requirement and the thin homotopy (non-full rank interpolation) equivalence. We can take the notion of e.g. "strong 2-track" in [98] and generalize it to higher dimensional paths.

[79]Any strict higher groupoid constructed out of a manifold, regardless of the detailed construction, is incapable of capturing the full homotopy information of that manifold [39, 99]. (In general, the information of Whitehead product and beyond will be lost. Our particular construction further losses all the homotopy n-type information for $n > d$.) We will mention more about this in footnote 98. In order to capture the full homotopy information, suitable notion of weak higher category must be used, and in Section 5.4 we will introduce one such notion, simplicial weak groupoid, i.e. Kan complex, that is widely used.

If we want to define a path n-groupoid that captures the full homotopy n-type information for some finite n, there are some other particular constructions. For instance, in order to construct a weak path 3-groupoid that captures the full homotopy 3-type information, in [98], identification of 2-paths under a "laminated" condition, which is more stringent than thin homotopy, is taken, so that some 2-paths identified under thin homotopy now become distinct under this laminated condition, and the path 3-groupoid becomes a less strict kind of category—a Gray 3-category, which will be introduced in Section 5.3. Holonomies valued in Gray 3-categories can hence be considered.

3. We can ask whether BG can be delooped once more into a strict 2-groupoid $B^2G :=$ $(G \rightrightarrows * \rightrightarrows *)$. This is only well-defined when G is abelian, due to the requirement of interchangeability between vertical and horizontal compositions. [80] Obviously, when G is abelian, it can be delooped arbitrarily many of times into B^nG. And obviously, this will be related to what we discussed in Section 2.3, that higher form gauge fields must be abelian.

4. More generally, a strict 2-groupoid with a single object, but not necessarily with a single 1-morphism, is called a *strict 2-group*. It can be proven that strict 2-groups always take the "crossed module" form $B\mathcal{G} := (G \ltimes H \rightrightarrows G \rightrightarrows *)$ where G, H are groups with a homomorphism $\widetilde{t}$ from H to G [46, 99, 100]:

$$(82)$$

(more general compositions can be derived using the associativity and interchangeability conditions, with the fact that $(1,1) \in G \ltimes H$ is the identity for both the vertical and the horizontal composition), and this is the delooping of an action groupoid $\mathcal{G} := (G \ltimes H \rightrightarrows G)$ equipped with some extra structures (that make G a group, to which H has a homomorphism $\widetilde{t}$, along with a G action back on H). The interchangeability between vertical and horizontal compositions requires $\ker(\widetilde{t})$ to a subgroup of the center $Z(H)$. It is apparent that the case of $U(1) \times \mathbb{R} \rightrightarrows U(1) \rightrightarrows *$ will be related to the d.o.f. in Villainized $U(1)$ gauge theory, and it deloops the groupoid $S^1 \times \mathbb{R} \rightrightarrows S^1$ that we have discussed before.

(Even more generally, a strict 2-category with single object can be viewed as the delooping of a 1-category equipped with extra structure, and a 1-category with such extra structure is called a *strict monoidal category*.)

5. The strict 2-groupoid $S^2 \times SU(2) \times \mathbb{R} \rightrightarrows S^2 \times SU(2) \rightrightarrows S^2$, whose elements look like

---

[80]This is the basic example of the Eckmann-Hilton argument.

will apparently be related to the spinon decomposition of S^2 nlσm. The structure (37) is contained in the maps involved in the definition of this strict 2-category. In particular, given source and target objects, $(S^2 \times SU(2))|_{\hat{n}',\hat{n}} \cong U(1)$, and given source and target 1-morphisms, $(S^2 \times SU(2) \times \mathbb{R})|_{(\hat{n},\mathcal{V}),(\hat{n},\mathcal{V}e^{i\theta\hat{n}\cdot\vec{\sigma}})} \cong \mathbb{Z}$. Unwinding more structurally in order to compare with (37), we have

$$
\begin{aligned}
(S^2 \times SU(2) \times \mathbb{R})^{[2]} &\rightrightarrows S^2 \times SU(2) \times \mathbb{R} \\
&\downarrow {\scriptstyle (\mathsf{s},\mathsf{t})} \\
(S^2 \times SU(2))^{[2]} &\rightrightarrows S^2 \times SU(2) \\
&\downarrow {\scriptstyle (\mathsf{s},\mathsf{t})} \\
(S^2)^2 &\rightrightarrows S^2
\end{aligned}
\tag{83}
$$

where $(S^2 \times SU(2))^{[2]} := (S^2 \times SU(2)) \times_{(S^2)^2}^{(\mathsf{s},\mathsf{t}),(\mathsf{s},\mathsf{t})} (S^2 \times SU(2)) \cong S^2 \times SU(2) \times U(1)$ and $(S^2 \times SU(2) \times \mathbb{R})^{[2]} := (S^2 \times SU(2) \times \mathbb{R}) \times_{S^2 \times SU(2) \times U(1)}^{(\mathsf{s},\mathsf{t}),(\mathsf{s},\mathsf{t})} (S^2 \times SU(2) \times \mathbb{R}) \cong S^2 \times SU(2) \times \mathbb{R} \times \mathbb{Z}$.

6. The structure (51) is captured by the strict 2-groupoid $\bar{\mathcal{P}}_2 S^3 \times U(1)/WZW \rightrightarrows \bar{\mathcal{P}}S^3 \rightrightarrows S^3$. (Here we have identified paths related by thin homotopy, while in (51) we did not; this does not matter because our purpose is to capture the WZW evaluation, which is indeed unaffected by any thin homotopy.) Including the Villainzation layer in (48) above (51), the structure is captured by the strict 3-groupoid $(\bar{\mathcal{P}}_2 S^3 \times U(1)/WZW) \times \mathbb{R} \rightrightarrows \bar{\mathcal{P}}_2 S^3 \times U(1)/WZW \rightrightarrows \bar{\mathcal{P}}S^3 \rightrightarrows S^3$. As mentioned there, the problem of using this structure for a lattice theory is that $\bar{\mathcal{P}}S^3$ is infinite dimensional. Our task is to find a finite dimensional 3-category in Section 5.5 which is equivalent to this infinite dimensional strict 3-category in a suitable sense. Understanding such "equivalence in a suitable sense" is why higher category theory is necessary; otherwise, without category theory, it is hard to move beyond (51).

So far we have described the general structure of strict higher categories. But more interesting is the relation between structures.

Given two 0-categories, i.e. sets, we would think about functions mapping between them, $D \xleftarrow{F} C$. Just from this notation, we realize a deep, interesting point, that all 0-categories together form a 1-category Set, or say $\mathsf{0Cat}$, where the objects in Set_0 are sets, and the morphisms in Set_1 are functions between sets. [81] This point of view is not only important purely mathematically, but is directly useful for the concept of *internalization* in Section 5.2, which will in turn underlie our construction of lattice d.o.f..

---

[81] The issue mentioned in footnote 70 appears here. To equate "all 0-categories" to "all sets", we should really mean "all small 0-categories". The same is understood in further discussions below. (Set itself is a large 1-category because the collection of all sets is not a set but a proper class. If we want—though often there is no intrinsic problem to work with large categories—we can always further restrict "all sets" to sets whose cardinalities are not too large, so that the collection of them is still a set, and the collection thus becomes a small 1-category. But all these should not matter in physics, because we do not expect sets with indefnitely large cardinalities to be directly involved in physics anyways.)

It is then natural to ask what maps between two 1-categories. The notion of *functor* naturally comes up (although for our application we will need a more general notion of functor, i.e. *anafunctor*, which we will explain in Section 5.2): A functor F from 1-category C to 1-category D, again denoted as $D \xleftarrow{F} C$, involves a function F_0 from C_0 to D_0 and a function F_1 form C_1 to D_1, pictorially

$$(84)$$

such that the source and target maps, the composition, and the identity specifications are all preserved. [82] (In the special case when C and D are BG and BH, a functor between BG and BH apparently is a group homomorphism between G and H.) Similar to functions between 0-categories, functors between 1-categories can be composed in the obvious manner, $(E \xleftarrow{G \circ F} C) := (E \xleftarrow{G} D \xleftarrow{F} C)$, and the composition is associative.

A fundamental reason that makes the notion of 1-category more powerful than the notion of set (0-category) is, now that we have two layers, there is a new kind of relation from C to D that has no counter-part in 0-categories: We can also consider a map from C_0 to D_1. But then what would C_1 map to? Recall we may view a 1-category as a special kind of 2-category which only has identity 2-morphisms, i.e. $D_2 = \{\mathbf{1}_h | h \in D_1\} \cong D_1$, therefore C_1 must somehow map to this D_2. This leads to the notion of *natural transformation*. We can think of a natural transformation Φ pictorially as

$$(85)$$

where the top and bottom surfaces reduce to two functors F, G mapping from C to D, and there is a function Φ_1 mapping from C_0 to D_1 such that it reduces to F_0 and G_0 when taking the source and target in D. Moreover there is a Φ_2 mapping from C_1 to D_2, where D_2 only contains identity 2-morphisms. More exactly, $f \in C_1$ is mapped to the rectangular shape on the left, which should represent a 2-morphism in D_2, and since the only available 2-morphisms in a 1-category are identity 2-morphisms, we conclude the only possibility is

---

[82]It is common to abbreviate both F_0 and F_1 as just F, but keeping the subscript in mind is helpful for generalizing towards the crucial notion of *anafunctor* in Section 5.2.

$\Phi_2(f) = \mathbf{1}_{\Phi_1(c') \circ F_1(f)} = \mathbf{1}_{G_1(f) \circ \Phi_1(c)} \in D_2$, which in turn implies $\Phi_1(c') \circ F_1(f) = G_1(f) \circ \Phi_1(c)$. Therefore, Φ_2 does not contain any more information than what is already contained in F_1, G_1, Φ_1, rather it provides a consistency constraint between these three functions. Such a Φ is said to be a natural transformation from functor F to functor G. Thus, apparently we should denote a natural transformation as

$$D \overset{F}{\underset{G}{\rightleftarrows}} \Downarrow \Phi \quad C$$

.

The vertical composition of natural transformations is obvious; with a little extra effort horizontal composition can be defined, too (called Godemant product). From here, we see all 1-categories together form a strict 2-category Cat, or say $\mathsf{1Cat}$.

A natural transformation Φ^{-1} from G to F is the inverse (under vertical composition) of Φ if $(\Phi^{-1})_1 = (\Phi_1)^{-1}$—and this may or may not exist for a given Φ. Just like how the equivalence (equipotence) between two sets is established by the existence of an invertible function between them, we can say two functors are equivalent if there is an invertible natural transformation (also called *natural isomorphism*) between them, though the two functors may not be equal.

With this notion of equivalence between functors, now we can define the notion of "inverse" for a functor at two levels of strictness. Intuitively we can define *the* inverse $C \overset{F^{-1}}{\longleftarrow} D$ of $D \overset{F}{\leftarrow} C$ by strictly requiring $C \overset{F^{-1} \circ F}{\longleftarrow} C = \mathbf{1}_C$ and $D \overset{F \circ F^{-1}}{\longleftarrow} D = \mathbf{1}_D$. If two categories are related by invertible functors in such a strict sense, the two categories are strictly isomorphic at each level (we will colloquially say they are the same). However, often a less strict notion is more useful, especially when the strict inverse does not exist. We say a functor $C \overset{\bar{F}}{\leftarrow} D$ is *an* inverse of a functor $D \overset{F}{\leftarrow} C$, if the composed functor $C \overset{\bar{F} \circ F}{\longleftarrow} C$ has an invertible natural transformation to $\mathbf{1}_C$, and $D \overset{F \circ \bar{F}}{\longleftarrow} D$ also has an invertible natural transformation to $\mathbf{1}_D$. We say the existence of such pair $F, \bar{F}$ establishes a *natural equivalence* between the 1-categories C and D.

This is the first scenario where the flexibility of category theory manifests—and we will need more kinds of flexibility later in order to arrive at the lattice construction we desire. It can be seen that the definition of natural equivalence between 1-categories looks remarkably similar to the definition of homotopy equivalence between topological spaces, whose contrast with the strict notion of homeomorphism shows the power of flexibility. Indeed, a homotopy between two manifolds induces a natural equivalence between their fundamental groupoids.

It is easy to prove that an equivalent—but often more useful in practice—way to state natural equivalence between C and D is to say F is "essentially surjective and fully faithful". "Essentially surjective" means while $D_0 \overset{F_0}{\leftarrow} C_0$ might not be surjective, any $d \in D_0$ must be related via some invertible morphism to (in generalization of being strictly equal to) some $F_0(c)$. "Fully faithful" means for any $a, b \in C_0$, the restriction of F_1 to $C_1|_{b,a}$ is a bijection between $C_1|_{b,a}$ and $D_1|_{F_0(b),F_0(a)}$. From these conditions it is not hard to construct an inverse functor $\bar{F}$ that is also essentially surjective and fully faithful. [83] Thus, the map between two

---

[83]An important caveat is that one needs a "choice function" to define $\bar{F}_0$ for the essentially surjective

naturally equivalent 1-categories is still bijective in the traditional sense at the morphism layer given the source and the target, but becomes more flexible at the object layer.

Let us discuss a simple example of natural equivalence relevant to lattice QFT. Consider two 1d lattice loops, but with different numbers of vertices. We feel they should be equivalent in some suitable sense, since they both discretized a 1d space(time) circle. Indeed, as 1-categories they are naturally equivalent, and both naturally equivalent to a lattice loop with a single vertex, i.e. $B\mathbb{Z}$—and this $\mathbb{Z}$ in the 1-morphism captures the π_1 of a loop. Readily from here, we can feel that natural equivalence is related to the invariant information under renormalization (coarse graining of lattice), and the notions of "same" versus "naturally equivalent" are, roughly speaking, respectively suitable for discussing UV versus IR. We will see more and more of such intuition in the proceeding.

With this example, we can introduce the concept of *skeletal* category, which means in such a category, if two objects are related by an invertible morphism, then these two objects must be the same object. Starting with a generic category, we can arrive at a skeletal category naturally equivalent to the original category, by identifying objects that are related by invertible morphisms. We will often use a skeletal category to represent its natural equivalence class, calling it the *skeleton* of the class. In the example above, $B\mathbb{Z}$ is the skeleton.

Now let us try to generalize the notions of functor and natural transformation for strict 2-categories. It is not hard to see we can still define functors, natural transformation between functors, and moreover we can define a new kind of relation called *modification* between natural transformations, which maps C_0 to D_2 (and C_1, C_2 to D_3, D_4 which contain only identity 3- and 4-morphisms). We will not delve into modification. But now, even for functors and natural transformations, there arise the possibility of having definitions at different levels of strictness. In the below we will discuss these different levels of strictness and see how they arise in the familiar lattice theories.

A strict 2-functor F is such that it has functions F_k $(k = 0, 1, 2)$ that map C_k to D_k and strictly preserve all the source, target, composition and identity maps. A strict 2-natural transformation is basically the same as a natural transformation for 1-category, i.e. Φ_2 still maps C_1 to the subset of identity morphisms $\{\mathbf{1}_h | h \in D_1\} \subseteq D_2$.

But even between two strict 2-functors, strict 2-natural transformations are not the only option. We can consider the more general notion of *lax 2-natural transformation*, where Φ_2 can map C_1 to D_2 in the generic way, i.e. $\Phi_2(f) \in D_2$ (the rectangular surface on the left of (85)) does not have to be any $\mathbf{1}_h \in D_2|_{h,h} \subset D_2$, and there are consistency constraints, whose details we will omit, provided by Φ_3 that maps C_2 to D_3 which contains only identity 3-morphisms. (It is sometimes desired to require $\Phi_2(f)$ to be an invertible 2-morphism which is not necessarily an identity 2-morphism, and such a "slightly stricter" version of lax 2-natural transformation is called a *pseudo 2-natural transformation*.)

And more general than strict 2-functors, there are *lax 2-functors*, where the composition of 1-morphisms and the assignment of identity 1-morphisms do not have to be preserved

---

F_0. The choice function will lead to discontinuity issue when we work with topological spaces, as we will discuss in the next subsection. This caveat makes the use of anafunctor necessary in many situations, in generalization of ordinary functor.

strictly, but only up to some 2-morphisms, i.e. we can specify $F_1(g \circ f) \xLeftarrow{\varphi_{g,f}} F_1(g) \circ F_1(f)$, $\mathbf{1}_{F_0(a)} \xLeftarrow{\psi_a} F_1(\mathbf{1}_a)$ in generalization of having equalities in the middle, and these 2-morphisms must be chosen to satisfy certain consistency constraints—whose details we will omit, but they finally come from the fact that D_3 contains identity 3-morphisms only. (Again, it is sometimes desired to require $\varphi_{g,f}$ and ψ_a to be invertible 2-morphisms, and such a 2-functor is called *pseudo 2-functor*.) The definition of lax 2-natural transformation between two lax 2-functors also requires some changes compared to when the 2-functors are strict, though the spirit is the same.

And thus we can envision that, more generally, for strict n-categories, there are (n, q)-*transfors* between two $(n, q-1)$-transfors, which map C_k to D_{k+q} (so $q = 0$ are n-functors and $q = 1$ are n-natural transformations), such that the maps for $k + q \leq n$ contains information that defines the transformation, and the map for $k + q = n + 1$ provides consistency constraints. The transfors can be defined at different levels of strictness. The collection of strict n-categories along with their strict (n, q)-transfors $(0 \leq q \leq n)$ form a strict $(n + 1)$-category, but often it is desired to include laxer transfors, which will in general result in a less strict $(n + 1)$-category. Plunging more deeply into this is beyond our scope.

Now it can be readily seen how functors and natural transformations are useful in lattice QFT:

1. In traditional lattice nlσm, a field configuration is a function from $\bar{\mathcal{L}}_0$ to $\mathcal{T}$.

2. In traditional lattice gauge theory, a field configuration is a functor from $\bar{\mathcal{L}}_{\leq 1}$ to BG. A gauge transformation is a natural transformation, which is invertible since all morphisms are invertible in BG. Hence field configurations that are related by gauge transformations are indeed equivalent as functors.

3. In Villainized S^1 nlσm, a field configuration is a functor from $\bar{\mathcal{L}}_{\leq 1}$ to the action groupoid $S^1 \times \mathbb{R} \rightrightarrows S^1$. [84] More generally, a field configuration in a Villainized nlσm is a functor from $\bar{\mathcal{L}}_{\leq 1}$ to $\widetilde{\mathcal{T}}^2/\Gamma \rightrightarrows \mathcal{T}$, where $\widetilde{\mathcal{T}}$, the universal cover of $\mathcal{T}$, is a Γ bundle over $\mathcal{T}$ for some discrete group Γ, and the mod Γ is by a Γ action on both $\widetilde{\mathcal{T}}$'s.

With this perspective we can systematically understand what it means for a lattice path integral to be local, especially in situations like Villainization (recall the discussion we had at the end of Section 2.1). Locality just means each field configuration sampled in the path integral is a functor from the lattice to some target category (possibly a higher category, which we will discuss later)—in generalization of the usual notion of target space—so that each vertex is mapped to some field valued in C_0, each link is mapped to some field valued in C_1, and so on. But the path integral is in general not locally factorizable, in the sense that C_1 in general cannot be factorized into the form $C_0 \times C_0 \times X$—either not in this form as a

---

[84]What we called a $\mathbb{Z}$ gauge transformation in Section 2.1 when viewing Villainization as gauging a $\mathbb{Z}$ symmetry from an $\mathbb{R}$-valued theory is *not* a natural transformation. In fact it does not act on this description, because $e^{i\theta} \in S^1$ and $\gamma \in \mathbb{R}$ are already physically meaningful variables. The categorical nature of the $\mathbb{Z}$ gauge transformation will be explained below (91) after we introduce anafunctor.

set, or not in this form as a manifold though as a set—and likewise for higher morphisms. However, C_1 does have the source and target maps to C_0, which can be viewed as local constraints (e.g. the $e^{i\gamma_l} = e^{id\theta_l}$ constraint in the Villain model). In practice, when sampling the fields, we parametrize C_1 by $C_0' \times C_0' \times X$ using some large enough C_0' and X (e.g. we write $\gamma_l = d\theta_l + 2\pi m_l \in \mathbb{R}$ in the Villain model, with $\theta_v, \theta_{v'} \in (-\pi, \pi]$ and $m_l \in \mathbb{Z}$).

While these descriptions above seem nice and systematic, they are not entirely satisfactory. Let us take the traditional nlσm case as example. Looking only at $\bar{\mathcal{L}}_0$ means we ignore which vertices are connected by links and which ones are not, but the path integral weight should be associated with links and cares about the difference of the fields between the two ends of each link. So it is desirable to be able to talk about the lattice $\bar{\mathcal{L}}$ entirely, instead of truncating it to, say, $\bar{\mathcal{L}}_0$. Can we say a field configuration in traditional lattice nlσm is a functor from $\bar{\mathcal{L}}$ to $\mathcal{T}$, instead of from $\bar{\mathcal{L}}_0$ to $\mathcal{T}$? It turns out the statement becomes incorrect, because $\bar{\mathcal{L}}_1$ contains all lattice paths, meanwhile, since $\mathcal{T}$ is a 0-category, $\mathcal{T}_1 = \{\mathbf{1}_a | a \in \mathcal{T}\}$ only contains identity 1-morphisms, thus a functor from $\bar{\mathcal{L}}$ to $\mathcal{T}$ must have a constant field over each connected component of the lattice, which is certainly not what we want in general.

The correct statement is:

1. In traditional lattice nlσm, a field configuration is a functor from $\bar{\mathcal{L}}$ to the *pair groupoid* $E\mathcal{T} := (\mathcal{T}^2 \rightrightarrows \mathcal{T})$ in which from any object to any other object there is exactly one morphism, $E\mathcal{T}_1|_{b,a} = \{(b,a)\}$. Physically, this just means an almost trivial fact in traditional lattice nlσm: the d.o.f. associated with a link $l = \langle v'v \rangle$ is just specified by the fields on the two vertices v and v' together, no more and no less.

It is easy to see any pair groupoid $E\mathcal{T}$ is naturally equivalent to the trivial category $*$ for arbitrary $\mathcal{T}$, because any functor between $E\mathcal{T}$ and $*$ are essentially surjective and fully faithful. (If $\mathcal{T}$ is furthermore a group G, then similar to the relation between the category BG and the classifying space $|BG|$, the category EG is also related to the universal bundle $|EG|$ via the procedure of geometric realization that we will introduce in Section 5.4. Just like the space $|EG|$ is a G bundle over $|BG|$, in a suitable sense the category EG is also a G bundle over BG. [85] And the fact that EG is naturally equivalent to the trivial category is related to the fact that $|EG|$ is contractible. More generally, the space $|E\mathcal{T}|$ can be constructed in the same way and is also contractible, although there is no $B\mathcal{T}$ when $\mathcal{T}$ does not have a group structure.) Does this natural equivalence between $E\mathcal{T}$ and $*$ mean the traditional lattice nlσm is a trivial theory? Certainly not—the theory is non-trivial because of the presence of the path integral weight. This is the crucial distinction between a generic lattice theory and a topological lattice theory that worths some detailed explanation.

A topological lattice theory (e.g. [19, 20, 23, 32, 33, 41], which should be viewed as an

---

[85]This means we have a functor from the 0-category G to the 1-category EG and then a functor from EG to BG, such that any 1-morphism $\mathbf{1}_{\tilde{g}} \in G_1$ is mapped to $(\tilde{g}, \tilde{g}) \in EG_1 = G^2$, which is in turn mapped to the identity $\tilde{g}\tilde{g}^{-1} = \mathbf{1} \in BG$. Alternatively, the pull-back category (which we did not systematically define) $EG \times_{BG} EG := (\{((g_1, g_2), (g_1', g_2'))\} | g_2 g_1^{-1} = g_2' g_1'^{-1}\} \rightrightarrows \{(g, g')\})$ has a functor to the 0-category G, given by $g'^{-1}g = \tilde{g} \in G_0 = G$, and consistently, $g_2'^{-1}g_2 = g_1'^{-1}g_1 = \tilde{g}$ for $\mathbf{1}_{\tilde{g}} \in G_1 \cong G$, specifying the G action on EG, just like $|EG| \times_{|BG|} |EG|$ has a function to G, specifying the G action on $|EG|$. The mathematical idea and physical interpretation behind this will become clearer as we proceed (in particular as discussed below (111)).

effective theory already coarse grained to the deep IR limit) has a key feature that the path integral weight can only take value 0 or a complex phase of magnitude 1, and is invariant under natural transformations of field configurations. Thus, for a topological theory, the target category being natural equivalent to a trivial category means the theory is necessarily trivial. By contrast, it is familiar that a dynamical traditional lattice nlσm can explore at least two phases by tuning the path integral weight, the trivial (disordered) phase and the spontaneous symmetry breaking (ordered) phase. Consider two topological limits, and the more generic and more physical situation in-between:

- If the link weight is 1 for any field configuration, then the path integral is sampling the functors from $\bar{\mathcal{L}}$ to $E\mathcal{T}$ freely, such that the weight is invariant under any natural transformation. This is the extreme case of the disordered (trivial) phase, as if we have replaced $E\mathcal{T}$ by its ananaturally equivalent skeleton, the trivial category $*$.

- If the link weight is a delta function, then only identity 1-morphisms are kept, in which case the target category becomes $\mathcal{T}$, a non-trivial subcategory of $E\mathcal{T}$. A functor from $\bar{\mathcal{L}}$ to $\mathcal{T}$ is indeed a completely ordered configuration, where each connected component of the lattice has a constant field. This is the extreme case of the ordered phase.

- For a traditional (dynamical) lattice nlσm, a generic link weight lies in-between these two topological limits—the weight does not respect invariance under natural transformation of the field configuration, but it has not gone as far as to reduce the target category from $E\mathcal{T}$ to $\mathcal{T}$. The phase transition between the ordered and disordered is not determined at the lattice level but only in the IR.

For this (physically very intuitive) reason, in dynamical lattice theory, we *must not* conclude the phase of the theory based on the natural equivalence class of the target category. We will have more thorough discussions of this in Sections 5.2 and 5.5.

After these explanations, we are ready to see how the known examples of lattice theories in Section 2 are described by strict higher categories along with functors and natural transformations (since the strict higher categories involved are strict higher groupoids where all morphisms are strictly invertible, hence all "lax" below are automatically "pseudo"):

1. In traditional lattice nlσm, a field configuration is a functor from $\bar{\mathcal{L}}$ to the pair groupoid $E\mathcal{T} := (\mathcal{T}^2 \rightrightarrows \mathcal{T})$, which means the field on a link $l = \langle v'v \rangle$ is just specified by the fields on v and v' together.

   A generic natural transformation is going to change the relative values of the fields across a link (since $\Phi_1(v \in \bar{\mathcal{L}}_0)$ can be any element in $E\mathcal{T}_1$), therefore physically we do not demand the path integral weight to be invariant under natural transformation.

2. In traditional lattice gauge theory, a field configuration is a functor from $\bar{\mathcal{L}}$ to $BEG := (G^2 \rightrightarrows G \rightrightarrows *)$ (the delooping of EG), which means the field on the plaquette bounded by two Wilson lines is just specified by the two Wilson lines together, or equivalently, it can be specified by one Wilson line along with the holonomy.

   If G is a discrete group (as is often the case in the effective theory of topological phase, which we will discuss more in Section 5.3), then it is physically possible to forbid the

gauge flux, in which case only identity 2-morphisms are left, so that the target category becomes just BG. But for continuous group it is not so physical to demand so.

A gauge transformation is a strict 2-natural transformation. The holonomy around a plaquette or a non-contractible loop remains invariant (up to conjugation by Wilson lines) because the image of Φ_2 in a strict 2-natural transformation only contains identity 2-morphisms. On the other hand, a generic lax 2-natural transformation changes the holonomy. Therefore, physically we demand the path integral weight to be invariant under strict 2-natural transformation, but not under generic lax 2-natural transformation.

3. In Villainized S^1 nlσm, a field configuration is a functor from $\bar{\mathcal{L}}$ to $S^1 \times \mathbb{R} \times \mathbb{Z} \rightrightarrows S^1 \times \mathbb{R} \rightrightarrows S^1$, where the 1-morphisms the $\mathbb{Z}$ in the space of 2-morphism represents the vorticity; in particular, it comes from $(S^1 \times \mathbb{R})^{[2]} := (S^1 \times \mathbb{R}) \times_{(S^1)^2}^{(s,t),(s,t)} (S^1 \times \mathbb{R}) \cong S^1 \times \mathbb{R} \times \mathbb{Z}$. If vortices are forbidden, i.e. only identity 2-morphisms are allowed, then the target category can be reduced to the action groupoid $S^1 \times \mathbb{R} \rightrightarrows S^1$. More generally, a field configuration in a Villainized nlσm is a functor from $\bar{\mathcal{L}}$ to $\widetilde{\mathcal{T}}^2/\Gamma \times \Gamma \rightrightarrows \widetilde{\mathcal{T}}^2/\Gamma \rightrightarrows \mathcal{T}$ (note that $\mathbb{R}^2/\mathbb{Z} \cong S^1 \times \mathbb{R}$); if the Γ vortices are forbidden, only the identity 2-morphisms are left.

Again, in a nlσm, physically we do not demand the path integral weight to be invariant under 2-natural transformation.

4. In Villainized $U(1)$ gauge theory, a field configuration is a functor from $\bar{\mathcal{L}}$ to the delooping of the target category above, $U(1) \times \mathbb{R} \times \mathbb{Z} \rightrightarrows U(1) \times \mathbb{R} \rightrightarrows U(1) \rightrightarrows *$, where the $\mathbb{Z}$ in the 3-morphism represents the monopole. If monoples are forbidden, i.e. only identity 3-morphisms are allowed, then the target category can be reduced to the 2-group $U(1) \times \mathbb{R} \rightrightarrows U(1) \rightrightarrows *$. More generally, a field configuration in a Villainized gauge theory is similar, as long as we replace $U(1) \times \mathbb{R}$ by $G \ltimes H$, and $\mathbb{Z}$ by $H/G = \ker(\widetilde{t})$ which must be abelian as explained before.

Now it becomes particularly interesting to ask if the path integral weight should be invariant under natural transformations at some certain level of strictness.

As a Villainized gauge theory, the path integral weight should be invariant under strict 3-natural transformations only, i.e. those where Φ_{k+1} maps C_k to identity $(k+1)$-morphisms in D_{k+1} for $k > 0$, and to generic 1-morphisms for $k = 0$. This are the usual gauge transformations on the lattice.

One can ask what if we impose a stronger requirement that the path integral weight be invariant under 3-natural transformations that are less strict? In particular, let us consider invariance under those laxer 3-natural transformations where Φ_{k+1} maps C_k to identity $(k+1)$-morphisms in D_{k+1} for $k > 1$ and generic $(k+1)$-morphisms for $k = 0, 1$. This is what is called *2-group gauge theory* that has been studied relatively early on as an application of category theory in physics [32–34, 44–46]. In this case the flux is no longer gauge invariant up to conjugation, but the $\ker(\widetilde{t})$-valued monopole in the 3-morphism is still physically well-defined. We will discuss more about this in Section 5.3.

From here, we can see that in general, even for the same target category, we can still demand invariance of the path integral weight under natural transformations of different levels of strictness. As usual, by tuning the path integral weight, we may access different phases of a theory; as we demand the path integral weight to remain invariant under laxer and laxer natural transformations, the accessible phases of a theory becomes more and more limited. This is why the Villainized $U(1)$ gauge theory can access both the confined and the Coulomb phases (for $d \geq 4$ [62]), while the 2-group gauge theory with the same target category only represents the confined phase [32–34].

5. Obviously, when both G and H in the target category above are abelian, we can deloop the category arbitrarily many times, and obtain Villainized higher form gauge theories.

6. In spinon decomposed S^2 nlσm, a field configuration is a functor from $\bar{\mathcal{L}}$ to $S^2 \times SU(2) \times \mathbb{R} \times \mathbb{Z} \rightrightarrows S^2 \times SU(2) \times \mathbb{R} \rightrightarrows S^2 \times SU(2) \rightrightarrows S^2$, where the $\mathbb{Z}$ in the 3-morphism represents the hedgehog (see (83)). If hedgehogs are forbidden, i.e. only identity 3-morphisms are allowed, then the target category reduces to $S^2 \times SU(2) \times \mathbb{R} \rightrightarrows S^2 \times SU(2) \rightrightarrows S^2$.

   Again, in a nlσm, physically we do not demand the path integral weight to be invariant under 3-natural transformation.

7. Consider two smooth functions f, g from manifold $\mathcal{M}$ to manifold $\mathcal{N}$. Function f determines a d-functor F from the strict path d-groupoid $\bar{P}_d \mathcal{M}$ to the strict path d-groupoid $\bar{P}_d \mathcal{N}$ (where d is the max of the dimensions of $\mathcal{M}, \mathcal{N}$), because knowing how every point on $\mathcal{M}$ maps to $\mathcal{N}$ determine how every path, surface and so on on $\mathcal{M}$ maps to that on $\mathcal{N}$. Likewise for g. A homotopy from f to g determines a lax d-natural transformation from F to G. Homotopy equivalence between $\mathcal{M}$ and $\mathcal{N}$ implies natural equivalence between $\bar{P}_d \mathcal{M}$ and $\bar{P}_d \mathcal{N}$. [86]

8. $\bar{\mathcal{L}}$ and $\bar{\mathcal{L}}'$ for two lattices that discretize the same space or two homotopically equivalent spaces are naturally equivalent, where the natural equivalence is again established by lax d-natural transformations.

From these discussions we can experience that category theory is a natural language for organizing our thoughts about lattice QFT and potentially their relation to continuum QFT. To describe the known lattice QFTs in Section 2, we only used strict higher categories; so we may indeed anticipate that the generalization problem discussed in Section 3 might find its solution when the more flexible higher categories are taken into consideration.

To go towards this direction, next we shall motivate the introduction of *anafunctors* as a necessary (and actually familiar and intuitive, as we shall see) generalization of the ordinary functors, whenever we are concerned with the continuity/smoothness of spaces and functions—which is indeed the very problem our work aims at.

---

[86]But the converse is not necessarily true, because any strict groupoids cannot capture the full homotopy information of the manifold. See footnote 98. To establish the converse, weak higher groupoids are needed, see Section 5.4 for one construction.

5.2 Internalization and anafunctor

Let us start with a motivating problem. In the above we have seen that the homotopy between two manifolds implies natural equivalence of their strict path d-groupoids; the same holds when both manifolds are discretized into lattices. But an obvious question to ask is: Consider the strict d-groupoid of a lattice and the strict path d-groupoid of the manifold that the lattice is discretizing, are they also naturally equivalent in some suitable sense?

The subtlety here lies in that the manifold is not only a set of points, but has the extra structure of being smooth. So, as mentioned before, it is intuitive to require whatever maps that are involved to be smooth maps. This intuition can be more systematically phrased in terms of *internalization*. Consider all the sets and functions involved in defining a category $C = (C_1 \rightrightarrows C_0)$,

$$
C_1 \times_{C_0}^{\mathsf{s},\mathsf{t}} C_1 \xrightarrow{\ \circ\ } C_1 \underset{t}{\overset{s}{\rightleftarrows}} \xleftarrow{i} C_0
\tag{86}
$$

where the functions satisfy some consistency constraints (such as $\mathsf{si} = \mathsf{ti} = \mathbf{1}_{C_0}$, associativity, etc.). While this diagram represents a category C, if we stare at this diagram, we realize it also represents a few objects and a few morphisms within some larger category—and this "larger category" is Set, because C_0, C_1 and $C_1 \times_{C_0}^{\mathsf{s},\mathsf{t}} C_1$ are indeed sets, and $\mathsf{s}, \mathsf{t}, \mathsf{i}$ and $\circ$ are indeed functions. So this diagram, along with the diagrams that describe the consistency constraints (which are straightforward to draw, and we omitted here), define a category C by picking certain objects and certain morphisms from the 1-category Set of sets. We say C is "a category *internal to* the ambient category Set"—which is what we often mean by default when we say "a category". [87]

With this perspective in mind, it is easy to generalize to 1-categories internal to other ambient 1-categories. For example, let the ambient 1-category be $\mathsf{Manifold}$ instead, where the objects are finite dimensional smooth manifolds, and the morphisms are smooth maps between them (so $\mathsf{Manifold}$ is a subcategory of Set). Then we can pick some objects and morphisms from $\mathsf{Manifold}$ to form the same diagram as above, satisfying the same consistency constraints (which are also presented as diagrams), and this will define a 1-category internal to $\mathsf{Manifold}$, which means C_0, C_1 and $C_1 \times_{C_0}^{\mathsf{s},\mathsf{t}} C_1$ are smooth manifolds, and all maps involved are smooth maps. [88] This is the systematic description of the intuition before. Likewise we can define the internalization of higher categories or other structures in $\mathsf{Manifold}$. A familiar example is a group internal to $\mathsf{Manifold}$, which is, apparently, a Lie group; similarly, a groupoid internal to $\mathsf{Manifold}$ is a Lie groupoid. [89]

---

[87]More precisely, C defined by such a diagram in Set must be a small category. So the answer to the self-referencing question of whether Set can be defined by such a diagram within Set is "No".

[88]s, t must be surjective submersions to ensure that, by transversality, $C_1 \times_{C_0}^{\mathsf{s},\mathsf{t}} C_1$ and $C_1 \times_{C_0}^{\mathsf{s},\mathsf{t}} C_1 \times_{C_0}^{\mathsf{s},\mathsf{t}} C_1$ also exist as smooth manifolds (the latter space is for describing the consistency condition of associativity).

[89]In the above we drew the diagram that defines a category. To draw the diagram for a groupoid, we have an additional arrow from C_1 to C_1 (satisfying suitable constraints) that assigns inverses. Further, to define a group, we require C_0 to be the manifold with a single point—if we break away from the set theoretic language, such a "single point manifold" can be described as being a *terminal object* in $\mathsf{Manifold}$, i.e. it is

We can also consider more general ambient categories, as long as products of the form $X \times_Z^{u,v} Y$ are defined in the ambient category. [90] [91] In particular, when discussing the relation between lattice and continuum, we will often need the spaces of paths, surfaces and so on in a manifold (such as in defining the strict path d-groupoid), and these spaces are infinite dimensional. Therefore we will need a notion of "smoothness" for infinite dimensional spaces. A notion suitable for our usage would be "diffeological", whose detailed definition we will not get into (see e.g. [90]) since we are not aiming at a comprehensive and rigorous mathematical exposition in this work. The category Difflg with diffeological spaces as objects and diffeological maps as morphisms will often be used as the ambient category, generalizing Manifold by including the infinite dimensional cases. In the below, we may colloquially use the familiar word "smooth" to mean diffeological when the space involved is infinite dimensional.

The problem we face is, the definitions of functor and natural transformation introduced in Section 5.1 are designed for categories internal to Set, but when applied to categories internal to Manifold or Difflg—as we do—i.e. when requiring the maps involved in the definitions of functor and natural transformation to be smooth, the definitions would become too restrictive to capture many interesting situations. So we must generalize the definitions.

Let us consider the simplest example in our motivating problem: Is $\bar{\mathcal{L}}$ for a 1d lattice loop (in the extreme case where the lattice loop has only one vertex and one link, we get $\bar{\mathcal{L}} = B\mathbb{Z}$, the skeleton) naturally equivalent to the path groupoid $\bar{P}_1 S^1 = (S^1 \times \mathbb{R} \rightrightarrows S^1)$ of the circle that it is discretizing? Indeed, we can have an essentially surjective and fully faithful functor from $\bar{\mathcal{L}}$ to $\bar{P}_1 S^1$, for instance
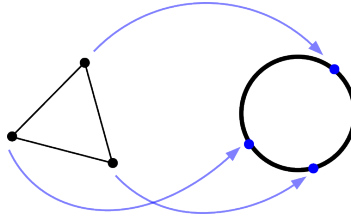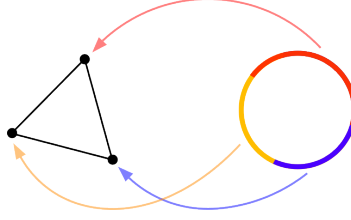

where we indicated how the lattice vertices map to points on the circle, and then the links are mapped to paths on the circle in the obvious way. Conversely, there must be an essentially surjective and fully faithful functor from $\bar{P}_1 S^1$ to $\bar{\mathcal{L}}$ that is an inverse of the functor above. One such inverse functor is

an object such that all objects in Manifold has a unique morphism to it (so it is easy to see that a terminal object is unique up to unique invertible morphisms).

[90]This is to require the ambient category to admit finite limits. Limit is a crucial general concept in category theory which we, nevertheless, did not introduce. One may consult relevant texts on this.

[91]The ambient category can also be a higher category. In that case, some equality signs in the consistency constraints can be replaced by invertible 2- or higher morphisms in the ambient category (where "invertible" itself may also be defined in a weak sense). Some of our discussions below can be phrased in this language, for example multiplicative bundle gerbe crucial to our main construction can be described as 2-group internalized in the bicategory of Lie groupoids [35]. But we will not go deeply into the details.

where we indicated how the points on the circle map to the lattice vertices, and then the paths on the circle are mapped to the paths on the lattice depending on the starting and ending point and the winding, in the intuitive way. This is all good if the categories are internal to Set, but now that we want them to be internal to Manifold, [92] there is a problem: The inverse functor obviously involves discontinuous functions.

A familiar treatment allows us to avoid such discontinuity, and will lead us towards the definition of anafunctor soon. Instead of thinking about the circle itself, we cover the circle with some patches (open charts) U_α, and map each patch to a lattice vertex.
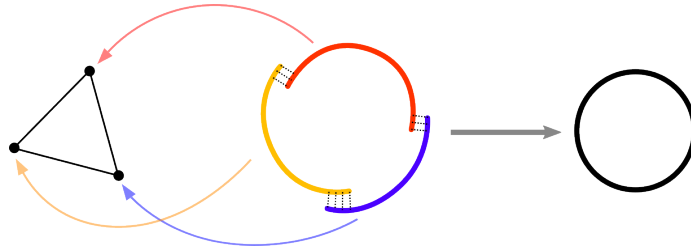
More particularly, given the patches we can form a category F, where $F_0 = \sqcup_\alpha U_\alpha$, the disjoint union of the patches, and F_1 contains two kinds of basic morphisms: one is the paths within each patch, and the other is the identification morphisms, specifying which points on different patches will be identified when mapped to $\mathcal{M}$ (denoting the map as $\sqcup_\alpha U_\alpha \xrightarrow{\Pi} \mathcal{M}$), i.e. there is a morphism from $x \in U_\alpha$ to $y \in U_\beta$ whenever $\Pi(x) = \Pi(y) \in S^1$; moreover, these two kinds of basic morphisms can be composed, up to the intuitive identification
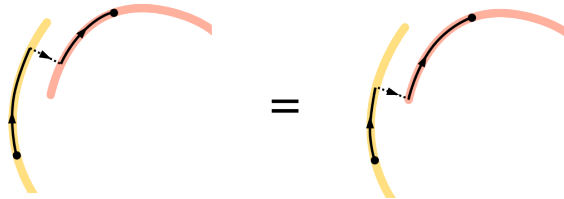
Such a category F has a surjective (rather than just essentially surjective) and fully faithful functor to each of $\bar{P}_1\mathcal{M}$ and $\bar{\mathcal{L}}$, and moreover all maps involved are smooth. In the below, we will extract the essence behind this familiar treatment to define the notions of *anafunctor*, *ananatural transformation*, and *ananatural equivalence*. The example here will turn out to be an ananatural equivalence between the lattice $\bar{\mathcal{L}}$ and the continuum $\bar{P}_1\mathcal{M}$, established by an invertible anafunctor F.

Before giving the precise definitions, it is helpful to look at another motivating example. Let us consider a manifold $\mathcal{M}$ with identity morphisms only, $\mathcal{M} \rightrightarrows \mathcal{M}$. What are the possible

---

[92] The lattice $\bar{\mathcal{L}}_0$ and $\bar{\mathcal{L}}_1$ are discrete, but discrete topology is a special case of topology.

functors from $\mathcal{M} \rightrightarrows \mathcal{M}$ to $BG = (G \rightrightarrows *)$? Somehow we feel there should be different possibilities, to do with different principal G bundles over $\mathcal{M}$. However, in fact there is only one possible functor—which maps each point on $\mathcal{M}$ to the single object $*$, and the identity morphism of each point on $\mathcal{M}$ to the identity element of G. This is not unexpected—the definition of functor is suitable for categories internal to Set, and if we view $\mathcal{M}$ as merely a set rather than a manifold, indeed there should be no distinction of different bundles—as a set without topology we only have $\mathcal{M} \times G$. In order to define different principal G bundles over $\mathcal{M}$, one familiar treatment is, again, to cover $\mathcal{M}$ by some patches $\sqcup_\alpha U_\alpha \xrightarrow{\Pi} \mathcal{M}$, and then specify the transition functions. In the category theory language, given the patches, along with the aforementioned identification morphisms, we form a category $F = (\mathcal{U} \times_{\mathcal{M}}^{\Pi,\Pi} \mathcal{U} \rightrightarrows \mathcal{U})$ where $\mathcal{U} := \sqcup_\alpha U_\alpha$. Via the Π, this category F has a smooth, surjective and fully faithful functor to $\mathcal{M} \rightrightarrows \mathcal{M}$. Moreover, this category F can now have different smooth functors to BG, which specify the transition functions. Thus we obtain different principal G bundles.

One step further, we can also take G connections over $\mathcal{M}$ into account, if we consider the path groupoid $\bar{P}\mathcal{M} = (\bar{\mathcal{P}}\mathcal{M} \rightrightarrows \mathcal{M})$ instead of $\mathcal{M} \rightrightarrows \mathcal{M}$ to begin with. We need an $\widetilde{F}$ such that $\widetilde{F}_0$ is still $\mathcal{U}$, while now in $\widetilde{F}_1$, in addition to the identification morphisms, we also have paths and their compositions with the identification morphisms, just like in the S^1 example described before. (For general $\mathcal{M}$, the spaces $\bar{\mathcal{P}}\mathcal{M}$ and $\widetilde{F}_1$ are infinite dimensional, and we need to internalize the discussion in Difflg instead of just Manifold.) Then a smooth (which really means diffeological) functor from $\widetilde{F}$ to BG not only specifies the transitions functions, but also the parallel transport (Wilson lines), hence the G connection.

Here we used patches and transition functions to describe a principal bundle, but there is another familiar way to describe a principal bundle, namely the total space $\mathcal{E} \xrightarrow{\Pi} \mathcal{M}$ of the bundle. It turns out that this corresponds to another choice F' that replaces the F above, given by $F' = (\mathcal{E} \times_{\mathcal{M}}^{\Pi,\Pi} \mathcal{E} \rightrightarrows \mathcal{E})$, which again has a surjective and fully faithful map to $\mathcal{M} \rightrightarrows \mathcal{M}$ via Π. On the other hand, note that $\mathcal{E} \times_{\mathcal{M}} \mathcal{E} \cong \mathcal{E} \times G$ through the G action on the fibres of $\mathcal{E}$, thus F' has a smooth functor to BG by just looking at the G part. Again, if we begin with the path groupoid $\bar{\mathcal{P}}\mathcal{M} \rightrightarrows \mathcal{M}$ instead of $\mathcal{M} \rightrightarrows \mathcal{M}$, then we need an $\widetilde{F}'$ such that $\widetilde{F}'_0 = F'_0 = \mathcal{E}$ and in $\widetilde{F}'_1$ we also need to include paths in a suitable way so that the functor from $\widetilde{F}'$ to $\bar{P}\mathcal{M}$ is fully faithful, and then a smooth functor from this $\widetilde{F}'$ to BG would specify a total G-connection over $\mathcal{E}$, which reduces to a G-connection over $\mathcal{M}$, with the remaining components along the fibres being the BRST (or Faddeev-Popov) ghosts [101].

Now we have two ways to describe a principal bundle as a functor from some intermediate category F covering $\mathcal{M}$ to BG, one is the Čech way where $F_0 = \sqcup_\alpha U_\alpha$ consists of patches, and the other is the BRST way where $F'_0 = \mathcal{E}$ is the total space of the bundle. But given a principal bundle, these two ways of description must be equivalent in a suitable sense. Likewise if we also specify the connection. This motivates us to define ananatural transformation in the below.

Gathering the experience from these familiar treatments, it is now clear that we should

define an *anafunctor* $D \xleftarrow{F} C$, in generalization to an ordinary functor, as

$$
\begin{array}{ccccc}
D_1 & \longleftarrow & F_1 & \xrightarrow{\text{f.\,f.}} & C_1 \\
\Downarrow & & \Downarrow & & \Downarrow \\
D_0 & \longleftarrow & F_0 & \twoheadrightarrow & C_0
\end{array}
\tag{87}
$$

where there is an intermediate category F, called the "span", such that it has a surjective (rather than just essentially surjective) and fully faithful ordinary functor to C (so that F is in some sense equivalent to C but "larger in appearance"), and another ordinary functor to D. [93] The notation F_0, F_1 here seems to be in conflict with the notation we used for ordinary functor before, but in fact this is a generalization rather than a conflict—the function F_0 in the ordinary functor can be viewed as a set of order pairs $\{(c, F_0(c)) | c \in C_0\}$, which is a set that has a bijective map to C_0, and now we are generalizing this to a set with a surjective map to C_0. Pictorially,

$$\tag{88}$$
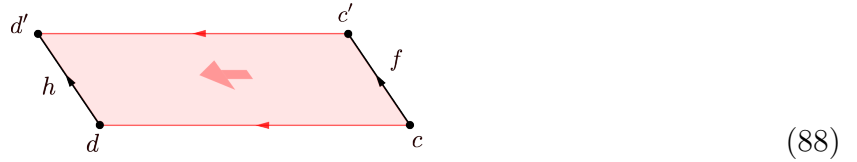
in an ordinary functor, there is a unique red arrow emanating from any given c, but in an anafunctor, there can be one or more choices of red arrows emanating from a given c, and moreover they can end at different ds; the collection of all such red arrows emanating from all c is F_0. On the other hand, the collection of all pink surfaces in the middle forms F_1; the requirement of "fully faithful" is that, given the black arrow f on the right and the two red arrows on the sides, the pink surface in the middle is uniquely determined, and hence so is the black arrow h on the left. [94]

The composition of anafunctors $D \xleftarrow{F} C$, $E \xleftarrow{G} D$ can be defined by such a category H (the arrows here represent ordinary functors)

$$
\begin{array}{c}
H \\
\swarrow \; = \; \searrow \\
E \;\leftarrow\; G \;\rightarrow\; D \;\leftarrow\; F \;\rightarrow\; C
\end{array}
\tag{89}
$$

where $H_0 = F_0 \times_{D_0} G_0$, and H_1 is determined by the condition that the ordinary functor from H to C via F is surjective and fully faithful; the ordinary functors from H to D via F

---

[93] Before the general notion of anafunctor was formulated, it has already had other names in specific contexts. In particular, anafunctor between Lie groupoids has been known as *bibundle* or *Hilsum-Skandalis morphism*. In this and some related contexts, anafunctor has also been known as *Morita morphism*, and ananatural equivalence has been known as *Morita equivalence*. In the broader context of higher homotopy theory, which is what we are concerned about, the span F is known as a *cofibrant resolution*. In this paper we will just use the general context terminology *anafunctor*.

[94] This description can be casted in the language of *double category*, where a category has the objects from both C_0 and D_0, and then there are two kinds of morphisms: those black arrows from C_1 and from D_1, and those red arrows from F_0. Although we will not directly use this language in the below, this perspective is helpful for understanding and unifying many concepts.

and via G are required to be equal. Then $E \xleftarrow{H} C$ is the resulting anafunctor. Note that the composition of anafunctors is not strictly associative, but the two results are equivalent up to invertible ananatural transformations, which we now introduce. [95]

Between two anafunctors $D \xleftarrow{F} C$, $D \xleftarrow{F'} C$ we can define an *ananatural transformation.* Consider the category H (the arrows here represent ordinary functors)

$$
\begin{array}{ccc}
 & F & \\
\swarrow & \uparrow & \searrow \\
D \quad {}^{\Phi}\Downarrow & H \quad \| & C \\
\nwarrow & \downarrow & \nearrow \\
 & F' &
\end{array}
\tag{90}
$$

where $H_0 = F_0 \times_{C_0} F'_0$, and H_1 is determined by the condition that the ordinary functor from H to C via F is equal to that via F', and is surjective and fully faithful. (Such an H is, intuitively, called the strict pull-back of $F \to C \leftarrow F'$, although we skipped the general definition of pull-back in category theory.) On the left half of the diagram, the two ordinary functors from H to D via F and via F' are in general distinct. If there is an ordinary natural transformation Φ between these two ordinary functors, then H together Φ defines are ananatural transformation between the two anafunctors of interest. If Φ is an invertible ordinary natural transformation, then H and Φ together is an invertible ananatural transformation (an ananatural isomorphism), and in this case the two anafunctors F and F' are considered equivalent.

Two anafunctors $D \xleftarrow{F} C$, $C \xleftarrow{\bar{F}} D$ are considered inverse of each other, if their compositions $\bar{F} \circ F$ and $F \circ \bar{F}$ are related to $\mathbf{1}_C$ and $\mathbf{1}_D$ respectively via invertible ananatural transformations; we say they establish an *ananatural equivalence* between C and D. It is not hard to see that if C and D are ananaturally equivalent, then there exists some span F such that there are strictly surjective and fully faithful functors from F to both C and D.

With these definitions, we can see that:

1. In our lattice loop versus circle example, $\bar{\mathcal{L}}$ and $\bar{\mathcal{P}}S^1 \rightrightarrows S^1$ are ananaturally equivalent, established by the span given by the patches, as we described.

   $\bar{\mathcal{P}}S^1 \rightrightarrows S^1$ is an example of fundamental groupoid. More generally, the fundamental groupoid of a manifold $\mathcal{M}$ is ananaturally equivalent to a skeletal category $\tilde{\pi}_1(\mathcal{M}) \rightrightarrows \pi_0(\mathcal{M})$, where the elements of $\pi_0(\mathcal{M})$ represent connected components of $\mathcal{M}$, and the elements of $\tilde{\pi}_1(\mathcal{M})$ represents classes of non-contractible loops on $\mathcal{M}$, such that $\tilde{\pi}_1(\mathcal{M})|_{a,a} \cong \pi_1(\mathcal{M}, x)$ is the usual fundamental group based at some point x on a given connected component a. Hence the name "fundamental groupoid". For $\mathcal{M} = S^1$, the skeleton is $B\mathbb{Z}$, a lattice loop with only one vertex.

2. In the principal bundle example, we have two choices of the span F for the anafunctor $BG \xleftarrow{F} \mathcal{M}$, one is the Čech choice $F = (\mathcal{U} \times_{\mathcal{M}} \mathcal{U} \rightrightarrows \mathcal{U})$ given by the patches, and the

---

[95]In the usual set theoretic construction, $(X \times Y) \times Z$ and $X \times (Y \times Z)$ are unequal as sets, but there is a bijection between them. Similarly as we involve fibre products now. Thus, the collection of all 1-categories internal to some ambient category, along with their anafunctors and ananatural transformations, form a *bicategory* which we will introduce in Section 5.3.

other is the BRST choice $F' = (\mathcal{E} \times_{\mathcal{M}} \mathcal{E} \rightrightarrows \mathcal{E})$ given by the total space, and the two choices of anafunctors are related by invertible ananatural transformation.

If we want to specify connections, then we use $BG \xleftarrow{\widetilde{F}} \bar{P}_1 \mathcal{M}$, where $\widetilde{F}_0 = F_0$ and $\widetilde{F}_1$ includes paths suitably so that $\widetilde{F} \to \bar{P}_1 \mathcal{M}$ is surjective and fully faithful. Likewise for $\widetilde{F}'$. And the two anafunctors are again related by invertible ananatural transformation.

What is the fundamental difference between Set and Manifold (or Difflg) that makes the notion of anafunctor, defined by the diagrams above, necessary when internalized in Manifold (or Difflg), but not in Set? Let us denote the anafunctor $D \xleftarrow{F} C$ by two ordinary functors $D \xleftarrow{\mathsf{F^t}} F \xrightarrow{\mathsf{F^s}} C$ where $\mathsf{F^s}$ is surjective and fully faithful. In Set, recall that this means $\mathsf{F^s}$ has an inverse $F \xleftarrow{\bar{\mathsf{F}}^s} C$. With this inverse, we can see the anafunctor $D \xleftarrow{\mathsf{F^t}} F \xrightarrow{\mathsf{F^s}} C$ of interest is equivalent (via invertible ananatural transformation) to the ordinary functor $D \xleftarrow{\mathsf{F^t} \circ \bar{\mathsf{F}}^s} C$. But crucially, the existence of such $\bar{\mathsf{F}}^s$ requires the axiom of choice; while the axiom of choice can be imposed (as is usually done) in set theory, it is in general violated upon the introduction of topology—simply because in general a projection cannot be lifted back to a continuous section. [96] Therefore, anafunctor is the more useful notion in generic ambient categories, and only in those ambient categories where the axiom of choice is respected can it be reduced to ordinary functors.

We can envision how n-anafunctor is to be defined for strict n-categories internal to some ambient category. Still $D \xleftarrow{F} C$ takes the form $D \xleftarrow{\mathsf{F^t}} F \xrightarrow{\mathsf{F^s}} C$ where $\mathsf{F^t}$ and $\mathsf{F^s}$ are ordinary, generally non-strict n-functors, and moreover $\mathsf{F^s}$ satisfies the condition that: Given the source and target $(k-1)$-morphisms $g, f \in F_{k-1}$ (it is implied that g, f themselves share the same source and target $(k-2)$-morphisms in F_{k-2}), the restriction of $\mathsf{F^s}_k$ from $F_k|_{g,f}$ to $C_k|_{\mathsf{F^s}_{k-1}(g), \mathsf{F^s}_{k-1}(f)}$ is a surjection for $k < n$, and a bijection for $k = n$. [97] And F establishes a higher ananatural equivalence between C and D if $\mathsf{F^t}$ also has these properties of $\mathsf{F^s}$.

---

[96] The axiom of choice is the statement that, if there is a collection of non-empty sets $S_a (a \in A)$, then there exists a "choice function" f from A to $\sqcup_a S_a$ such that $f(a)$ is an element of S_a (so the image $\mathrm{Im}(f)$ contains exactly one element from each S_a). This statement is obviously true (as f can be explicitly constructed) if A is a finite set, but when A is an infinite set, whether such f exists depends on whether we impose the existence as an axiom—and either way is consistent in set theory.

Even if we did impose the axiom of choice in set theory, when we introduce extra structures such as topology to the sets involved, the axiom of choice may become incompatible with the extra structures. For a familiar example, suppose S_a are fibres in a fibre bundle E that project to points a on a base manifold A. Then in general there does not exist a continuous lifting function f from A to E.

Given an essentially surjective and fully faithful ordinary functor F in some ambient category, the axiom of choice in that ambient category is needed for (and is, in fact, equivalent to) the existence of an essentially surjective and fully faithful inverse ordinary functor $\bar{\mathsf{F}}$, roughly because F_0 is in general non-injective.

[97] We want to make sure this is a sensible definition. In particular, we shall make sure that, if we view an n-category as an $(n+1)$-category with identity $(n+1)$-morphisms only, then "bijection" for $k = n$ can be replaced by "surjection", as long as we have "bijection" for $k = n+1$. This is indeed true. Given ϕ, ψ in F_n, the restriction $F_{n+1}|_{\phi,\psi}$ is empty if $\phi \neq \psi$ and has a unique element $\mathbf{1}_\psi$ if $\phi = \psi$; likewise for C_{n+1}. So, first of all, for $k = n+1$, the map from $F_{n+1}|_{\phi,\psi}$ via $\mathsf{F^s}_{n+1}$ to $C_{n+1}|_{\mathsf{F^s}_n(\phi), \mathsf{F^s}_n(\psi)}$ is automatically injective; then the only non-trivial requirement is that it is also surjective. It being surjective means, whenever $C_{n+1}|_{\mathsf{F^s}_n(\phi), \mathsf{F^s}_n(\psi)} = \{\mathbf{1}_{\mathsf{F^s}_n(\phi)} = \mathbf{1}_{\mathsf{F^s}_n(\psi)}\}$, we have $F_{n+1}|_{\phi,\psi} = \{\mathbf{1}_\psi = \mathbf{1}_\phi\}$, which indeed establishes the injectivity for $k = n$, as desired.

Compositions and ananatural transformations of higher anafunctors are essentially defined by the same diagrams (89), (90) as before. But there is some new ingredient. Consider two strict 2-categories C and D. Even between two ordinary 2-functors from C to D, we can have 2-ananatural transformations that are beyond the ordinary 2-natural transformations. This is because an ordinary 2-natural transformation involves a functor from $C_1 \rightrightarrows C_0$ to $D_2 \rightrightarrows D_1$ (in a generic 2-category D_2 not only contains identity 2-morphisms), and now we can consider the possibility that this becomes an anafunctor. This is intuitive if we think pictorially (along the lines of (88)): In (85), even if the F_0 red arrow and the G_0 red arrow emanating from c are unique for each given c (so that F, G are ordinary 2-functors), the blue surface in between (and hence the black arrow on the left) might still admit multiple choices. Of course, in a more general 2-ananatural transformation between two anafunctors, both red arrows and the blue surface emanating from any given c may admit non-unique choices. We will see how such new ingredient is relevant in our main construction in Section 5.5, in particular in (121).

Coming back to the motivating problem at the beginning of this subsection, we can say if two manifolds are homotopic, then the strict path d-groupoids of the manifolds, the strict path d-groupoids of patches covering the manifolds (with the identification morphisms between different patches), and the strict d-groupoids of the lattices discretizing the manifolds, are all ananaturally equivalent to each other as strict d-groupoids. [98]

Clearly, in physics, not only does the homotopy information of the spacetime (as a continuum manifold or a lattice) matter. Besides the topological properties, usually we are also interested in the non-topological correlations of observables at generic energy/length scales. (For example, how confinement happens in Yang-Mills theory is an important problem at the intermediate energy scale Λ_{QCD}.) Therefore, we need to care about both the ananatural equivalence class as well as more details of a category, depending on the problem of interest. Roughly speaking, towards the IR limit, the ananatural equivalence class becomes more important, because this is the information that is kept unchanged under coarse graining.

The discussions above are about the spacetime, appearing as the source category of a field configuration. Similar situation is happening on the side of the target category, too. This has already been explained in Section 5.1, except "natural equivalence" should more precisely be "ananatural equivalence": Unlike in topological lattice field theory, in a generic dynamical lattice field theory, the physics is not determined by the ananatural equivalence class of the target category, because the weight of the path integral does not respect invariance under general ananatural transformations.

That said, in the context of "topologically refining" a lattice QFT, [99] it is still useful

---

[98]Using higher category theory to lay the foundation of homotopy theory is an important program in mathematics [39], and weak higher categories must be used. Roughly speaking, the skeleton (under ananatural equivalence) of a strict n-groupoid (may as well take $n \to \infty$) can be expressed in terms of *crossed complex*, which is a generalization of the crossed module introduced before that describes the strict 2-group [102, 103], and it does not contain information about the higher order mappings between the π_m's (such as the Whitehead product). That is why weak higher category is in general needed to capture the full homotopy information [39, 99]. The Kan complex to be introduced in Section 5.4 is one such construction.

[99]We have not yet defined "topological refinement" mathematically, but we already have the experience

to consider the ananatural equivalence class of a target category. It turns out the (higher category analogue of) skeleton of the ananatural equivalence class tells us which topological operators the topological refinement of the lattice theory enables us to explicitly define, regardless of the detailed dependence of the path integral weight on the target category (in particularly, regardless of the physical dynamics or fugacity of these topological operators).

1. Consider a Villainized nlσm. For simplicity let us first assume the vortices are forbidden. The target category is $\widetilde{\mathcal{T}}^2/\Gamma \rightrightarrows \mathcal{T}$, which is ananaturally equivalent to the skeleton $B\Gamma = (\Gamma \rightrightarrows *)$, established by

$$
\begin{array}{ccccc}
\Gamma & \longleftarrow & \widetilde{\mathcal{T}}^2 \times \Gamma & \longrightarrow & \widetilde{\mathcal{T}}^2/\Gamma \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & \widetilde{\mathcal{T}} & \longrightarrow & \mathcal{T}
\end{array} \quad , \tag{91}
$$

where the span has a surjective and fully faithful ordinary functor to the right by identifying $(x, y, \gamma) \in \widetilde{\mathcal{T}}^2 \times \Gamma$ with $(\gamma' x, \gamma'' y, \gamma' \gamma \gamma''^{-1}) \in \widetilde{\mathcal{T}}^2 \times \Gamma$ for any $\gamma', \gamma'' \in \Gamma$ (and this is the categorical nature of what we called the Γ gauge invariance in Section 2.1 where $\mathcal{T} = S^1$ and $\Gamma = \mathbb{Z}$ and in Section 2.3 for more general $\mathcal{T}$ and Γ), and a surjective and fully faithful ordinary functor to the left by collapsing $\widetilde{\mathcal{T}}$ to $*$. The Γ at the 1-morphism in $B\Gamma$ is originates from the fact that $\pi_1(\mathcal{T}) \cong \Gamma$. Physically, it means the Villainized nlσm allows us to explicitly describe Γ-valued windings, regardless of whether it is important or not in the dynamics due to the path integral weight.

It is particularly illuminating to think about the deep IR limit, where the lattice is so coarse grained such that, the $\bar{\mathcal{L}}_1 \rightrightarrows \bar{\mathcal{L}}_0$ part becomes the skeleton of the fundamental groupoid, $\widetilde{\pi}_1(\mathcal{M}) \rightrightarrows \pi_0(\mathcal{M})$. If we also reduce the target category to its skeleton $B\Gamma$, then a field configuration is a homomorphism from the non-contractible loops to Γ, as expected.

When the vortices are not forbidden, the target category is $(\widetilde{\mathcal{T}}^2/\Gamma) \times \Gamma \rightrightarrows \widetilde{\mathcal{T}}^2/\Gamma \rightrightarrows \mathcal{T}$ (recall $(\widetilde{\mathcal{T}}^2/\Gamma)^{[2]} := (\widetilde{\mathcal{T}}^2/\Gamma) \times_{\mathcal{T}^2}^{(s,t),(s,t)} (\widetilde{\mathcal{T}}^2/\Gamma)$), which is ananaturally equivalent to $BE\Gamma = (\Gamma^2 \rightrightarrows \Gamma \rightrightarrows *)$. The extra Γ at the 2-morphism describes the vortices. This category is in turn ananaturally equivalent to the trivial category $*$, which physically suggests that the theory describes a trivial phase if the plaquette weight is sufficiently insensitive (just like in the traditional lattice gauge theory mentioned above).

This is one way to mathematically motivate the target category to be used for Villainized nlσm. In Section 5.5, in particular (114), we will introduce a closely related and somewhat more systematic way towards the same goal, started out by allowing rather than forbidding vortices.

2. A Villainized gauge theory is similar as long as we replace the 0-category $\mathcal{T}$ by the 1-category BG—or we can say, as long as we take $\mathcal{T} = G$ and deloop everything said above. In particular, Γ must be abelian. "$\pi_1(\mathcal{T}) \cong \Gamma$" stays "$\pi_1(G) \cong \Gamma$", but "a homomorphism from $\pi_1(\mathcal{M})$ to Γ" becomes "a homomorphism from $\pi_2(\mathcal{M})$ to Γ".

what this means from the previous sections. An important goal of the remaining parts of this paper is to lead towards a suitable definition, in Section 5.5 and Section 6.

3. If G itself abelian, we can further deloop arbitrarily many times.

4. Consider the spinon decomposed S^2 nlσm. For simplicity let us first assume the hedgehogs are forbidden. The target category is $S^2 \times SU(2) \times \mathbb{R} \rightrightarrows S^2 \times SU(2) \rightrightarrows S^2$ (recall (83)), which is ananaturally equivalent to the skeleton $B^2\mathbb{Z} = (\mathbb{Z} \rightrightarrows * \rightrightarrows *)$, established by

$$
\begin{array}{ccccc}
\mathbb{Z} & \longleftarrow & SU(2)^2 \times \mathbb{R}^2 \times \mathbb{Z} & \longrightarrow & S^2 \times SU(2) \times \mathbb{R} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & SU(2)^2 \times \mathbb{R} & \longrightarrow & S^2 \times SU(2) \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & SU(2) & \longrightarrow & S^2
\end{array}
\qquad , \qquad (92)
$$

where the span has a surjective (at the lower morphisms given any source and target) and fully faithful (at the top morphism) ordinary functor to the right by identifying $(\mathcal{U}, \mathcal{U}', a, a', s) \in SU(2)^2 \times \mathbb{R}^2 \times \mathbb{Z}$ with $(\mathcal{U}e^{i\alpha\sigma^z}, \mathcal{U}'e^{i\alpha'\sigma^z}, a + \alpha - \alpha' + 2\pi k, a' + \alpha - \alpha' + 2\pi k', s + k - k') \in SU(2)^2 \times \mathbb{R}^2 \times \mathbb{Z}$ for any $k' \in \mathbb{Z}$ and $\alpha, \alpha' \in \mathbb{R}$ (this is the categorical nature of the 1-form $\mathbb{Z}$ gauge invariance and the $\mathbb{R}$ mod $2\pi\mathbb{Z}$ gauge invariance in Section 2.4), and a surjective and fully faithful ordinary functor to the left by collapsing $SU(2)$ and $\mathbb{R}$ to $*$. Similar to the Villainization case, the $\mathbb{Z}$ in the 2-morphism is related to the fact that $\pi_2(S^2) \cong \mathbb{Z}$, and physically it means the spinon decomposition allows us to explicitly describe $\mathbb{Z}$-valued skyrmions.

When the hedgehogs are not forbidden, the target category has the space of 3-morphisms being $S^2 \times SU(2) \times \mathbb{R} \times \mathbb{Z}$, where the extra $\mathbb{Z}$ (compared to the space of 2-morphisms) describes the hedgehogs. The target category is ananaturally equivalent to $B^2E\mathbb{Z}$, which is in turn ananaturally equivalent to the trivial category, and this physically suggests the theory can describe the trivial phase if the cube weight is sufficiently insensitive.

Again, in Section 5.5, in particular (117), we will introduce a closely related and somewhat more systematic way towards the same goal, started out by allowing rather than forbidding hedgehogs.

From these discussions, it becomes clear that to tackle the main problems we aim at, for nlσm we need a topological refinement for $\mathcal{T} = S^3$ so that the target category is internal to Manifold (which implies finite dimensional) and has a ananatural equivalence—similar to the ones in the examples above—to $B^3\mathbb{Z}$ (when baryon non-conserving hedgehogs are forbidden) or $B^3E\mathbb{Z}$ (when baryon non-conserving hedgehogs are allowed); then, for Yang-Mills theory we take $\mathcal{T} = SU(N)$ and suitably deloop the refined target category, to obtain one that has a ananatural equivalence to $B^4\mathbb{Z}$ (when Yang monopoles are forbidden) or $B^4E\mathbb{Z}$ (when Yang monopoles are allowed). If we do not care about the d.o.f. being finite dimensional—so that we are internalizing in Difflg instead of Manifold—then we can turn (51) (along with the Villainizing layer at the top of (48)) into a strict higher category (just like how (46) is related to (92)) to fulfill the goal, as briefly explained in Section 5.1. However, for an actual lattice model, the d.o.f. being finite dimensional is important. To satisfy all these conditions, it turns out we have to work with more flexible higher categories, in generalization to the strict higher categories that we have been working with so far.

5.3 Weak categories

Let us introduce some weak categories that are more flexible than the strict ones, but not as flexible as what we will finally need. In particular we will focus on the weak 2- and 3-categories called *bicategories* and *tricategories*. They have been extensively used in the study of topological phases and generalized symmetries, which we will briefly mention but not go deeply into. We will mainly emphasize the conceptual aspects which will lead us towards our final construction in the next subsections.

In this subsection we will ignore topology, so that all the structures are internalized in Set. In fact, our very reason to go towards even more flexible definitions of categories in the next subsection is to take topology into account.

From the definitions of lax 2-natural transformation and lax 2-functor, we have learned that, when non-trivial 2-morphisms are available, we may replace the equality signs that appear in some consistency conditions between 1-morphisms by more general (i.e. possibly non-identity) 2-morphisms between 1-morphisms. Now, we note that even in the definition of category itself, there are some equality signs describing consistency conditions between 1-morphisms—the associativity condition $(h \circ g) \circ f = h \circ (g \circ f)$, and the unital condition $f \circ \mathbf{1}_a = f = \mathbf{1}_b \circ f$. These equality signs can be understood as identity 2-morphisms, which are the only 2-morphisms available in a 1-category. However, if we have a 2-category with more general 2-morphisms, it is possible to relax these conditions on 1-morphisms, by replacing the equality signs (identity 2-morphisms) with more general 2-morphisms:

$$h \circ (g \circ f) \xLeftarrow{\alpha_{h,g,f}} (h \circ g) \circ f, \qquad f \xLeftarrow{\lambda_f^L} \mathbf{1}_b \circ f, \qquad f \xLeftarrow{\lambda_f^R} f \circ \mathbf{1}_a \tag{93}$$

where $\alpha_{h,g,f}$ is called the *associator* for h, g, f, and λ_f^L and λ_f^R are the *left and right unitors* for f; we require these 2-morphisms to be invertible under vertical composition. We expect the associators and unitors to satisfy suitable consistency conditions which ultimately follow from the fact that the only available 3-morphisms are identity 3-morphisms. It is helpful to explain the details of these conditions, because over the process we will develop some important perspectives.

To begin, we picture an associator as

$$\tag{94}$$
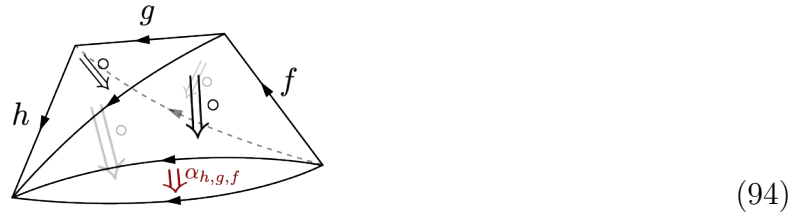
which looks like a tetrahedron but with one edge becoming a slit filled with the associator 2-morphism.

- The most crucial point to read-off from this picture is that, the composition $\circ$ should be thought of as a generalized kind of 2-morphism, or say a 2-cell, which has a triangular shape that takes two source 1-morphisms to one target 1-morphism, rather than

the previous globular shape (78) which takes one source 1-morphism to one target 1-morphism. Let us denote by $[g \circ f]$ the triangular shape 2-morphism that takes g, f to $g \circ f$.

To appreciate the importance of this point of view, let us consider such a situation (which will actually appear in our discussion soon). It is possible that, as 1-morphisms, $h \circ (g \circ f)$ and $(h \circ g) \circ f$ are equal. In that case, however, we can still have an associator $\alpha_{h,g,f}$ which is not the identity 2-morphism. What would the associator mean if the two 1-morphisms are equal already? The point of view above explains it: While $h \circ (g \circ f)$ denotes a 1-morphism, we shall also view the process as a trapezoidal shaped 2-morphism, denoted as $[h \circ (g \circ f)]$, that takes three source 1-morphisms h, g, f to one target 1-morphism which we called $h \circ (g \circ f)$; likewise for $[(h \circ g) \circ f]$. Thus, regardless of whether $h \circ (g \circ f)$ and $(h \circ g) \circ f$ are equal as 1-morphisms, $[h \circ (g \circ f)]$ and $[(h \circ g) \circ f]$ may still be different as trapezoidal shaped 2-morphisms, and the difference is captured by the associator $\alpha_{h,g,f}$. So (94) means

$$[h \circ (g \circ f)] = \alpha_{h,g,f} \circ_{\mathsf{v}} [(h \circ g) \circ f] \,, \tag{95}$$

where the equality is made sense of as trapezoidal shaped 2-morphisms taking three sources to one target. This is what the picture (94) really means.

In particular, the "equality as 2-morphisms" is because there is no non-identity 3-morphisms—in the picture, the bounded 3d volume represents the equality sign in the formula above. [100]

- Since the only available 3-morphisms are identity 3-morphisms, it is easy to see that the 2-morphisms satisfy strict associativity under consecutive vertical compositions, and strict interchangeability between vertical and horizontal compositions. On the other hand, the associativity under consecutive horizontal compositions is slightly modified. Replacing the arrows for h, f, g in (94) by slits $h' \overset{\varphi}{\Leftarrow} h, g' \overset{\psi}{\Leftarrow} g, f' \overset{\rho}{\Leftarrow} f$ under consecutive horizontal compositions, it is not hard to see in the end we will be left with a 3d volume bounded by four slits, which represents the equality between 2-morphisms

$$(\varphi \circ_{\mathsf{h}} (\psi \circ_{\mathsf{h}} \rho)) \circ_{\mathsf{v}} \alpha_{h,g,f} = \alpha_{h',g',f'} \circ_{\mathsf{v}} ((\varphi \circ_{\mathsf{h}} \psi) \circ_{\mathsf{h}} \rho) \,. \tag{96}$$

- When four 1-morphisms j, h, g, f are consecutively composed, starting with the upper left in the diagram below, by using associators and whiskering, we conclude that the two results at the lower right must be equal as 2-morphisms, as indicated by the blue equal sign. Since their triangular parts are the same, the equality becomes that of the

---

[100]Our perspective becomes closer and closer to that of a *simplicial set*, which is indeed what we will get to in the next subsection. Intuitively, it becomes more and more similar to a lattice theory (which is desired), or to a high dimensional tiling game with certain rules, such as what kind of tiles are available and which ones can join together. This is indeed the nature of it.

vertical compositions of the (whiskered) associators.

$$
\begin{array}{c}
\alpha_{j,h,g\circ f} \\[2pt]
\text{left whiskering} \quad \alpha_{h,g,f} \\[2pt]
\mathbf{1}_j \circ_{\mathsf h} \alpha_{h,g,f} \\[2pt]
\mathbf{1}_j \circ_{\mathsf h} \alpha_{j,h\circ g,f} \\[2pt]
\alpha_{j\circ h,g,f} \circ_{\mathsf v} \alpha_{j,h,g\circ f} \\[2pt]
\alpha_{j,h,g} \circ_{\mathsf h} \mathbf{1}_f \circ_{\mathsf v} \alpha_{j,h\circ g,f} \circ_{\mathsf v} \mathbf{1}_j \circ_{\mathsf h} \alpha_{h,g,f} \\[2pt]
\alpha_{j,h,g} \; \alpha_{j,h\circ g,f} \; \mathbf{1}_j \circ_{\mathsf h} \alpha_{h,g,f} \quad \text{right whiskering}
\end{array}
\tag{97}
$$

This is often called the "pentagon equation" of the associators (the pentagon refers to the five red equal signs, at which associators are introduced).

The pentagon equation can also be thought of as five tetrahedra (94) piecing up to a 4d simplex, where the slits are taken care of by filling in two extra 3d volumes representing the whiskerings. And the existence of such an equation is simply because the only available 4-morphisms are identites ones.

- In (94) or (95), if $g = \mathbf{1}_b$ is some identity 1-morphism, by applying the unitors and suitable whiskerings, we obtain

$$
\mathbf{1}_h \circ_{\mathsf h} (\lambda_f^L)^{-1} = \alpha_{h,\mathbf{1}_b,f} \circ_{\mathsf v} ((\lambda_h^R)^{-1} \circ_{\mathsf h} \mathbf{1}_f) \tag{98}
$$

which is often called the "triangle equation", similar to the pentagon equation above.

This explains how a single, simple fact that all available 3- and higher morphisms are identities lead to a set of seemingly complicated consistency conditions satisfied by the associators and the unitors—so these conditions can be thought of as being derived, rather than being imposed at will. Such a 2-category is called a *weak 2-category*, or a *bicategory*. Between weak 2-categories, it is generally impossible to define strict 2-functors, so we must use pseudo or lax 2-functors and 2-natural transformations.

There is a coherence theorem for 2-categories stating that every weak 2-category is naturally equivalent to some strict 2-category (and this is not true for higher categories), but practically there are many advantages to work with weak 2-categories [40, 100].

In our main construction we do not directly use bicategories. However, they are widely used in both mathematics and theoretical physics. Here we briefly review some applications in physics related contexts. Most of these applications concentrate on bicategories with a single object. (A bicategory with a single object can be viewed as the delooping BM of a 1-category M equipped with suitable extra structures; such an M is called a *monoidal category*.)

89

One major application is on the classification of 2-groups [104]. It is proven that every 2-group (recall we ignore topology for now) is naturally equivalent to a skeletal weak 2-group $K \ltimes A \rightrightarrows K \rightrightarrows *$ where A is abelian, and being skeletal at the 1-morphism level means $\mathsf{s}((k,a)) = \mathsf{t}((k,a)) = k$; [101] the unitor is trivial, and the associator $\widetilde{\alpha} : K^3 \to A$ (where $\alpha_{k,k',k''} = (kk'k'', \widetilde{\alpha}_{k,k'k''}) \in K \ltimes A$, so $\mathsf{s}(\alpha_{k,k',k''}) = \mathsf{t}(\alpha_{k,k',k''}) = kk'k''$ and the non-trivial content in α is $\widetilde{\alpha}$) is well-defined as an element of the group cohomology $H^3_{\text{group}}(K; A)$. This classification is important in physics because, as we have seen before, the phase of a system is characterized by the natural equivalence class of the target category of the low energy effective theory (which is in general different from that of the original target category in a UV lattice theory), and thus the phases of those systems described by 2-group symmetries or 2-group gauge theories at low energies are classified using the skeletal weak 2-groups [34, 42, 43]. In particular, given a strict 2-group $G \ltimes H \rightrightarrows G \rightrightarrows *$, the naturally equivalent skeletal weak 2-group has $A = \ker(\widetilde{\mathsf{t}}), K = \operatorname{coker}(\widetilde{\mathsf{t}})$, forming the exact sequence $* \to A \to H \xrightarrow{\widetilde{\mathsf{t}}} G \to K \to *$; the associator arises from the fact that, as groups, in general $H \neq A \times (H/A)$ and $G \neq K \times (H/A)$. [102] (Conversely, by the coherence theorem mentioned above, given a weak skeletal 2-group, there always exists a naturally equivalent strict 2-group; more particularly, given A and K in the exact sequence, the "2-extension problem" of finding the possible choices of H and G is indeed classified by $H^3_{\text{group}}(K; A)$, the data encoded by the associator.) In the Villainized gauge theories we discussed, K is trivial (as we said, our work is interested in generic dynamics, rather than just the low energy phases, therefore we not only care about the natural equivalence class of the target category, but also the target category itself, so K being trivial does not mean the theory is trivial), but there are physical applications with non-trivial K [32–34, 42, 43], including studies on the possible low energy phases after Yang-Mills confinement and/or Higgsing.

It can also be noted that, when $A = U(1)$, we may use the associator as the Dijkgraaf-Witten phase [19] for a 3d topological order with K lattice gauge field (recall we ignore topology here, so K is discrete, and thus we may forbid its flux, as is assumed in finite group Dijkgraaf-Witten theory), or as the WZW phase for a 2d symmetry protected topological order with K global symmetry [23]. But in these applications, it is better not to view α as an associator 2-morphism, but (equivalently) as a non-identity 3-morphism, i.e. (95) becoming

$$[h \circ (g \circ f)] \overset{\alpha_{h,g,f}}{\Longleftarrow} [(h \circ g) \circ f] \,, \tag{99}$$

---

[101]So this is an example of the situation we explained before, that $k \circ (k' \circ k'') \overset{\alpha_{k,k',k''}}{\Longleftarrow} (k \circ k') \circ k''$ has the same source and target 1-morphisms, but the associator is still meaningful.

[102]While we will not rigorously prove the natural equivalence here, we can explain how the associator arises in more details. Let $g_k \in G$ denote a chosen lift of $k \in K$. We have $g_{kk'} = \widetilde{\mathsf{t}}(\beta_{k,k'})g_k g_{k'}$ where $\beta_{k,k'} \in H$; it is in general impossible to simultaneously make all $\widetilde{\mathsf{t}}(\beta_{k,k'}) = 1$ as $G \neq K \times (H/A)$ in general. When three elements in K are composed, we have $g_{kk'k''} = \widetilde{\mathsf{t}}(\beta_{kk',k''})\widetilde{\mathsf{t}}(\beta_{k,k'})g_k g_{k'} g_{k''} = \widetilde{\mathsf{t}}(\beta_{k,k'k''})(g_k \widetilde{\mathsf{t}}(\beta_{k',k''})g_k^{-1})g_k g_{k'} g_{k''}$. In the language of strict 2-group we write $g_{kk'} \overset{\beta_{k,k'}}{\Longleftarrow} g_k g_{k'}$, and $g_{kk'k''} \overset{\beta_{kk',k''}\beta_{k,k'}}{\Longleftarrow} (g_k g_{k'})g_{k''}$ and $g_{kk'k''} \overset{\beta_{k,k'k''}(^k\beta_{k',k''})}{\Longleftarrow} g_k(g_{k'} g_{k''})$. Thus we find $\widetilde{\alpha}_{k,k',k''} := (\beta_{k,k'k''}\beta_{k',k''})^{-1}(\beta_{kk',k''}(^k\beta_{k,k'}))$ satisfies $\widetilde{\mathsf{t}}(\widetilde{\alpha}_{k,k',k''}) = 1$, i.e. $\widetilde{\alpha}_{k,k',k''} \in A$. The pentagon equation and the facts that in general $\beta \notin A$ but has an ambiguity parametrized by A implies $\widetilde{\alpha} \in H^3_{\text{group}}(K; A)$.

or more pictorially, (94) becomes a tetrahedron (this is possible since $(h \circ g) \circ f = h \circ (g \circ f) \in K$) with the associator 3-morphism $\alpha_{h,g,f}$ filling the 3d volume—indeed this is how we usually think of the Dijkgraaf-Witten phase. Moreover this will make better connection to the cases of continuous-valued d.o.f. to be discussed in the next subsection. (More general topological orders will also be mentioned there, and further discussed in Section 7.)

Now let us briefly introduce weak 3-category, or *tricategory*. In our main construction, when we go from S^3 nlσm to $SU(2)$ gauge theory (which, as we can tell now, is some kind of delooping process) in Section 4.2, recall there is a potential issue involving Yang-Baxter equation that we could have had encountered. Now we can understand the origin of this problem in terms of tricategory.

When non-identity 3-morphisms are available, the consistency conditions between 2-morphisms in a strict or weak 2-category can be relaxed by replacing equalities (identity 3-morphisms) with more general invertible 3-morphisms. These conditions include the unital law for identity 2-morphisms, the interchangeability, the vertical associativity, the modified horizontal associativity (96), the pentagon equation (97), and the triangle equation (98).

Let us first look at the case where a 3-category is almost like a strict 3-category, except the interchangeability of 2-morphisms—the last diagram of (80)—is weakened. Such weak 3-categories are called *Gray 3-categories*. There is a coherence theorem for 3-categories stating that every weak 3-category is naturally equivalent to some Gray 3-category, but not to any strict 3-category in general [105]. Such a level of strictness, that makes the n-categories "as strict as possible" but still able to be equivalent to the generic weak n-categories, is called "semi-strict". For $n = 2$, semi-strict is strict. For $n = 3$, semi-strict can be Gray, but there are also other options.

In a Gray 3-category, the last diagram of (80) is relaxed, i.e. composing vertically first and then horizontally and composing horizontally first and then vertically may result in different 2-morphisms, but we may specify an invertible *interchanger* 3-morphism between them. (Even when the two resulting 2-morphisms are equal, a non-identity interchanger is still meaningful, just like the associator case we discussed before.) But practically it is more convenient to do the following. We first define vertical composition and left and right whiskerings, and then use the whiskerings and vertical composition to define two kinds horizontal compositions (instead of defining one kind),

$$\tag{100}$$

91

and then introduce the interchanger 3-morphism $\rho_{\psi,\phi}$ that relates these two kinds of horizontal compositions:

$$\psi \circ_{\mathsf{r}}^{\mathsf{l}} \phi \quad \overset{\rho_{\psi,\phi}}{\Longleftarrow} \quad \psi \circ_{\mathsf{l}}^{\mathsf{r}} \phi \,. \tag{101}$$

It is not hard to see that $\rho_{\psi,\phi \circ_{v} \chi}$ is given by the composition of 3-morphisms (in the third direction other than horizontal and vertical)

$$\psi \circ_{\mathsf{r}}^{\mathsf{l}} (\phi \circ_{\mathsf{v}} \chi) \quad \overset{\mathbf{1}_{\mathsf{t}(\psi) \circ_{\mathsf{h}} \phi} \circ_{\mathsf{v}} \rho_{\psi,\chi}}{\Longleftarrow} \quad (\psi \circ_{\mathsf{l}}^{\mathsf{r}} \chi) \circ_{\mathsf{r}}^{\mathsf{l}} \phi = (\psi \circ_{\mathsf{r}}^{\mathsf{l}} \phi) \circ_{\mathsf{l}}^{\mathsf{r}} \chi \quad \overset{\rho_{\psi,\phi} \circ_{\mathsf{v}} \mathbf{1}_{\mathsf{s}(\psi) \circ_{\mathsf{h}} \chi}}{\Longleftarrow} \quad \psi \circ_{\mathsf{l}}^{\mathsf{r}} (\phi \circ_{\mathsf{v}} \chi) \tag{102}$$

where we used the assumption that all composition rules other than interchangeability are strict. From this we can further derive that, when three 2-morphisms are consecutively "horizontally composed", there will be $3! = 6$ different definitions related by six interchangers (with suitable 2-whiskerings), and they satisfy a consistency constraint. Such a "hexagon equation" [103] constraint is in fact the Yang-Baxter equation; it is due to the fact that the only available 4-morphisms are identities, just like the pentagon equation (97) for the associators comes from the fact that the only available 3-morphisms are identities.

If we go beyond Gray 3-category by weakening more composition rules (such as the associativities of 1- and/or 2-morphisms), then the form of (102) and hence the form of the Yang-Baxter equation will change, but the spirit is the same—all constraints come from the fact that the only available 4-morphisms are identities.

One major application of weak 3-category in physics occurs in delooping, which, as we have seen, is important for gauge theory. Recall a (delooped) group BG can be delooped to a 2-category B^2G only if G is abelian, due to the interchangeability condition. If we really do want to deloop, we may discard elements of G by only keeping its center $Z(G)$ and deloop to $B^2Z(G)$. However, if we have a (delooped) strict 2-group $G \ltimes H \rightrightarrows G \rightrightarrows *$, we may deloop it to a Gray 3-category not by discarding information, but by specifying more information—the interchanger (even if G is abelian, specifying an interchanger is still meaningful, for reason explained before). The same is true for more general 2-categories. [104]

In physics, delooping occurs in two common ways. One is when we take $\mathcal{T} = G$ in the target category of a nlσm and deloop it to the target category of a gauge theory, as we have seen before. Another is when we start with a gauge theory, but look at the gauge invariant (up to conjugation) fluxes, so that G-valued link d.o.f. (as 1-morphisms) are ignored, and we only look at the G-valued plaquette fluxes (as 2-morphisms). The second way is commonly

---

[103]This is different from what is usually called the "hexagon equation" in topological order. There, the "hexagon equation" is a generalized version of (102), when the associators are non-identity.

[104]Of particular interest in topological order (see e.g. [41, 106]) is a (possibly weak) 2-categories with a single object, which is by definition a delooped monoidal category. By further specifying the interchanger, it can be further delooped to a weak 3-category. In this context, a monoidal category with the interchanger specified (so that it can be delooped twice) is called a *braided monoidal category*, and the interchanger is also called *braiding*. But there are different choices of braiding/interchanger specification. A construction called *Drinfeld center* turns a monoidal category to a larger one by essentially including all possible consistent ways of braiding specifications, and this larger monoidal category will naturally be braided. In a suitable sense, the Drinfeld center is the generalization of the notion of "center" for usual groups.

seen in the study of anyons. [105] Here we will focus on the first way.

Consider the strict 2-groupoid $\bar{\mathcal{P}}_2 S^3 \times U(1)/WZW \rightrightarrows \bar{\mathcal{P}}S^3 \rightrightarrows S^3$ which, as we said in Section 5.1, re-expresses the structure (51)—which describes WZW curving of S^3 nlσm in the continuum. [106] Now we view the space S^3 as a group $SU(2)$ and try to deloop the category to understand CS in $SU(2)$ gauge theory in the continuum. We note that the 1-category part $\bar{\mathcal{P}}SU(2) \rightrightarrows SU(2)$ cannot be delooped because the interchangeability would not be satisfied, as $\bar{\Omega}_* SU(2) = \{\gamma \in \bar{\mathcal{P}}SU(2) | \mathsf{t}(\gamma) = \mathsf{s}(\gamma) = \text{some fixed } g\}$ is non-abelian under concatenation. [107] More explicitly, when we attempt to deloop to something like "$\bar{\mathcal{P}}SU(2) \rightrightarrows SU(2) \rightrightarrows *$", we can define the vertical composition as the concatenation of paths, and define the left/right whiskering as the group multiplication of a group element on the left/right of a path in the group. Then the two horizontal compositions in (100) indeed yield two different 2-morphisms. This is nothing but what we have already seen in (75). Fortunately, originally we also have non-trivial 2-morphisms, $\bar{\mathcal{P}}_2 S^3 \times U(1)/WZW$, which will become 3-morphisms after delooping to $\bar{\mathcal{P}}_2 S^3 \times U(1)/WZW \rightrightarrows \bar{\mathcal{P}}S^3 \rightrightarrows S^3 \rightrightarrows *$, so we can choose suitable elements from $\bar{\mathcal{P}}_2 S^3 \times U(1)/WZW$ as interchangers, which explains what we discussed below (77). (But a crucial point emphasized there is that practically we did not have to really make effort to choose these interchangers that satisfy the Yang-Baxter equation. The reason will be discussed in Section 6.2.)

5.4   Simplicial weak categories, Kan complexes

The crucial perspective brought to us by (94), that we should think of the composition $\circ$ of 1-morphisms as a 2-morphism of triangular shape rather than globular shape, opens up an obvious new possibility: Can we consider 2-categories with more general triangular shaped 2-morphisms?

The simplest case is to consider triangular shaped 2-morphisms obtained by vertically composing $\circ$ and a globular shaped 2-morphism:

$$\tag{103}$$

If this covers all the cases, then of course we achieve nothing new. We want to consider scenarios where a triangular shaped ψ cannot be naturally decomposed into an ordinary composition $\circ$ and a globular shaped ϕ.

---

[105] Given a Dijkgraaf-Witten theory with K gauge field (which we mentioned above), an anyon type is described by an element of the Drinfeld center (see the previous footnote), which specifies the flux as well as the charge (the charge comes from the braiding specification) of an anyon. Likewise for defects in symmetry protected topological order. The cases of more general topological orders, as well as some conceptual questions, will be discussed in the next subsection and in Section 7.

[106] At the beginning of this subsection we said we will ignore smoothness in this subsection. Although this example involves smoothness, the issue we will describe now is largely independent of smoothness.

[107] Since we are talking about concatenation, this is unrelated to whether G itself is non-abelian or not. The same problem exists for e.g. $G = U(1)^2$ too.

Continuity/smoothness requirement makes such more general scenarios necessary. Suppose the space of all triangular shaped 2-morphisms ψ form a fibre bundle (or some more general covering) over the space $C_1 \times_{C_0}^{\mathsf{s},\mathsf{t}} C_1$ of the two source 1-morphisms g, f. Being able to define a unique $\circ$ in a continuous manner means this bundle (or more general covering) has a continuous section. But then we can conceive scenarios in which the bundle has no continuous section—so that there is no good way to define a unique composition "$\circ$" that is continuous in its two source 1-morphisms; rather, we should just think about generic triangular shaped 2-morphisms, interpreted as non-unique compositions, such that the different compositions can be related by globular shaped 2-morphisms:

$$
\begin{array}{ccc}
\raisebox{-1em}{$g \Downarrow \psi\; f$} & = & \raisebox{-1em}{$g\; \Downarrow \psi'\; h'\; f\; \Downarrow \phi$}
\end{array}
\qquad (104)
$$

(The composition of a triangular shaped 2-morphism and a globular shaped 2-morphisms appended on one of the source 1-morphisms, rather than appended on the target 1-morphism as shown above, should also be specified.)

This can be casted in terms of anafunctor—which, as we have seen in Section 5.2, also becomes necessary when continuity/smoothness is required. Recall the hom-category $C|_{b,a}$ introduced in Section 5.1, with $(C|_{b,a})_0 = C_1|_{\mathsf{t}=b,\mathsf{s}=a}$ the space of objects, and $(C|_{b,a})_1 = C_2|_{\mathsf{tt}=b,\mathsf{ss}=a}$ (the globular shaped 2-morphisms between 1-morphisms in $C_1|_{\mathsf{t}=b,\mathsf{s}=a}$) the space of 1-morphisms. The composition $\circ$ along with $\circ_\mathsf{h}$ forms an ordinary functor from $C|_{c,b} \times C|_{b,a}$ to $C|_{c,a}$. However, once continuity/smoothness is taken into consideration, we should also consider anafunctor to capture more interesting possibilities, and "the triangular shaped 2-morphisms" are nothing but the objects of the span F in this anafunctor. (In more compact notations, $C_{\geq 1}$ from a 2-category is a 1-category, $\circ$ and $\circ_\mathsf{h}$ form an ordinary functor from $C_{\geq 1} \times_{C_0}^{\mathsf{s},\mathsf{t}} C_{\geq 1}$ to $C_{\geq 1}$, while more generally we need an anafunctor.)

A problem familiar in topological order exemplifies the necessity to introduce such generality. When we introduced (94), continuity/smoothness was not part of the consideration. Once continuity/smoothness is taken into account, we may want the associator $\alpha_{h,g,f}$ to be continuous/smooth in h, g, f. But this requirement is too restrictive to capture many interesting cases. For instance, recall the symmetry protected nlσm [23] mentioned in the previous subsection, where the associator $\widetilde{\alpha}_{h,g,f}$ specifies an element of the group cohomology $H^3_{\text{group}}(K; U(1))$ and serves as 2d WZW phase. When K is discrete, everything is fine. When K is a Lie group, it is well-known that to suitably define $H^3_{\text{group}}(K; U(1))$, we should not only consider those $\widetilde{\alpha}$ that are continuous from K^3 to $U(1)$, but also those that are only piecewise continuous (Borel), in order to capture many interesting cases. [108] While such definition of group cohomology is mathematically consistent, the discontinuity makes the corresponding lattice model unphysical. Part of the problem here is indeed that we have defined a unique notion of composition. What we will explain below, (128), that leads to the construction in Section 4.1 for the 2d WZW phase of lattice S^3 nlσm, is the solution to this kind of problem.

---

[108] More mathematically, $H^3_{\text{group}}(K; U(1))$ defined under this piecewise continuous condition, rather than the strictly continuous condition, is isomorphic to $H^4(|BK|; \mathbb{Z})$.

(If we go to 3d and gauge the global symmetry K by introducing dynamical gauge field, we obtain a Dijkgraaf-Witten theory [19]. But K is a Lie group now, so unlike the cases of finite groups, it is unphysical to demand the K gauge field to be flat. Then the problem indeed becomes that of defining K CS [19] on the lattice, and the solution will be (130) in the below, that leads to the construction in Section 4.2.)

With these motivations in mind, we are now ready to introduce the notion of *simplicial weak category*, [109] which is sufficiently weak so that it is powerful enough to fulfill our goal.

We first define a *simplicial set*—which is not a single set, but a collection of sets related in a "simplicial fashion". To begin, we still have a set C_0 of objects, pictured as points, and a set C_1 of 1-morphisms (or 1-cells), pictured as arrows between objects. But now C_2 is a set of 2-morphisms (or 2-cells) not of globular shape between two 1-morphisms, but of triangular shape between three 1-morphisms. Likewise, C_3 is a set of tetrahedral shape 3-morphisms (or 3-cells) between four 2-morphisms, and so on. Thus, just like the source and target maps in Section 5.1:

- We have $k+1$ maps ∂_i $(i = 0, 1, \cdots, k)$ from C_k to C_{k-1}, called the *face maps*. We may think of ∂_i as removing the ith vertex from a k-simplex to obtain a $(k-1)$-simplex. For example for $k = 2$:
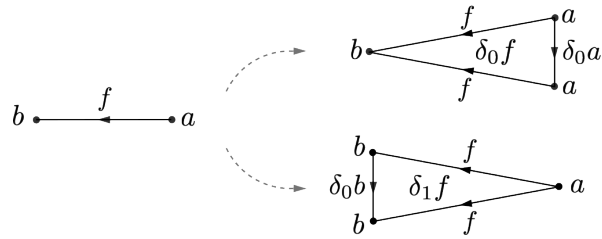
$$\begin{array}{c}\includegraphics{} \end{array} \tag{105}$$

Moreover, we would like to be able to view a $(k-1)$-cell as some special kind of k-cell, much like the identity maps in Section 5.1:

- We have k maps δ_i $(i = 0, 1, \cdots, k-1)$ from C_{k-1} to C_k, called the *degeneracy maps*. There are k of them, because we may think of δ_i as repeating the ith vertex of a $(k-1)$-simplex to obtain a k-simplex. For example for $k = 2$:

$$\tag{106}$$

The face maps and degeneracy maps satisfy some pictorially obvious constraints (sometimes called simplicial identities), such as $\partial_1 \partial_0 \psi = \partial_0 \partial_2 \psi$ and $\partial_0 \delta_1 f = \delta_0 \partial_0 f (= \delta_0 b)$ in the diagrams above. We may truncate at some $k \leq n$ if we want, which means all higher cells are essentially expressing identities.

---

[109]There is a potential confusion of terminology. Here we mean "weak category modeled by simplicial set", rather than "simplicial object internalized in some category of weak categories". This will be clear from our definition below.

Just like (86), a simplicial set can thus be viewed as a diagram

$$\cdots \ C_2 \ \substack{\longrightarrow \\ \longleftarrow \\ \longrightarrow \\ \longleftarrow \\ \longrightarrow} \ C_1 \ \substack{\longrightarrow \\ \longleftarrow \\ \longrightarrow} \ C_0 \tag{107}$$

internalized in the ambient category Set. Now, if we internalize the same diagram in the ambient category Manifold instead, we get a simplicial set whose C_k are manifolds, and whose maps in between are smooth maps. Such a simplicial set is called a *simplicial manifold*—which must not be confused with a manifold discretized into a simplicial complex.

Sometimes we may want to impose some additional conditions on a simplicial set so that it becomes more like a usual category. Consider such a "horn",
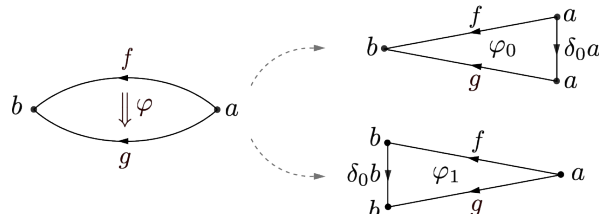

$$\tag{108}$$

which means two 1-cells are about to compose. In a usual category, there is a unique composition $[g \circ f]$. We motivated by saying we want more possibilities. But the definition of simplicial set also allows the scenario where we end up with less possibility—as there might just exist no 2-cell ψ that satisfies $\partial_0 \psi = g, \partial_2 \psi = f$. Sometimes we may want to avoid such scenario. For many of our applications, we will impose the *Kan condition* that,

- For any "horn" formed by k many $(k-1)$-cells that looks like k out of the $k+1$ faces of a k-cell, there indeed exists at least one k-cell that takes them as k out of its $k+1$ faces. The k-cell, possibly non-unique, can be viewed as one way of composing these $(k-1)$-cells, and the result of this particular composition is the remaining face.

Such a simplicial set is called a *Kan complex*, which is a simplicial version of higher groupoid where all cells are "invertible", because the condition does not distinguish between "source" and "target"—given a k-cell, we can view any k faces of its as source and the remaining one face as target. [110] In our application we will also consider Kan complexes internalized in Manifold, sometimes called *Kan simplicial manifolds*.

Note that we did not need to separately include globular shaped higher cells. This is because the role played by globular shaped n-cells can be effectively covered by simplicial shaped n-cells. For instance for $n = 2$:

$$\tag{109}$$

---

[110]If we impose other (often less stringent) conditions in replacement of Kan condition, then we get other kinds of simplicial weak categories, such as quasi-category.

Now it appears that simplicial weak categories are weakened to such an extent that they become conceptually simple to understand again. The strict categories and strict functors in Section 5.1 were easy to understand because the rules are all dictated by strict equality signs. As we began to involve lax functors, anafunctors, bicategories and tricategories, the rules became a little harder to follow, because there are some weakened rules as well as some strict equalities. But now, as we further weaken the rules and arrive at simplicial set, all the rules are essentially of the form "given the $(k-1)$-cells around, which k-cells are we allowed to fill-in", much like some kind of tiling game, so the concept becomes simple to comprehend again. Different simplicial sets just have different details in the "rules of the game".

The analogue of an ordinary functor between simplicial sets is obviously a *simplicial map* such that each C_k is mapped to D_k with the face maps and degeneracy maps preserved. On the other hand, when internalized in Manifold, we should use the analogue of anafunctor to map between simplicial manifolds; roughly speaking, the idea is the essentially the same as what we described below (88), but applied to higher dimensional cells as well. We will see how this works in our main construction below.

And it is not hard to sense that being "simplicial" is not of crucial importance here. We can as well consider "cubical sets/manifolds", whose definition is obvious from the name. A simplicial set and a cubical set might be equivalent (contain the same essential information) in a suitable sense, essentially seen by dividing an n-cube into multiple n-simplices.

Let us relate the previously discussed categories to the more general notion of simplicial weak categories.

- Given a strict n-category, or a weak one with globular shaped higher morphisms, such as those introduced in Section 5.3, we can naturally build an associated simplicial set or simplicial manifold essentially by (103), (94) and analogues for higher morphisms, known as the *nerve* of the category. (By appending an extra ∘ triangle to (103) we can build an associated cubical set or cubical manifold.) The nerve of a category C is commonly denoted as $\mathcal{N}(C)$, but in the below, by a slight abuse of notation, we will often use just C to denote the nerve.

  This covers all the examples discussed from Section 5.1 to 5.3.

- For a manifold $\mathcal{M}$, the collection $\Delta_m \mathcal{M}$ of singular simplicial m-cells (those which we use as the basis for singular m-chain in singular homology) for all m form a Kan complex $S\mathcal{M}$ internal to Difflg, with $S\mathcal{M}_m = \Delta_m \mathcal{M}$. This is more powerful than the strict path groupoid $\bar{P}_{n\to\infty}\mathcal{M}$, in that $S\mathcal{M}$ captures all the homotopy information of $\mathcal{M}$; in fact, $S\mathcal{M}$ is one way to realize the notion of the (fully fledged rather than strict) fundamental groupoid $\Pi_{n\to\infty}\mathcal{M}$ [99]. Likewise we can consider singular cubical m-cells.

- A cubic lattice with given branching structure [111] is naturally a cubical set with C_m the set of all m-dimensional cubes, and a lattice in the form of a simplicial complex with given branching structure is naturally a simplicial set with C_m the set of all m-dimensional simplices. We will denote the cubic/simplicial set as $\mathcal{L}$. Note that $\mathcal{L}$

---

[111]I.e. an ordering of vertices, which is needed when defining cup product.

is not the nerve of the strict d-category $\bar{\mathcal{L}}$ that we introduced before. In particular, $\mathcal{L}$ captures the full homotopy d-type information of the manifold that the lattice is discretizing, while the strict $\bar{\mathcal{L}}$ does not.

An interesting subtlety is noteworthy. $\mathcal{L}$ as defined above (say, in the simplicial case) is not a Kan complex, because two consecutive links on a simplicial complex might not be two edges of any triangle. As a consequence, $\mathcal{L}$ does not fully contain the nerve of $\bar{\mathcal{L}}$, as the later is a Kan complex—any two consecutive links can be composed to a lattice path in $\bar{\mathcal{L}}$ made of two links. Likewise for higher dimensional cells. Of course, we can enlarge $\mathcal{L}$ into a Kan complex that contains $\bar{\mathcal{L}}$, by including into $\mathcal{L}_1$ the lattice paths with more than one link, and into $\mathcal{L}_2$ the "degenerate triangles" representing the composition of paths, and so on. [112] But it turns out that we do not want such enlargement when we describe a lattice field configuration, i.e. a simplicial map from $\mathcal{L}$ to some target Kan simplicial manifold to be introduced below. Indeed, recall the explicit description in Section 4; the d.o.f. in the path integral only live on the actual lattice cells in $\mathcal{L}$ without the enlargement. We want to understand more deeply why this is the case in the future.

Before we move on, we can finally introduce the procedure of *geometric realization*, which has been mention before. Given a simplicial set C (or a category C, and then take its nerve), we can construct such a simplicial complex that the m-dimensional simplices are labeled by elements of C_m, and these simplices are geometrically glued together according to the face maps and degeneracy maps of C to form a topological space $|C|$, the geometric realization of C, which is in general infinite dimensional. [113] If C is a simplicial manifold to begin with, then the topology on C_m is also inherited onto the set of m-dimensional simplices, on top of the topology of each simplex (thus, the set of m-dimensional simplices, before the gluing, has a topology of $C_m \times$ (one m-dimensional simplex)).

One may note that the procedure of taking the geometric realization of a simplicial set/manifold and the procedure of taking the singular simplicial complex of a topological space seem to be some kind of inverse of each other. The former is a functor from the category of simplicial sets to the category of topological spaces, while the later is a functor form the category of topological spaces to the category of simplicial sets. In fact, these two functors are not inverses of each other in any sense; rather, the later functor is a *right adjoint* functor to the former, which is an important generalization of the notion of inverse. This means given any simplicial set C and any topological space X, the hom-set of simplicial maps from C to SX is isomorphic to the hom-set of continuous functions from $|C|$ to X. [114] While this is in general a crucial mathematical point, in the below we will only briefly mention this fact near the end of the section.

---

[112]This is to enlarge $\mathcal{L}$ by pushing out $\mathcal{L} \hookleftarrow \mathcal{L} \cap \bar{\mathcal{L}} \hookrightarrow \bar{\mathcal{L}}$ in the category of simplicial sets.

[113]Sometimes the geometric realization is only defined up to homotopy equivalence. For example, $|B\mathbb{Z}|$ constructed by this procedure is infinite dimensional, but it is homotopic to a circle, so we may as well say S^1 is a realization of $|B\mathbb{Z}|$. In most cases there are only infinite dimensional realizations.

[114]Adjunction is a central concept of category theory that we nevertheless did not introduce at length. An adjoint functor is more general than an inverse functor, and is motivated by adjoint operators familiar in linear algebra (indeed, a linear operator has adjoint even if it might not be invertible). The definition we give above is equivalent to the following definition. $D \xleftarrow{F} C$ and $C \xleftarrow{G} D$ form an adjoint pair (with G the left adjoint to F and F the right adjoint to G) if there are natural transformations $\mathbf{1}_C \xLeftarrow{\eta} G \circ F$ and

5.5   Topological refinement from higher anafunctor

Now we tackle our main problem. Everything discussed below will be internalized in Manifold, or Difflg when necessary. Let us think more closely what we really want when we say we want to define the skyrmion that counts $\pi_3 \cong \mathbb{Z}$. The continuum expression (52) of the skyrmion density in fact represents the generator of $H^3(\mathcal{T}; \mathbb{Z})$. [115] This is related to π_3 via the universal coefficient theorem and the Hurewicz theorem, which in the case of $\mathcal{T} = S^3$ imply $H^3(S^3; \mathbb{Z}) \xrightarrow{\sim} \mathrm{Hom}(H_3(S^3; \mathbb{Z}); \mathbb{Z})$ and $H_3(S^3; \mathbb{Z}) \xleftarrow{\sim} \pi_3(S^3)$ respectively. Now let us understand the topological classes from the perspective of anafunctor.

We begin with the simple case $H^1(S^1; \mathbb{Z}) \cong \mathbb{Z}$. Recall we can represent the basic $\mathbb{Z}$ bundle over S^1 as an anafunctor from S^1 to $B\mathbb{Z}$:

$$
\begin{array}{ccccc}
\mathbb{Z} & \longleftarrow & \mathbb{R} \times_{S^1} \mathbb{R} & \longrightarrow & S^1 \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & \mathbb{R} & \longrightarrow & S^1
\end{array}
\tag{110}
$$

where $\mathbb{R} \times_{S^1} \mathbb{R} \cong \mathbb{R} \times \mathbb{Z}$, and the $\mathbb{Z}$ here maps to the left identically. Anafunctors from a manifold $\mathcal{T}$ to $B\mathbb{Z}$ are classified by $H^1(\mathcal{T}; \mathbb{Z})$—here "classify" means up to ananatural isomorphism between anafunctors—which agrees with the usual classification of $\mathbb{Z}$ bundles. [116] And the particular anafunctor (110) realizes the generator of the classification $H^1(S^1; \mathbb{Z}) \cong \mathbb{Z}$.

This looks almost like the Villainzation process. From the Villain model we can observe that the Villainization process is described by the following 2-anafunctor (we may often omit the "2-" or "higher" in the below) from ES^1 to $B^2\mathbb{Z}$:

$$
\begin{array}{ccccc}
\mathbb{Z} & \longleftarrow & S^1 \times (\mathbb{R} \times_{S^1} \mathbb{R}) & \longrightarrow & S^1 \times S^1 \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & S^1 \times \mathbb{R} & \longrightarrow & S^1 \times S^1 \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & S^1 & \longrightarrow & S^1
\end{array} \quad .
\tag{111}
$$

The right column ES^1 is nothing but the target category used in traditional lattice nlσm. The left column $B^2\mathbb{Z}$ is the category characterizing the topological defects that we want to describe—the vortices on the plaquettes. The middle column, i.e. the span, is the desired target category to be used for the Villain model. (Also, as discussed below (91), this target category, i.e. the middle column in (111), has an ananatural equivalence to $BE\mathbb{Z}$. This $BE\mathbb{Z}$ then maps to the left column $B^2\mathbb{Z}$ by picking up the holonomies in the 2-morphisms.)

$\mathbf{1}_D \xRightarrow{\eta} F \circ G$, where ε and η are not necessarily invertible but only required to satisfy a weaker condition that under natural transformation composition, $(\mathbf{1}_F \circ_\mathsf{h} \varepsilon) \circ_\mathsf{v} (\eta \circ_\mathsf{h} \mathbf{1}_F) = \mathbf{1}_F$, and $(\varepsilon \circ_\mathsf{h} \mathbf{1}_G) \circ_\mathsf{v} (\mathbf{1}_G \circ_\mathsf{h} \eta) = \mathbf{1}_G$.

[115]More precisely, as mentioned in footnote 39, the continuum WZW curving and the transition functions introduced from (52) to (55) form the generator of the *Deligne-Beilinson double cohomology* $H^2_{\mathrm{DB}}(\mathcal{T}; U(1))$, where the double cochain has a de Rham exterior derivative coboundary operator and a Čech transition function coboundary operator. $H^2_{\mathrm{DB}}(\mathcal{T}; U(1))$ contains the topological classification information in $H^3(\mathcal{T}; \mathbb{Z})$ as well as the flat 2-holonomy information [36]. (The case of $H^1_{\mathrm{DB}}(X; U(1))$ is presented in details in e.g. [60].)

[116]Roughly speaking, the 1-cocycle condition corresponds to the fact that $B\mathbb{Z}$ only has identity 2-morphisms, and 1-coboundaries corresponds to ananatural isomorphisms

That is to say, we can extract from the Villain model what "topological refinement" means: *The target category of a "topologically refined theory" is the span of the suitable anafunctor that maps from the target category of the traditional theory to the category characterzing the topological defects that we want to describe.*

How is the Villainization anafunctor (111) related to the anafunctor (110) that realizes the generator of the classification $H^1(S^1; \mathbb{Z}) \cong \mathbb{Z}$? In this example, roughly speaking, we can spot that the (110) is "part of" (111) if in (111) we ignore the S^1 in the right and middle columns and ignore the bottom row—physically this amounts to fixing the S^1 d.o.f. on one vertex and looking at the physics on the links and plaquettes in the same connected component of the lattice. [117]

Notably, $H^1(S^1; \mathbb{Z}) \cong \mathbb{Z}$ does *not* classify anafunctors from ES^1 to $B^2\mathbb{Z}$, as such a classification would have been trivial because ES^1 is ananaturally equivalent to the trivial category $*$. Thus we are coming back to the important conceptual problem discussed in Section 5.1: Why we in general do not just treat ES^1 as trivial. We gave the physical explanation there. We now discuss the more formal aspect of it, which will in turn allow us to generalize the relation between (110) and (111) to generic Villainization procedures (this is non-trivial because for the universal cover $\widetilde{\mathcal{T}}$ of a generic $\mathcal{T}$, in general $\widetilde{\mathcal{T}}^2/\Gamma$ in (91) cannot be expressed as $\mathcal{T} \times \widetilde{\mathcal{T}}$, unlike when $\mathcal{T} = S^1$ where $\mathbb{R}^2/\mathbb{Z} \cong S^1 \times \mathbb{R}$, even though $\mathcal{T}^2/\Gamma|_{\text{fixing s or t}} \cong \widetilde{\mathcal{T}}$).

Let us think of not just $E\mathcal{T}$, but the inclusion functor $\mathcal{T} \hookrightarrow E\mathcal{T}$, where the objects map identically and the identity morphisms in $\mathcal{T}$ map to those in $E\mathcal{T}$. Physically, this means we have a well-defined notion of when the two d.o.f. across a lattice link "take the same value". [118] In this purview, let us look back at the ananatural equivalence between $E\mathcal{T}$ and $*$, equipping the latter with the trivial functor $\mathcal{T} \to *$. While any pair of functors between $*$ and $E\mathcal{T}$ establish their (ana)natural equivalence, if we impose the additional requirement that the equipped $\mathcal{T} \to *$ and $\mathcal{T} \hookrightarrow E\mathcal{T}$ must be preserved, then we can see there is no functor from $*$ to $E\mathcal{T}$ that respects this requirement (unless $\mathcal{T}$ itself is a single point). Related to this, while any functor from $E\mathcal{T}$ to $*$ preserves the equipped functor from $\mathcal{T}$ and is surjective and fully faithful, if we carefully inspect the bijection in the definition of "fully faithful", we find the map from $\{\mathbf{1}_*\}$ to $\{(a, b)\}$ for $a \neq b \in \mathcal{T}$ does not preserve the image of any identity 1-morphism in $\mathcal{T}$. Therefore, we conclude that *while $E\mathcal{T}$ and $*$ are equivalent as categories, they are inequivalent as categories equipped with the specified functors from $\mathcal{T}$.* We call categories with such specified functors from $\mathcal{T}$ "categories under $\mathcal{T}$". [119]

Now with this perspective in mind, we shall think of (111) as an "anafunctor under $\mathcal{T} = S^1$", i.e. with each column equipped with an obvious functor from $\mathcal{T} = S^1$ (which maps identically to the objects of the right and middles columns, and trivially to the left column)

---

[117] If the model has $U(1)$ global symmetry over S^1, it is natural and familiar to do so. But even if the global symmetry is slightly broken, the topological considerations here should not be spoiled.

[118] And the physical intuition for the lattice nlσm is that the link weight will be maximized there.

[119] In general, given a category C, the "under category" c/C (for some specified $c \in C_0$) is made of objects $(c/C)_0 := \{f \in C_1 | \mathsf{s}(f) = c\}$ and morphisms $(c/C)_1|_{f', f} := \{g \in C_1 | g \circ f = f'\}$ ("over category" C/c is if we replace s with t). This can be generalized to higher categories. Here, $\mathcal{T} \hookrightarrow E\mathcal{T}$ and $\mathcal{T} \to *$ are objects in the 2-category of Lie groupoids under $\mathcal{T}$, and there is no 1-morphism ($\mathcal{T}$-preserving anafunctor) between them.

that is being preserved along the horizontal functors in (111); moreover, any notion of "surjection" and "bijection" in the definition of anafunctor are also $\mathcal{T}$-preserving. Anafunctors as such are classified, up to ananatural isomorphisms under similar $\mathcal{T}$-preserving conditions, are classified by the relative cohomology $H^2(ES^1, S^1; \mathbb{Z}) \cong H^2(|ES^1|, S^1; \mathbb{Z})$. Since ES^1 is trivial (or say $|ES^1|$ is contractible), by the long exact sequence of relative cohomology

$$\cdots \to H^n(X, Y; A) \to H^n(X; A) \to H^n(Y; A) \to H^{n+1}(X, Y; A) \to H^{n+1}(X; A) \to \cdots \tag{112}$$

we have an isomorphism $H^2(ES^1, S^1; \mathbb{Z}) \overset{\sim}{\leftarrow} H^1(S^1;\ Z) \cong \mathbb{Z}$, hence explaining the relation to (110), and (111) realizes the generator of this classfication.

More generally, let $\widetilde{\mathcal{T}}$ be the universal cover of $\mathcal{T} = \widetilde{\mathcal{T}}/\Gamma$ with Γ is discrete, we have

$$
\begin{array}{ccccc}
\Gamma & \longleftarrow & \widetilde{\mathcal{T}} \times_{\mathcal{T}} \widetilde{\mathcal{T}} & \longrightarrow & \mathcal{T} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & \widetilde{\mathcal{T}} & \longrightarrow & \mathcal{T}
\end{array}
\tag{113}
$$

(where $\widetilde{\mathcal{T}} \times_{\mathcal{T}} \widetilde{\mathcal{T}} \cong \widetilde{\mathcal{T}} \times \Gamma$). If Γ is discrete and abelian, the Villainization procedure is described by such an anafunctor under $\mathcal{T}$

$$
\begin{array}{ccccc}
\Gamma & \longleftarrow & (\widetilde{\mathcal{T}}^2/\Gamma) \times \Gamma & \longrightarrow & \mathcal{T} \times \mathcal{T} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & \widetilde{\mathcal{T}}^2/\Gamma & \longrightarrow & \mathcal{T} \times \mathcal{T} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & \mathcal{T} & \longrightarrow & \mathcal{T}
\end{array}
\tag{114}
$$

(in general $\widetilde{\mathcal{T}}^2/\Gamma$ cannot be expressed as $\mathcal{T} \times \widetilde{\mathcal{T}}$, even though $\mathcal{T}^2/\Gamma|_{\text{fixing s or t}} \cong \widetilde{\mathcal{T}}$) realizing suitable generator of the classification $H^2(E\mathcal{T}, \mathcal{T}; \Gamma) \cong H^1(\mathcal{T}; \Gamma)$. For discrete non-abelian Γ, the left column that describes the topological defects shall be replaced by the 2-group $\text{InnAut}\Gamma \ltimes \Gamma \rightrightarrows \text{InnAut}\Gamma \rightrightarrows *$, where the inner automorphism InnAut arises because non-abelian holonomies are only well-defined up to conjugation. The cohomology classification will take value in this 2-group instead. These more complicated non-abelian cases have little to do with what we plan to introduce below.

When $\mathcal{T}$ in the above is itself a Lie group G and Γ is abelian, we can deloop the above and obtain the Villainized gauge theory. The classification becomes $H^3(BEG, BG; \Gamma) \overset{\sim}{\leftarrow} H^2(BG; \Gamma) \cong H^2(|BG|; \Gamma)$. If G itself is abelian, we can deloop the above arbitrary many times. This completes the discussion of general Villainization.

Next, consider $H^2(\mathcal{T}; \mathbb{Z})$; for this case we may keep $\mathcal{T} = S^2$ in mind as the basic example. $H^2(\mathcal{T}; \mathbb{Z})$ classifies the $U(1)$ bundles on $\mathcal{T}$. Often times (including our Sections 2 and 3) we would represent a $U(1)$ bundle as $U(1) \to \mathcal{E} \to \mathcal{T}$ where $\mathcal{E}$ is the total space, but this has two disadvantages: the map $U(1) \to \mathcal{E}$ is non-canonical, and moreover it obscures the relation between a $U(1)$ bundle and a $U(1)$ function. It is more natural to represent a $U(1)$ bundle

as an anafunctor

$$
\begin{array}{ccccc}
U(1) & \longleftarrow & \mathcal{E} \times_{\mathcal{T}} \mathcal{E} & \longrightarrow & \mathcal{T} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & \mathcal{E} & \longrightarrow & \mathcal{T}
\end{array}
\tag{115}
$$

where the map $U(1) \leftarrow \mathcal{E} \times_{\mathcal{T}} \mathcal{E} \cong \mathcal{E} \times U(1)$ represents the $U(1)$ action on $\mathcal{E}$ and is therefore canonical, and moreover this is obviously a "higher version" of a $U(1)$ function $U(1) \leftarrow \mathcal{T}$. [120] We can say a $U(1)$ function is a $U(1)$ 0-bundle, while a $U(1)$ bundle is a $U(1)$ 1-bundle. To relate this to $H^2(\mathcal{T}; \mathbb{Z})$, we deloop the anafunctor (110), and compose it on the left of this anafunctor, and obtain

$$
\begin{array}{ccccc}
\mathbb{Z} & \longleftarrow & \mathcal{E} \times \mathbb{R} \times \mathbb{Z} & \longrightarrow & \mathcal{T} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & \mathcal{E} \times \mathbb{R} & \longrightarrow & \mathcal{T} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & \mathcal{E} & \longrightarrow & \mathcal{T}
\end{array}
\quad .
\tag{116}
$$

Obviously we should call such an anafunctor a $\mathbb{Z}$ 2-bundle over $\mathcal{T}$.

Analogous to the "$\mathcal{T} \hookrightarrow E\mathcal{T}$ step" from (113) to (114), the target category to use for topological refinement would be the span of a suitable anafunctor under $\mathcal{T}$ from $E\mathcal{T}$ to $B^3\mathbb{Z}$:

$$
\begin{array}{ccccc}
\mathbb{Z} & \longleftarrow & (\mathcal{E}^2/U(1)) \times \mathbb{R} \times \mathbb{Z} & \longrightarrow & \mathcal{T} \times \mathcal{T} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & (\mathcal{E}^2/U(1)) \times \mathbb{R} & \longrightarrow & \mathcal{T} \times \mathcal{T} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & \mathcal{E}^2/U(1) & \longrightarrow & \mathcal{T} \times \mathcal{T} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & \mathcal{T} & \longrightarrow & \mathcal{T}
\end{array}
\quad ,
\tag{117}
$$

which is classified by $H^3(E\mathcal{T}, \mathcal{T}; \mathbb{Z}) \xleftarrow{\sim} H^2(\mathcal{T}; \mathbb{Z})$. For our familiar example $\mathcal{T} = S^2$, the $B^3\mathbb{Z}$ on the left represents the hedgehog defects; we have $\mathcal{E} = SU(2)$ and $SU(2)^2/U(1) \cong S^2 \times SU(2)$, [121] and the anafunctor realizes the generator of $H^3(ES^2, S^2; \mathbb{Z}) \xleftarrow{\sim} H^2(S^2; \mathbb{Z}) \cong \mathbb{Z}$. (Also, as discussed below (92), the target category of the refined nlσm, i.e. the middle column above, has an ananatural equivalence to $B^2 E\mathbb{Z}$, which in turn maps to the left column $B^3\mathbb{Z}$ by picking up the holonomies in the 3-morphisms.)

We can also arrive at (117) from (115) with an interchanged order of steps. From (115) we can first perform the "$\mathcal{T} \hookrightarrow E\mathcal{T}$ step" from (113) to (114) but with the discrete Γ replaced

---

[120] Also, as we explained in Section 5.2, in this formulation the span does not have to use the total space $\mathcal{E}$; it can as well use other coverings of $\mathcal{T}$, such as the disjoint union of charts $\mathcal{U}$. This unification of different ways of representing a principal bundle is yet another advantage of the anafunctor presentation.

[121] Where $(\mathcal{U}', \mathcal{U}) \in SU(2)^2$ is mapped to $(R_{\mathcal{U}}\hat{z}, \mathcal{U}'\mathcal{U}^{-1}) \in S^2 \times SU(2)$, which is invariant under $\mathcal{U}'e^{i\psi\sigma^z}, \mathcal{U}e^{i\psi\sigma^z}$ for $e^{i\psi} \in U(1)$.

with $U(1)$, arriving at

$$
\begin{array}{ccccc}
U(1) & \longleftarrow & (\mathcal{E}^2/U(1)) \times U(1) & \longrightarrow & \mathcal{T} \times \mathcal{T} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & \mathcal{E}^2/U(1) & \longrightarrow & \mathcal{T} \times \mathcal{T} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & \mathcal{T} & \longrightarrow & \mathcal{T}
\end{array}
\qquad (118)
$$

which corresponds to the first step of spinon decomposition in Section 2.4 and the $B^2 U(1)$ describes the Berry curvature on plaquettes. Then we compose on its left the twice delooping of (110), which corresponds to the second step that Villainizes the Berry curvature in Section 2.4, to arrive at (117).

After these discussions, it becomes obvious that if we want $\mathcal{T} = S^3$ or $\mathcal{T} = |SU(N)|$ and capture the physics due to $H^3(\mathcal{T}; \mathbb{Z})$, we shall begin with a $U(1)$ 2-bundle on $\mathcal{T}$, i.e. an anafunctor

$$
\begin{array}{ccccc}
U(1) & \longleftarrow & L \times_{Y \times_{\mathcal{T}} Y} L & \longrightarrow & \mathcal{T} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & L & \longrightarrow & \mathcal{T} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & Y & \longrightarrow & \mathcal{T}
\end{array}
\qquad (119)
$$

In particular, we will let Y be a surjective submersion covering $\mathcal{T}$, so that $Y \times_{\mathcal{T}} Y$ exists as a manifold, and L is the total space of a $U(1)$ bundle over $Y \times_{\mathcal{T}} Y$, and thus $U(1) \leftarrow L \times_{Y \times_{\mathcal{T}} Y} L \cong L \times U(1)$ is the $U(1)$ action on L. Such a construction for $U(1)$ 2-bundle over $\mathcal{T}$ is called a *bundle gerbe* over $\mathcal{T}$ [37] (with basic idea from [107]); the Lie groupoid part $L \rightrightarrows Y$ in the span is the analogue of the total space $\mathcal{E}$ for a $U(1)$ 1-bundle (115). (Just like a $U(1)$ bundle can be presented as $U(1) \to \mathcal{E} \to \mathcal{T}$ but with the map from $U(1)$ to $\mathcal{E}$ non-canonical, a $U(1)$ 2-bundle can also be presented in a similar way, except each entry becomes a Lie groupoid, i.e. there is a non-canonical anafunctor from $BU(1)$ to $L \rightrightarrows Y$ and then the projection from $L \rightrightarrows Y$ to $\mathcal{T}$.)

Just like the equivalence of two 1-bundles (115) is established by an invertible ananatural transformation, the same is true for 2-bundles. More explicitly, to establish the equivalence of two 1-bundles (115), we use an invertible ananatural transformation (90) that involves a function

$$
\begin{array}{l}
BU(1)_1 = U(1) \\
\qquad \qquad \nwarrow \\
\qquad \qquad \qquad H_0 = \mathcal{E} \times_{\mathcal{T}} \mathcal{E}'
\end{array}
\qquad (120)
$$

meaning that two equivalent $U(1)$ bundles only differ by a $U(1)$ function (which in physics is just gauge transformation). Now, to establish the equivalence of two 2-bundles (119), we use an invertible 2-ananatural transformation that involves an anafunctor (instead of functor, in

general)

$$
\begin{array}{ccc}
B^2U(1)_2 = U(1) & & \\
\Downarrow & \nwarrow & \\
B^2U(1)_1 = * & & (\Sigma \times_{\mathcal{T}} \Sigma) \times_{Y \times_{\mathcal{T}} Y' \times_{\mathcal{T}} Y \times_{\mathcal{T}} Y'} (L \times_{\mathcal{T}} L') \longrightarrow L \times_{\mathcal{T}} L' \\
& \nwarrow & \Downarrow \qquad\qquad\qquad\qquad \Downarrow \\
& & \Sigma \qquad\qquad\qquad\qquad \longrightarrow Y \times_{\mathcal{T}} Y'
\end{array}
\qquad (121)
$$

which means two equivalent $U(1)$ 2-bundles only differ by a $U(1)$ 1-bundle Σ over $Y \times_{\mathcal{T}} Y'$, and moreover two Σ's along with L and L' can together "piece up" to a trivialized $U(1)$ bundle over $Y \times_{\mathcal{T}} Y' \times_{\mathcal{T}} Y \times_{\mathcal{T}} Y'$. (By "piece up" we mean the two Σ's and L and L' each pulls-back to a $U(1)$ bundle over $Y \times_{\mathcal{T}} Y' \times_{\mathcal{T}} Y \times_{\mathcal{T}} Y'$, and then we take the tensor product of these four $U(1)$ bundles into one $U(1)$ bundle. The same when we say "piece up" of $U(1)$ bundles in the below.) We can illustrate this as

$$ (122) $$

where y_1, y_2, y_1', y_2' all project to a same element in $\mathcal{T}$, and a $U(1)$ value specifying the trivialization (the upper left of (121)) is assigned to the quadrangle as a function of $\sigma_1, \sigma_2, \ell, \ell'$. This anafunctor (121) in the 2-ananatural transformation is often called a *stable isomorphism* between the two bundle gerbes [108].

Originally, in [37] Y was restricted to be a fibre bundle over $\mathcal{T}$, and with such restriction it can be proven that, in order for the principal 2-bundle to be non-trivial in $H^3(\mathcal{T}; \mathbb{Z})$, [122] Y must be infinite dimensional (hence internalized in Difflg rather than Manifold). One such example is the *tautological bundle gerbe*, with $Y = \bar{\mathcal{P}}_* \mathcal{T}$, $Y \times_{\mathcal{T}} Y \cong \bar{\mathcal{P}}_* \mathcal{T} \times \bar{\Omega}_* \mathcal{T}$, and $L \cong \bar{\mathcal{P}}_* \mathcal{T} \times \left(\bar{\mathcal{P}}_* \bar{\Omega}_* \mathcal{T} \times U(1)/WZW \right)$ in the sense of (51), which realizes the generator of $H^3(\mathcal{T}; \mathbb{Z})$. [123] (The identity map from Y to L makes use of the naturally trivial element of $\left(\bar{\mathcal{P}}_* \bar{\Omega}_* \mathcal{T} \times U(1)/WZW \right)$, i.e. given a path in Y, we take the trivial surface in $\bar{\mathcal{P}}_* \bar{\Omega}_* \mathcal{T}$ sweeping from this path to itself, and take the identity in $U(1)$.) This puts (51), which we would expect from continuum QFT, in the present context, and note that this can be obtained from the strict 2-groupoid $\bar{\mathcal{P}}_2 \mathcal{T} \times U(1)/WZW \rightrightarrows \bar{\mathcal{P}} \mathcal{T} \rightrightarrows \mathcal{T}$ (which describes (51) as we said in Section 5.1) by fixing a source object (the starting point of the paths).

Later, Y that are more general surjective submersions covering $\mathcal{T}$ (as opposed to having to be fibre bundles over $\mathcal{T}$) have been considered, and such Y can be finite dimensional. This explains Section 3. A particularly nice choice of finite dimensional Y for $\mathcal{T} = S^3 \cong |SU(2)|$ is $Y = (SU(2) \backslash \{-\mathbf{1}\}) \sqcup (SU(2) \backslash \{+\mathbf{1}\})$; then $Y \times_{\mathcal{T}} Y$ has four patches, given by $Y \times_{\mathcal{T}} Y =$

---

[122]Non-trivial means the bundle gerbe $(L \rightrightarrows Y)$ is "non-exact", which means it is not stably isomorphic to any $(L' \rightrightarrows Y')$ such that the $U(1)$ bundle L' over $Y' \times_{\mathcal{T}} Y'$ is formed by piecing up (i.e pulling-back and then taking tensor product) two copies of some $U(1)$ bundle E' over Y'.

[123]It does not matter whether we take identification under thin homotopy or not. We can as well take $Y = \mathcal{P}_* \mathcal{T}$, since the WZW evaluation over a surface will be the same.

$(SU(2)\backslash\{-\mathbf{1}\}) \sqcup (SU(2)\backslash\{\pm\mathbf{1}\}) \sqcup (SU(2)\backslash\{\pm\mathbf{1}\}) \sqcup (SU(2)\backslash\{+\mathbf{1}\})$, and L is the $U(1)$ bundle over it such that, over the $(SU(2)\backslash\{-\mathbf{1}\})$ and $(SU(2)\backslash\{+\mathbf{1}\})$ patches which are topologically trivial, the $U(1)$ bundle is necessarily trivial, while over the $(SU(2)\backslash\{\pm\mathbf{1}\}) \cong S^2 \times [0,1]$ patches, the $U(1)$ fibre forms S^3 over the S^2 [89]. [124] [125] (The identity map from Y to L is choosing a trivialization section over the $(SU(2)\backslash\{-\mathbf{1}\}) \sqcup (SU(2)\backslash\{+\mathbf{1}\})$ part of $Y \times_\mathcal{T} Y$.) Why this choice is "nice" and why, in our actual construction in Section 4, we used a slightly more complicated Y (with an extra S^2 multiplied to the second patch) will be discussed later. This finite dimensional bundle gerbe is stably isomorphic to the infinite dimensional tautological bundle gerbe; we will demonstrate and make use of the stable isomorphism below (the stable isomorphism will basically follow from our geometrical interpretation of Y in Section 4.1).

This completes our introduction to the $U(1)$ bundle gerbe, or say $U(1)$ principal 2-bundle, (119), that is the higher analogue of (115). Based on the experience of the previous examples, clearly there are two follow-up steps in order to arrive at the desired topological refinement: the "$\mathcal{T} \hookrightarrow E\mathcal{T}$ step", and the "Villainization step", and the order of these two steps is interchangeable. For instance, starting from (115), if we first perform the Villainization step and then the $\mathcal{T} \hookrightarrow E\mathcal{T}$ step, we will go through (116) before arriving at (117); if we first perform the $\mathcal{T} \hookrightarrow E\mathcal{T}$ step and then the Villainization step, we will go through (118) before arriving at (117). In the below, starting with (119), we will proceed with the latter order, because this is closer to how we construct the actual physical model, and also makes closer connection to the existing mathematical literature, as we shall see now.

It turns out performing the "$\mathcal{T} \hookrightarrow E\mathcal{T}$ step" to the $U(1)$ 2-bundle (119) is crucially more non-trivial than doing the same to the $U(1)$ 1-bundle (115). Doing it to (115) will lead to (118) which is still an anafunctor of strict groupoids, while doing it to (119) will in general necessarily lead to an anafunctor of simplicial manifolds, i.e. smooth Kan complexes, (128). This is what we shall explain now. For simplicity, in the below we will focus on $\mathcal{T} = |G|$ for connected and simply connected semi-simple Lie group G, primarily with $|SU(N)|$ or more specifically $S^3 \cong |SU(2)|$ in mind. This is sufficient for our primary purpose of this paper.

A closely related problem has been well-studied in the literature: When $\mathcal{T}$ is the space of such a Lie group G, for example $SU(2)$, what is the delooping of (119)? Since delooping is roughly speaking appending a row below (119) with a single object $*$, this problem is obviously related to our "$\mathcal{T} \hookrightarrow E\mathcal{T}$ step" of interest. For instance, if we deloop (110) where $S^1 \cong |U(1)|$, the result is almost the same as (111), except in (111) we need to fix a source object in S^1 (fixing $e^{i\theta}$ in $(e^{i\theta'}, e^{i\theta}) \in S^1 \times S^1$, the space of the remaining d.o.f. can be naturally denoted by $e^{i(\theta'-\theta)} \in U(1)$); the physical meaning of this has been explained in the discussion below (111).

The delooping of (119) with $\mathcal{T} = |G|$ is known as a *multiplicative bundle gerbe* [36]. It is

---

[124]The non-trivial bundle part of L can be interpreted as the following [89]. When $g \neq \pm\mathbf{1} \in SU(2)$, the diagonalization $g = \mathcal{U}e^{i\lambda\sigma^z}\mathcal{U}^{-1}$ is non-degenerate, so $\mathcal{U} \in SU(2)$ is well-defined up to a $e^{i\kappa\sigma^z} \in U(1)$ action on the right, parametrizing an S^2. The $SU(2) \ni \mathcal{U}$ is the desired $U(1)$ bundle over the S^2. (Note the $SU(2) \ni \mathcal{U}$ is not the $SU(2) \ni g$ that we started with.)

[125]The generalization to $\mathcal{T} \cong SU(N > 2)$ in terms of the Weyl alcove is as explained at the end of Section 4.1.

an anafunctor but with a simplicial manifold as span:

$$
\begin{array}{ccccc}
U(1) & \longleftarrow & \Lambda^{(4)} & \longrightarrow & G^3 \\
\Downarrow & & \Downarrow\!\Downarrow\!\Downarrow\!\Downarrow & & \Downarrow\!\Downarrow\!\Downarrow\!\Downarrow \\
* & \longleftarrow & \Lambda & \longrightarrow & G^2 \\
\Downarrow & & \Downarrow\!\Downarrow\!\Downarrow & & \Downarrow\!\Downarrow\!\Downarrow \\
* & \longleftarrow & Y & \longrightarrow & G \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & * & \longrightarrow & *
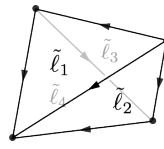\end{array}
\qquad (123)
$$

Here the right column is just the category BG presented as a simplicial manifold by taking the nerve (with triangular 2-cells being the group composition $\circ$, which forms $(G \times G) \times_G^{\circ,\mathrm{id}} G \cong G^2$, and so on). In the middle column, Λ is the manifold of all triangular shaped 2-cells, and is a $U(1)$ bundle over the triangular loop $(Y \times Y)_G^{\circ\Pi^2,\Pi} Y$ (where we have denoted the covering $Y \to G$ as Π, and $(Y \times Y)_G^{\circ\Pi^2,\Pi} Y$ is the submanifold of Y^3 that satisfies the G composition rule after the Π projection), representing the non-unique multiplicative structure on Y. We will explain later how to construct Λ given L from (119). $\Lambda^{(4)}$ is the manifold of all tetrahedral shaped 3-cells formed by four triangular shaped 2-cells:

$$
\Lambda^{(4)} := \left((\Lambda \times_Y^{\partial_2,\partial_2} \Lambda) \times_{Y \times Y}^{(\partial_1,\partial_1),(\partial_2,\partial_1)} \Lambda \right) \times_{Y \times Y \times Y}^{(\partial_0,\partial_0,\partial_0),(\partial_2,\partial_1,\partial_0)} \Lambda
\qquad (124)
$$

$$
\cong \left((\Lambda \times_Y^{\partial_2,\partial_2} \Lambda) \times_{Y \times Y}^{(\partial_1,\partial_1),(\partial_2,\partial_1)} \Lambda \right) \times U(1) .
\qquad (125)
$$

The second line expresses the fact that the four $U(1)$ bundles Λ can together piece up to a trivialized $U(1)$ bundle over the manifold of the six 1-cells around the tetrahedral 3-cell, and the trivialization is specified by the map from $\Lambda^{(4)}$ to $U(1)$,

$$
\text{lattice WZW curvature} \longleftarrow
\qquad (126)
$$

which will later be interpreted as assigning the lattice WZW curvature.

Why it becomes necessary here for the span to be a simplicial manifold in general? Since we are delooping (119), we want to introduce some notion of composition on Y, as well as some notion of horizontal composition on L (the original composition of L becomes the vertical composition). This corresponds to an ordinary functor $(\circ_h \rightrightarrows \circ)$ from $(L \rightrightarrows Y)^2$ to $(L \rightrightarrows Y)$. But to capture the general interesting cases, we should consider anafunctors, and this leads to the use of simplicial manifold, as explained in Section 5.4. The globular shaped 2-cells in L can be viewed as special kind of triangular shaped 2-cell in Λ. The $(\Lambda \overset{\rightarrow}{\rightrightarrows} Y \rightrightarrows *)$ part of the span is a generalized notion of Lie 2-group, modeled as a simplicial manifold—it has a single object and moreover satisfies the "invertibility" condition, i.e. the Kan condition. (And (123) can be rephrased as a Lie 2-group extension of BG—which is a Lie group, i.e. a Lie 1-group, a special case of Lie 2-group—by $B^2U(1)$, i.e. there is an

anafunctor from $B^2U(1)$ to $(\Lambda \rightrightarrows Y \rightrightarrows *)$, and then the projection to BG, forming a short exact sequence in a suitable sense [35]. [126])

Multiplicative bundle gerbes (123), or say Lie 2-group extensions by $B^2U(1)$, are classified by the cohomology $H^4(BG;\mathbb{Z}) \cong H^4(|BG|;\mathbb{Z})$ [35, 36], [127] manifested by Villainizing it, i.e. composing on its left the trice delooping of (110). This classification maps to the $H^3(\mathcal{T} = |G|;\mathbb{Z})$ that classifies (119) by forgetting about the multiplicative structure. (At the level of the ordinary singular cohomology, $H^4(|BG|;\mathbb{Z})$ maps to $H^3(\mathcal{T} = |G|;\mathbb{Z})$ by *transgression*, which is familiar in the CS and WZW context [36]. [128]) For $G = SU(N)$, the transgression is an isomorphism, $H^4(|BG|;\mathbb{Z}) \xrightarrow{\sim} H^3(\mathcal{T} = G;\mathbb{Z}) \cong \mathbb{Z}$. On the other hand, the relation of $H^4(|BG|;\mathbb{Z})$ to the group cohomology $H^3_{\text{group}}(G;U(1)) \cong H^4_{\text{group}}(G;\mathbb{Z})$ will be discussed later.

What we really want for the topologically refined lattice model is slightly different from the multiplicative bundle gerbe (123). Instead of delooping (119), we want to perform the "$\mathcal{T} \hookrightarrow E\mathcal{T}$ step", which should lead to an anafunctor from $E\mathcal{T}$ (equipped with the inclusion from $\mathcal{T}$) to $B^3U(1)$.

If we do not mind using infinite dimensional d.o.f. and had used the tautological bundle gerbe for (119), then the "$\mathcal{T} \hookrightarrow E\mathcal{T}$ step" would give us

$$
\begin{array}{ccccc}
U(1) & \longleftarrow & (\bar{\mathcal{P}}_2\mathcal{T} \times U(1)/WZW) \times U(1) & \longrightarrow & \mathcal{T} \times \mathcal{T} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & \bar{\mathcal{P}}_2\mathcal{T} \times U(1)/WZW & \longrightarrow & \mathcal{T} \times \mathcal{T} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & \bar{\mathcal{P}}\mathcal{T} & \longrightarrow & \mathcal{T} \times \mathcal{T} \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & \mathcal{T} & \longrightarrow & \mathcal{T}
\end{array}
\quad , \tag{127}
$$

[129] where fixing a source (or target) object gives $\bar{\mathcal{P}}\mathcal{T}|_{\text{fixing s}} = \bar{\mathcal{P}}_*\mathcal{T} = Y'$ and $(\bar{\mathcal{P}}_2\mathcal{T} \times U(1)/WZW)|_{\text{fixing ss}} \cong \bar{\mathcal{P}}_*\mathcal{T} \times (\bar{\mathcal{P}}\bar{\Omega}_*\mathcal{T} \times U(1)/WZW) = L'$, reducing to the tautological bundle gerbe. This is intuitive if we have a continuum nlσm and think of the lattice as being embedded in the continuum, which is to say, this is the categorical presentation of the

---

[126] More systematically, consider the ambient bicategory formed by Lie groupoids, anafunctors and ananatural transformations (which are automatically invertible since morphisms in Lie groupoids are invertible in a suitable sense). A Lie 2-group is a 2-group internalized in this ambient bicategory [35], hence compositions are in general given by anafunctors. A Lie 2-group extension is a short exact sequence in this ambient bicategory.

[127] This is related to the *bundle 2-gerbe* on $|BG|$ that we will introduce later [36].

[128] The transgression map is constructed in the following intuitive way. A singular n-cochain ϕ on $|BG|$ can be pulled-back to an n-cochain φ in $|EG|$. Since $|EG|$ is contractible, $H^n(|EG|;A)$ must be trivial, so $\varphi = d\rho$ for some $(n-1)$-cochain ρ in $|EG|$. Restricting ρ to some fibre $F \cong |G|$, we have some $(n-1)$-cochain ϱ in $|G|$ that satisfies $d\varrho = 0$, since $d\rho = \varphi$ is constant on the fibre F. Thus ϱ defines a class in $H^{n-1}(G;A)$.

The transgression map can be refined to a map from $H^3_{\text{DB}}(|BG|;U(1))$ to $H^2_{\text{DB}}(G;U(1))$ [36], which is essentially the familiar calculation of how the gauge transformation of CS 3-form is given by WZW curving 2-form, with the transition functions suitably taken care of (as briefly mentioned below (55)).

[129] Here $(\bar{\mathcal{P}}_2\mathcal{T} \times U(1)/WZW) \times U(1)$ can be recognized as $(\bar{\mathcal{P}}_2\mathcal{T} \times U(1)/WZW) \times_{\bar{\mathcal{P}}\mathcal{T} \times_{\mathcal{T}^2} \bar{\mathcal{P}}\mathcal{T}} (\bar{\mathcal{P}}_2\mathcal{T} \times U(1)/WZW)$ or as $\bar{\mathcal{P}}_3\mathcal{T} \times U(1)^2/WZW^2$. In particular, the last $U(1)$ in $(\bar{\mathcal{P}}_2\mathcal{T} \times U(1)/WZW) \times U(1)$ is the WZW integral over the volume swiped out by the element in $\bar{\mathcal{P}}_3\mathcal{T}$.

idea introduced around (51). The left column $B^3 U(1)$ is nothing but the continuum WZW curvature over a 3d region. [130] Note this is still an anafunctor of strict higher categories.

However, for an actual lattice model, we want the d.o.f. to be finite dimensional, i.e. we want to use a bundle gerbe (119) with a finite dimensional Y. Then, just like in the delooping problem (123), in general the "$\mathcal{T} \hookrightarrow E\mathcal{T}$ step" will lead to an anafunctor of simplicial manifolds, of the form

$$
\begin{array}{ccccc}
U(1) & \longleftarrow & \mathcal{T} \times \Lambda^{(4)} & \longrightarrow & \mathcal{T} \times G^3 \\
\Downarrow & & \Downarrow\Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow\Downarrow \\
* & \longleftarrow & \mathcal{T} \times \Lambda & \longrightarrow & \mathcal{T} \times G^2 \\
\Downarrow & & \Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow \\
* & \longleftarrow & \mathcal{T} \times Y & \longrightarrow & \mathcal{T} \times G \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & \mathcal{T} & \longrightarrow & \mathcal{T}
\end{array}
\tag{128}
$$

Here our target space $\mathcal{T} \cong |G|$ for some connected and simply connected semi-simple Lie group G, primarily $|SU(N)|$ or more specifically $S^3 \cong |SU(2)|$. [131] The right column is the target category $E\mathcal{T} \cong EG$ used in traditional lattice nlσm, presented as a simplicial manifold by taking the nerve, and as in Section 4.1, we represented $(g_1, g_2) \in \mathcal{T}^2$ as $(g_1, g_2 g_1^{-1}) \in \mathcal{T} \times G$, $(g_1, g_2, g_3) \in \mathcal{T}^3$ as $(g_1, g_2 g_1^{-1}, g_3 g_2^{-1}) \in \mathcal{T} \times G^2$, and so on. The left column $B^3 U(1)$ is nothing but the lattice WZW curvature e^{idW_c}. The middle column is what we want for the target category of the topologically refined lattice nlσm, before the last simple step of Villainization (i.e. composing the trice delooping of (110) on the left of (128)). The construction of Λ in (128) does not have to be exactly the same in details as that in (123), [132] but they need to be topologically equivalent in the sense that we want both (128) and (123) to implement the second Chern class—more particularly, anafunctors of the form (128) are classified by $H^4(E\mathcal{T}, \mathcal{T}; \mathbb{Z}) \overset{\sim}{\longleftarrow} H^3(\mathcal{T}; \mathbb{Z}) \overset{\sim}{\longleftarrow} H^4(BG; \mathbb{Z}) \cong \mathbb{Z}$ (where the latter classifies (123)), and we want to realize the generator of it.

Now we come to the crucial technical point of how to construct the multiplicative structure Λ in (128), given a finite dimensional bundle gerbe $(L \rightrightarrows Y)$ from (119).

It is, again, helpful to begin with what we expect from the continuum QFT, (127), which, although involving infinite dimensional d.o.f., gives us the crucial intuition of what we really want. We first present (127) as an anafunctor of simpilicial manifolds by taking the nerve; the space of 2-cells form $\Delta_2 \mathcal{T} \times U(1)/WZW$, where elements of $\Delta_2 \mathcal{T}$ are (singular) 2-chains in $\mathcal{T}$. Then we can induce a multiplicative structure on the finite dimensional $(L \rightrightarrows Y)$ using the stable isomorphism between this $(L \rightrightarrows Y)$ and the tautological bundle gerbe $(L' \rightrightarrows Y')$.

---

[130] In the previous examples, (111) is related to the continuum loop space since $S^1 \times \mathbb{R} \cong \bar{\mathcal{P}} S^1$, i.e. (45), and (117) is related to the continuum loop space since $S^2 \times SU(2) \cong \bar{\mathcal{P}} S^2 / Berry$, i.e. (44) and (46).

[131] The cases of more general $\mathcal{T}$ will be studied in future works. A particularly physically relevant case is when $\mathcal{T} = S^2$ and we want to capture both the π_2 and π_3 physical effects.

[132] We will comment on the detailed difference later, after we introduce the construction of Λ in (128) below.

The general idea is illustrated as
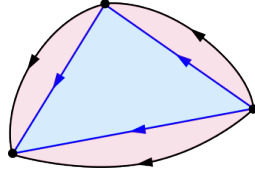
$$(129)$$

where each 0-cell is an element from $\mathcal{T} \cong |G|$, each black 1-cell is from $\mathcal{T} \times Y$, each blue 1-cell is from $\bar{\mathcal{P}}\mathcal{T}$ (which is a $Y' = \bar{\mathcal{P}}_* \mathcal{T}$ bundle over $\mathcal{T}$), whose "multiplicative structure" is represented by the blue 2-cell from $\Delta_2 \mathcal{T} \times U(1)/WZW$, and each pink 2-cell between the black and blue 1-cell is from (a suitable pullback of) Σ, the stable isomorphism between Y and Y'. Since the space of each 2-cell in the above is a $U(1)$ bundle over the space of the 1-cells on around it, the spaces of the four 2-cells together can piece up to a $U(1)$ bundle over the space of the three black 1-cells, and this will be identified as the desired $\mathcal{T} \times \Lambda$. [133] In practice, the construction is simpler. When the groupoid $(L \rightrightarrows Y)$ can be embedded in $(L' \rightrightarrows Y')$ by an ordinary functor while preserving the map from Y to $\mathcal{T}$, the stable isomorphism in pink 2-cells are just the pull-back of L' along the embedding. As a result, the pull-back of the $U(1)$ bundle $\Delta_2 \mathcal{T} \times U(1)/WZW$ over the space of the blue triangular loop $(\bar{\mathcal{P}}\mathcal{T} \times_{\mathcal{T}}^{\mathsf{s,t}} \bar{\mathcal{P}}\mathcal{T}) \times_{\mathcal{T}^2}^{(\mathsf{s,t}),(\mathsf{s,t})} \bar{\mathcal{P}}\mathcal{T}$ along the embedding directly gives rise to the desired $U(1)$ bundle $\mathcal{T} \times \Lambda$ over the space of the black triangular loop $\mathcal{T} \times ((Y \times Y) \times_{\mathcal{T}} Y)$. Obviously, the "geometrical interpretation of Y" described in details in Section 4.1 is nothing but such an embedding of $\mathcal{T} \times Y$ into $\bar{\mathcal{P}}\mathcal{T}$.

We almost finished the explanation of our construction in Section 4.1, except we still need to explain why we do not completely follow [89] which uses $Y = (SU(2)\backslash\{-\mathbf{1}\}) \sqcup (SU(2)\backslash\{+\mathbf{1}\})$, but use $Y = (SU(2)\backslash\{-\mathbf{1}\}) \sqcup (SU(2)\backslash\{+\mathbf{1}\} \times S^2)$ instead. While we can establish a stable isomorphism between $(SU(2)\backslash\{-\mathbf{1}\}) \sqcup (SU(2)\backslash\{+\mathbf{1}\})$ and $\mathcal{P}_* S^3$ thanks to the embedding, there is no canonical choice for the embedding of $(SU(2)\backslash\{-\mathbf{1}\}) \sqcup (SU(2)\backslash\{+\mathbf{1}\})$ into $\mathcal{P}_* S^3$, hence no canonical choice for the stable isomorphism—in particular for the upper left trivialization arrow in (121). [134] On the other hand, for $(SU(2)\backslash\{-\mathbf{1}\}) \sqcup$

---

[133]Similar idea can be applied to multiplicative bundle gerbe (123). This is the transgression-regression method introduced in [90]. We will remark about this in the below.

[134]To understand this, it may be helpful to consider the simpler case of $U(1)$ bundle over S^2. We can present the same $U(1)$ bundle using three different but equivalent anafunctors (115): the total space $SU(2)$ (with $SU(2) \times_{S^2} SU(2) \cong SU(2) \times U(1)$ specifying the $U(1)$ action on the fibre), the pointed path space $\mathcal{P}_* S^2$ (with $\mathcal{P}_* S^2 \times_{S^2} \mathcal{P}_* S^2 \cong \mathcal{P}_* S^2 \times \Omega_* S^2$ mapping to $U(1)$ by the area, or say Berry phase, bounded by the loop in $\Omega_* S^2$), or the patches $\mathcal{U} = (S^2\backslash\{-\hat{z}\}) \sqcup (S^2\backslash\{+\hat{z}\})$ (with a $U(1)$ transition function of winding number 1). For our current purpose, let us focus on the patches $\mathcal{U}$ versus the pointed path space $\mathcal{P}_* S^2$. There is no canonical choice for the transition function in the $\mathcal{U}$ anafunctor, for example we can choose where the pre-image of $1 \in U(1)$ lies. Accordingly, the ananatural isomorphism (120) between the $\mathcal{U}$ anafunctor and the $\mathcal{P}_* S^2$ anafunctor is also non-canonical. The intuitive understanding is, while the $(S^2\backslash\{-\hat{z}\})$ patch can be canonically embedded in $\mathcal{P}_* S^2$ as the shortest geodesic from $+\hat{z}$ to the desired point in $(S^2\backslash\{-\hat{z}\})$, the $(S^2\backslash\{-\hat{z}\})$ patch does not have a canonical embedding; we may consider a non-canonical embedding where we first go from $\hat{z}$ to $-\hat{z}$ along some fixed longitude—whose choice is non-canonical—and then go from $-\hat{z}$ to the desired point in $(S^2\backslash\{+\hat{z}\})$ via the shortest geodesic. However, if we use $(S^2\backslash\{-\hat{z}\}) \sqcup (S^2\backslash\{+\hat{z}\} \times S^1)$ instead, where the extra S^1 labels which longitude to use, then the embedding becomes rotationally covariant; in terms of the transition function, the S^1 specifies which longitude is the pre-image of $1 \in U(1)$.

$(SU(2)\backslash\{+\mathbf{1}\} \times S^2)$, the geometrical interpretation in Section 4.1 is a "good" embedding into $\mathcal{P}_*S^3$, leading to a "good" stable isomorphism. What we mean by "good or not" is that, $(SU(2)\backslash\{-\mathbf{1}\}) \sqcup (SU(2)\backslash\{+\mathbf{1}\})$ is like $(SU(2)\backslash\{-\mathbf{1}\}) \sqcup (SU(2)\backslash\{+\mathbf{1}\} \times S^2)$ with the $\hat{n} \in S^2$ d.o.f. fixed to some given direction, hence the embedding into $\mathcal{P}_*S^3$ is not rotationally covariant; while with the extra S^2 labeling this direction, the embedding becomes rotationally covariant. (For [89], the multiplicative structure and the stable isomorphism to tautological bundle gerbe were not part of the consideration. While for [90], the stable isomorphism to the tautological bundle gerbe was the main point, but being canonical, more particularly rotationally covariant, was not part of the consideration. But for our application we would like there to be physically natural—covariant under global symmetry—stable isomorphism, that induces the multiplicative structure from the continuum tautological bundle gerbe onto the finite dimensional bundle gerbe.) [135]

This is the derivation for our construction in Section 4.1—as long as we compose the trice delooping of (110) on the left of (128), completing the last simple step of Villainzation. (Of course, here we described the derivation in terms of simplicial manifolds, while in Section 4.1 we used cubical manifolds because the lattice is cubic, but the idea and the essential topological information are the same.)

Let us make two further remarks in relation to the mathematical literature.

- The first remark is on the detailed difference between our construction of the multiplicative structure Λ in (128) versus the construction in [90] of that in (123), although both implement the generator of $H^4(E\mathcal{T}, \mathcal{T}; \mathbb{Z}) \cong H^4(BG; \mathbb{Z})$ (where $\mathcal{T} = |G|$ in our discussion). The difference is because here we are performing the "$\mathcal{T} \hookrightarrow E\mathcal{T}$ step" from (119) to (128), while in [90] the delooping is being performed from (119) to (123).

  In our construction of Λ, we started with the infinite dimensional (127) (obtained by performing the "$\mathcal{T} \hookrightarrow E\mathcal{T}$ step" to the tautological bundle gerbe), in which the 1-morphisms in $\bar{\mathcal{P}}\mathcal{T}$ simply compose by concatenation. By contrast, if we try to deloop the tautological bundle gerbe, elements in $Y' = \bar{\mathcal{P}}_*\mathcal{T}$ cannot compose by concatenation, since the starting point is fixed while the ending point is arbitrary. Therefore, in [90], to facilitate the delooping, $Y'' = \mathcal{P}_*G$ (recall $\mathcal{T} = |G|$ in our discussion) with the starting point fixed at $\mathbf{1} \in G$ and without identification under thin homotopy is used, so that the composition of 1-morphisms can be defined by pointwise group mutiplication. To define the 2-morphisms of the delooping, in [90] the Mickelsson product is used for horizontal composition and geometrical concatenation is used for vertical composition before modding out WZW (while in [38] the Mickelsson product is used for both horizontal and vertical compositions), hence defining a 2-group $\mathcal{P}_*G \ltimes (\mathcal{P}_*\Omega_*G \times U(1)/WZW) \rightrightarrows \mathcal{P}_*G \rightrightarrows *$ in comparison to the 2-groupoid part $\bar{\mathcal{P}}_2\mathcal{T} \times U(1)/WZW \rightrightarrows \bar{\mathcal{P}}\mathcal{T} \rightrightarrows \mathcal{T}$ in the span of (127). This non-topological, detailed difference is finally carried over to the definitions of Λ via the transgression-regression process (129).

---

[135]It seems this extra S^2 may be some reminiscent (though not the same as) of the extra data in the more general bundle gerbe construction introduced in [91, 92]. But we are not sure about this yet and this should be further investigated.

- The second remark is on other choices of finite dimensional Y. This discussion will lead to the relation to Segal double cohomology [109], which is a smooth version of group cohomology for Lie groups.

  Our choice of Y is nice in that the patches are invariant under conjugation, which is important for manifesting the global symmetry, as explained in Section 4.1. However, if we do not care about the global symmetry but only the topology, then we can choose Y to be some finely cut patches, such that each of Y, $Y^{[2]} := Y \times_G Y$, $Y^{[3]} := Y \times_G Y \times_G Y$ and so on is a disjoint union of contractible spaces. In this case, the $U(1)$ bundle L over $Y \times_G Y$ is automatically trivial, and we can choose an arbitrary trivialization. While such choice of Y has the disadvantage of not manifestly respecting the global symmetry (hence not suitable for our construction of lattice QFT), it has an advantage that we can now consider the space $K_{n,m}$ as the pull back of G^n (viewed as the space of n-cells of the nerve of BG) along the covering of $Y^{[m]}$ over G, and then consider smooth mappings from $K_{n,m}$ to $U(1)$, which forms a *Segal double cochain complex* $C^{n,m}_{\text{Segal}}(G; U(1))$, from which we can define the *Segal double cohomology* $H^k_{\text{Segal}}(G; U(1))$ (for which a representative element involves one element from each of $C^{n,m}_{\text{Segal}}(G; U(1))$ that satisfies $n+m = k$), which is in turn isomorphic to $H^{k+1}(BG; \mathbb{Z}) \cong H^{k+1}(|BG|; \mathbb{Z})$. Now it is not hard to check that with such choice of Y, a multiplicative bundle gerbe is described by a representative element of a class in $H^3_{\text{Segal}}(G; U(1))$ (and we want to in particular realize the generator of the classification) [35, 110].

  The Segal double cohomology is a generalization of the usual group cohomology, with the advantage that we only consider those mappings from $K_{n,m}$ to $U(1)$ that are smooth, which is why Segal double cohomology was reinvented under the name *differentiable group cohomology* [110]. [136] When G is discrete, we can let $Y = G$ and Segal double cohomology reduces to usual group cohomology. On the other hand, in the usual group cohomology for Lie group G, in order for $H^n_{\text{group}}(G; U(1))$ to be isomorphic to $H^{n+1}(|BG|; \mathbb{Z})$, we must in general include mappings from G^n to $U(1)$ that are only piecewise continuous (Borel), which makes the lattice model thus built manifestly discontinuous [23].

  Therefore, in summary, in order to generalize the group cohomology based lattice models to the cases of Lie groups while making things smooth, we want to use Segal double cohomology in generalization of group cohomology, and further, in order for the construction to manifestly respect global symmetry, we want to go beyond Segal double cohomology and consider more generic multiplicative bundle gerbes that realize the anafunctor (123).

Now that in the second remark above we mentioned the relation to the usual group cohomology, let us further discuss the relation between our (128) for continuous-valued field nlσm and the familiar group cohomology treatment for discrete-valued field nlσm [23]. In fact, (128) encompasses the previous discrete cases. When G is discrete and $\mathcal{T} = |G|$, we can simply use $Y = G$ and $\Lambda = G^2 \times U(1)$ (of the form (103)), and $\Lambda^{(4)} \cong G^3 \times U(1)^4$; then we

---

[136]Indeed, very recently Segal double cohomology (differentiable group cohomology) has been used to carefully study anomalies of Lie group symmetries [111].

map $G^3 \times U(1)^4$ to $U(1)$ (assigning the WZW curvature) by multiplying the four $U(1)$ phases along with an extra $U(1)$ phase that depends on G^3—the associator from $H^n_{\text{group}}(G; U(1))$ seen in the previous section. In practice, since the extra $U(1)$ in Λ (and hence the extra $U(1)^4$ in $\Lambda^{(4)}$) does not contain topological information, it is usually discarded, [137] leaving only the associator. This explains why in (99) we said it is more natural to view the associator as a non-identity 3-morphism.

Related to this, the Dijkgraaf-Witten theory [19] with discrete gauge group—so that the flat connection condition can be (and is indeed) imposed—is encompassed by (123), with Y, Λ and the relevant discussions the same as the nlσm discussed in the previous paragraph. This is quite different from the case of continuous gauge group, where the flat connection condition becomes unphysical—so that we cannot start with BG but must start with BEG—and we need more involved treatment to be introduced in (130) below.

We can further note that, more general topological orders with discrete d.o.f. beyond group theory are also encompassed by lattice models with simplicial sets as target categories. We have in mind the Turaev-Viro theory [20] and generalizations (see e.g. [41]). Given the set of 1-cell link variables, the "admissible conditions" in the Turaev-Viro theory determine the set of 2-cells; they together form a discrete simplicial set that generalizes the right column BG (for discrete G in Dijkgraaf-Witten theory) in (123). Then, the functor to $B^3 U(1)$ specifies the F-symbols on the 3-cells. Usually, topological orders are phrased in terms of tensor categories (see e.g. [41, 106]) instead of simplicial sets. In Section 7 we will discuss this again in hope for a future unification of the usage of category theory in UV dynamical QFT and in IR topological QFT.

Finally we discuss the construction for lattice Yang-Mills theory. Clearly, what we need is a suitable definition for delooping (128) with $\mathcal{T} = |G|, G = SU(N)$, resulting in an anafunctor from the nerve of BEG (equipped with the inclusion $BG \hookrightarrow BEG$ for flat connections), the target category of traditional lattice gauge theory, to $B^4 U(1)$

$$
\begin{array}{ccccc}
U(1) & \longleftarrow & G^4 \times \widetilde{\Lambda}^{(5)} & \longrightarrow & G^{10} \\
\Downarrow & & \Downarrow\Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow\Downarrow \\
* & \longleftarrow & G^3 \times \widetilde{\Lambda} & \longrightarrow & G^6 \\
\Downarrow & & \Downarrow\Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow \\
* & \longleftarrow & G^2 \times Y & \longrightarrow & G^3 \\
\Downarrow & & \Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow \\
* & \longleftarrow & G & \longrightarrow & G \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & * & \longrightarrow & *
\end{array}
\tag{130}
$$

that realizes the generator of $H^5(BEG, BG; \mathbb{Z}) \xleftarrow{\sim} H^4(BG; \mathbb{Z}) \cong \mathbb{Z}$, the second Chern class. Here G^3 is the three 1-cells around a triangular 2-cell, and we can equivalently say $G^3 \cong G^2 \times G$ where the last G is the holonomy around the 2-cell, and Y covers this holonomy G;

---

[137] This discarded $U(1)$ can however be recognized as the complex phase of the $\mathbb{C}$-linear enrichment in the definition of a tensor category—a standard language used to describe of topological theories with discrete d.o.f. (e.g. [106]). See slightly more detailed discussions in Section 7.

likewise, G^6 is the six 1-cells around a tetrahedral 3-cell, and we can say $G^6 \cong G^3 \times G^3$ where the last $G^3 \cong G^3 \times_G G$ are the holonomies around four 2-cells around the tetrahedral 3-cell, and $\widetilde{\Lambda}$ is a suitable $U(1)$ bundle over the four triangular 2-cells $Y^3 \times_G Y$ covering the $G^3 \times_G G$; five such tetrahedral 3-cells piece up to a 4-cell with $\Lambda^{(5)}$, which maps to the $U(1)$ at the upper left that represents the $e^{id\mathcal{C}_h} = e^{i2\pi\mathcal{I}_h}$ on the lattice. The $G^3 \times \widetilde{\Lambda} \rightrightarrows^{\rightarrow} G^2 \times Y \rightrightarrows^{\rightarrow} G \rightrightarrows *$ part of the span is a weak 3-group, modeled as a Kan simplicial manifold with a single object, that extends BEG (equipped with the inclusion $BG \hookrightarrow BEG$) by $B^3 U(1)$.

To complete the construction, we perform the last Villainization step by composing on the left of this anafunctor the quarce (i.e. four times) delooping of (110), such that the $\mathbb{Z}$ of 5-morphisms represents the Yang monopole. The d.o.f. in the span thus becomes a weak 4-group, modeled as a Kan simplicial manifold with a single object, that extends BEG (equipped with the inclusion $BG \hookrightarrow BEG$) by $B^4 \mathbb{Z}$.

Section 4.2, and the follow-up paper [10] in greater details, describe how the intuitions from continuum QFT, partially involving ideas from the previous efforts [8, 96], help us construct such a structure (130) (except there we used cubical cells rather than simplicial cells); issues like the Yang-Baxter equation consistency constraint from the delooping are automatically resolved, as we will try to explore why in Section 6.2. As far as we are aware of, such a structure (130) seems not to have been formally introduced in the mathematical literature. We speculate that (130) should, in some suitable sense, be a finite dimensional realization of *CS bundle 2-gerbe.*

Let us briefly explain our speculation. A CS bundle 2-gerbe, introduced in [36] along with multiplicative bundle gerbe, is basically a multiplicative bundle gerbe over the G fibre of the universal bundle $|EG|$ over $|BG|$—and $|BG|$ is most often infinite dimensional. In terms of anafunctors, the universal bundle

$$
\begin{array}{ccccc}
G & \longleftarrow & |EG| \times_{|BG|} |EG| & \longrightarrow & |BG| \\
\Downarrow & & \Downarrow & & \Downarrow \\
* & \longleftarrow & |EG| & \longrightarrow & |BG|
\end{array}
\tag{131}
$$

(where $|EG| \times_{|BG|} |EG| \cong |EG| \times G$) is an anafunctor from $|BG|$ to BG. Note that the right column of a multiplicative bundle gerbe (123) is indeed BG (taken the nerve). So a CS bundle 2-gerbe can be viewed as the composition of the multiplicative bundle gerbe anafunctor (123) on the left of the universal bundle anafunctor (131). Now, our rough idea is to replace the infinite dimensional space $|BG|$ appearing in (131) by a finite dimensional structure. We have not figured out how this really works, but the rough idea is that the $BG \hookrightarrow BEG$ understood in (130) is a finite dimensional implementation of (131). Another possible direction is to take the singular simplicial complex $S|BG|$ of $|BG|$, which is the right adjoint to taking the geometrical realization. Then (131) should lift to an anafunctor from $S|BG|$ to (the nerve of) BEG, where the 1-cell G in BEG represents the value of a Wilson line along a path in $|BG|$—which is an element of $\Delta_1|BG| = S|BG|_1$—using the universal connection on the universal bundle $|EG|$, and the 2-cell G^3 represents the Wilson lines along the three edges of a triangular region in $|BG|$—which is an element of $\Delta_2|BG| = S|BG|_2$—in general with non-trivial holonomy. This should be roughly how the CS bundle 2-gerbe, that involves (131), might be related to the target category (130) we used in Section 4.2 to define CS on the lattice. We hope an actual understanding can be developed in future works.

6 Sketching a Relation between Continuum QFT and Lattice QFT

From the Villain model to our constructions, and from the explicit descriptions in physical terms to the systematic derivations in categorical terms, we have seen the geometrical intuition from the continuum played a crucial role. This is natural, because the very purpose of our work is to realize in lattice QFT those topological operators that are present in continuum QFT.

In Section 5.5 we derived our constructions in Section 4, and gave of mathematical notion of what "topological refinement" means. We started from the desired algebraic information $H^3(\mathcal{T}; \mathbb{Z})$ or $H^4(|BG|; \mathbb{Z})$, and the geometrical intuition from continuum is to facilitate the realization of the desired algebraic information. In this section, we want to reverse the emphasis. We want to begin with the geometrical picture from continuum QFT and come up with a corresponding lattice QFT, such that the algebraic information is to facilitate a suitable truncation of the geometrical details. This should lead to a systematic relation between continuum QFT and lattice QFT.

To motivate in another way, traditionally, we are familiar with the idea that a lattice QFT in the UV leads to a continuum QFT when renormalized towards the IR. However, there are also many situations—such as lattice QCD—in which we want to do the reverse, i.e. we want to find a lattice QFT that suitably describes some given continuum QFT. For TQFT, such a connection has been well-developed [19,22,23], simply because the UV and the IR really are not that different if the QFT is topological. Now we want to explore whether such a connection can be drawn for more general QFTs with dynamical d.o.f..

At this stage, such a broader picture, extending beyond our primary goal of arriving at the constructions in Section 4, is only a sketched one. We however do believe this is a good starting point for more systematic exploration of the relation between continuum QFT and lattice QFT.

6.1 Non-linear sigma models

When we say "a field configuration" in a continuum nlσm, we simply mean a smooth function from the spacetime manifold to the target manifold,

$$\mathcal{M} \to \mathcal{T}. \tag{132}$$

The path integral is intended to integrate over the space of all such functions. But this space is infinite dimensional and the path integral is not well-defined.

Let us ask what we intend to mean when we say "a field configuration" in a lattice nlσm. Of course, in traditional lattice nlσm, it is just a function from the lattice vertices, $\mathcal{L}_0$, to the target manifold $\mathcal{T}$. As we have seen in the previous sections, it is helpful to think of the lattice as being embedded in the continuum, then we can say, traditionally, a field configuration on the lattice is just a sampling of the continuum field at certain points on $\mathcal{M}$,

$$\mathcal{L}_0 \hookrightarrow \mathcal{M} \to \mathcal{T}. \tag{133}$$

Obviously a lot of information in the continuum field configuration is lost after the sampling.

To solve this problem, it turns out useful to think not only of $\mathcal{M}$ the manifold itself, but also its higher path spaces, which together form a higher (or infinite) groupoid. The realization is non-unique. We can use the singular simplicial complex $(\cdots \Delta_2\mathcal{M} \rightrightarrows\!\!\!\!\rightarrow \Delta_1\mathcal{M} \rightrightarrows \mathcal{M})$, or the cubical analogue $(\cdots \mathcal{P}^2\mathcal{M} \rightrightarrows\!\!\!\!\rightrightarrows \mathcal{P}\mathcal{M} \rightrightarrows \mathcal{M})$ where $\mathcal{P}$ means taking the path space without any thin homotopy identification, [138] or the weak higher category $(\cdots \mathcal{P}_2\mathcal{M} \rightrightarrows \mathcal{P}\mathcal{M} \rightrightarrows \mathcal{M})$ where $\mathcal{P}_n\mathcal{M} \subset \mathcal{P}^n\mathcal{M}$ is the space of interpolation of two elements of $\mathcal{P}_{n-1}\mathcal{M}$ that share boundaries in $\mathcal{P}_{n-2}\mathcal{M}$. [139] These realizations can capture the full homotopy information of $\mathcal{M}$; while for many physical applications, such as those considered in the present work which only concern the lowest non-trivial π_n, using the strict higher path groupoid $(\cdots \bar{\mathcal{P}}_2\mathcal{M} \rightrightarrows \bar{\mathcal{P}}\mathcal{M} \rightrightarrows \mathcal{M})$ would also be sufficient. [140] Similarly for the target manifold $\mathcal{T}$. We then have the simplicial map (assuming we used the simplicial realization, but we can also use the other realizations mentioned before; same below)

$$
\begin{array}{ccc}
\ldots & \rightarrow & \ldots \\
\Downarrow\!\Downarrow\!\Downarrow\!\Downarrow & & \Downarrow\!\Downarrow\!\Downarrow\!\Downarrow \\
\Delta_2\mathcal{M} & \rightarrow & \Delta_2\mathcal{T} \\
\Downarrow\!\Downarrow\!\Downarrow & & \Downarrow\!\Downarrow\!\Downarrow \\
\Delta_1\mathcal{M} & \rightarrow & \Delta_1\mathcal{T} \\
\Downarrow\!\Downarrow & & \Downarrow\!\Downarrow \\
\mathcal{M} & \rightarrow & \mathcal{T}
\end{array}
\tag{134}
$$

induced from $\mathcal{M} \to \mathcal{T}$, simply because the paths, surfaces and so on are all made of points. This simplicial map contains exactly the same amount of information as the original function $\mathcal{M} \to \mathcal{T}$.

The reason why we make things seemingly more complicated by including the higher path spaces is so that we can make better connection to the lattice. While $\mathcal{L}_0 \hookrightarrow \mathcal{M}$ is sampling some points in the continuum and lost the interpolation information, $(\mathcal{L}_1 \rightrightarrows \mathcal{L}_0) \hookrightarrow (\Delta_1\mathcal{M} \rightrightarrows \mathcal{M})$ is sampling some paths, hence retrieving more information about how the field interpolates from point to point. We can repeat this for higher dimensional cells (assuming the lattice is also a simplicial complex), until the d-dimensional cells completely

---

[138] $\mathcal{P}^2\mathcal{M}$ has four rather than two arrows to $\mathcal{P}\mathcal{M}$, because a path between two paths in general swipes out a square shape rather than a globular shape (which would be the case if we require the end points to be fixed—and that would be what we denoted as $\mathcal{P}_2\mathcal{M}$.

[139] This higher category is weak because without identification under thin homotopy, there is no identity in the strict sense, and composition is not strictly associative [39].

[140] See footnote 98.

fills up the continuum manifold. We obtain

$$
\begin{array}{ccccc}
\mathcal{L}_d & \hookrightarrow & \bar{\Delta}_d\mathcal{M} & \to & \bar{\Delta}_d\mathcal{T} \\
\downarrow\cdots\downarrow & & \downarrow\cdots\downarrow & & \downarrow\cdots\downarrow \\
\cdots & \hookrightarrow & \cdots & \to & \cdots \\
\Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow \\
\mathcal{L}_2 & \hookrightarrow & \Delta_2\mathcal{M} & \to & \Delta_2\mathcal{T} \\
\Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow \\
\mathcal{L}_1 & \hookrightarrow & \Delta_1\mathcal{M} & \to & \Delta_1\mathcal{T} \\
\Downarrow\Downarrow & & \Downarrow\Downarrow & & \Downarrow\Downarrow \\
\mathcal{L}_0 & \hookrightarrow & \mathcal{M} & \to & \mathcal{T}
\end{array}
\tag{135}
$$

where we have truncated the $\mathcal{M}$ column and the $\mathcal{T}$ column to the dth layer by taking identification of d-cells up to thin $(d+1)$-homotopy. After the truncation, the $\mathcal{L}$ column and the $\mathcal{M}$ column become ananaturally equivalent, [141] which roughly speaking means the d-dimensional lattice captures all the essential information of this truncated path d-groupoid of $\mathcal{M}$. We indeed do not expect the lattice theory to be able to capture those information that we truncated away—which, we believe, are physically unimportant anyways, as those truncated information are either unimportant UV details within each lattice cell (geometrically a tiny region), or higher homotopy information in $\mathcal{T}$ that seem not to be accessible by a d-dimensional QFT even in the continuum.

The above describes a functor from the lattice $\mathcal{L}$ to a target category, the $\mathcal{T}$ column, so it is almost interpretable as a lattice field configuration. Except there is one problem—the configuration is still essentially a continuum configuration, in the sense that, in general, the higher layers $\Delta_1\mathcal{T}, \Delta_2\mathcal{T}, \cdots, \bar{\Delta}_d\mathcal{T}$ in the target category are infinite dimensional spaces that came from the continuum picture (134), [142] which is undesired for a lattice theory.

What we gained is that now it becomes clear how the vague physical problem of defining a desired "topologically refined" lattice QFT should be turned into a well-posed mathematical problem:

$$
\begin{array}{cccccc}
\mathcal{L}_d & \hookrightarrow & \bar{\Delta}_d\mathcal{M} & \to & \bar{\Delta}_d\mathcal{T} & \mathbf{ET}_d \\
\downarrow\cdots\downarrow & & \downarrow\cdots\downarrow & & \downarrow\cdots\downarrow & \downarrow\cdots\downarrow \\
\cdots & \hookrightarrow & \cdots & \to & \cdots & \cdots \\
\Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow & \Downarrow\Downarrow\Downarrow \\
\mathcal{L}_2 & \hookrightarrow & \Delta_2\mathcal{M} & \to & \Delta_2\mathcal{T} & \overset{\text{equiv up to}}{\underset{\text{what we care}}{\longrightarrow}} \mathbf{ET}_2 \\
\Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow & \Downarrow\Downarrow\Downarrow \\
\mathcal{L}_1 & \hookrightarrow & \Delta_1\mathcal{M} & \to & \Delta_1\mathcal{T} & \mathbf{ET}_1 \\
\Downarrow\Downarrow & & \Downarrow\Downarrow & & \Downarrow\Downarrow & \Downarrow\Downarrow \\
\mathcal{L}_0 & \hookrightarrow & \mathcal{M} & \to & \mathcal{T} & \mathcal{T}
\end{array}
\tag{136}
$$

We want to reduce the third column, the simplicial path d-groupoid of $\mathcal{T}$, which in general involves infinite dimensional spaces, to an ananaturally equivalent (perhaps up to whatever

---

[141] Established by taking as the span the pullback of a Čech nerve over $\mathcal{M}$ (such that each patch is labeled by a lattice vertex) wtih the $\mathcal{M}$ column itself.

[142] Except for when $\mathcal{T} = S^1$, in which case we are basically done by now.

topological information we care about) but finite dimensional Kan simplicial manifold $\mathbf{ET}$*, with the objects* $\mathcal{T}$ *in the third column mapping identically to* $\mathbf{ET}_0 = \mathcal{T}$*. A topologically refined lattice nlσm field configuration is a functor (a simplicial map) from the lattice* $\mathcal{L}$ *to the target category* $\mathbf{ET}$*, which covers E$\mathcal{T}$, the target category of traditional lattice nlσm. Moreover, if* $\mathcal{T}$ *admits a global symmetry action* $G \times \mathcal{T} \to \mathcal{T}$*, then the action extends to an automorphism of simplicial manifold* $G \times \mathbf{ET} \to \mathbf{ET}$*. We make three crucial remarks:*

- By "up to what we care about", we mean, if $d > n$ but we only care about up to the π_n physics in a nlσm, then we can first further reduce the third column to a fundamental n-groupoid by taking identification in $\Delta_n \mathcal{T}$ under any $(n+1)$-homotopy, and then demand $\mathbf{ET}$ to only be ananaturally equivalent to this fundamental n-groupoid. In practice, for $n = 2$, we realize this by integrating the continuum Berry curvature over a 2d surface in $\Delta_2 \mathcal{T}$, as we did in (46). Similarly, for $n = 3$, in practice we first further reduce the third column to a fundamental 3-groupoid by integrating the WZW curvature over a 3d surface in $\Delta_3 \mathcal{T}$, as we essentially did in (51).

- We demand the continuum target space, i.e. the objects $\mathcal{T}$ of the third column, to map identically to $\mathbf{ET}_0 = \mathcal{T}$ because we still want to keep the ordinary vertex observables that take value in $\mathcal{T}$, acted on by the global symmetry in the ordinary way. If we do not demand this, that we will lose the dynamical information. For instance, suppose $d = 1$ and $\mathcal{T} = S^1$, we have $(\bar{\Delta}_1 \mathcal{T} \rightrightarrows \mathcal{T}) = (S^1 \times \mathbb{R} \rightrightarrows S^1)$ (which is already finite dimensional and can be readily used as $\mathbf{ET}$), but recall in (91) we said this is ananaturally equivalent to $B\mathbb{Z} = (\mathbb{Z} \rightrightarrows *)$. In the Villain model, we use $S^1 \times \mathbb{R} \rightrightarrows S^1$ as the target category, rather than $B\mathbb{Z}$, because we want to keep the dynamics of the S^1 d.o.f..

- The lattice nlσm field configurations constructed according to (136) forbid topological defects, simply because the construction started from smooth field configurations in the continuum, which do not contain defects. In many situations this is desired, if we want the lattice nlσm to represent a continuum nlσm which does not contain defect up to any accessible energy scale; by comparison, in a traditional lattice nlσm, the effects from defect fluctuation cannot be forbidden because the defects are not well-defined on the lattice. [143]

  In other situations, we might want to include the effects of defects on the lattice (meanwhile still being able to explicitly define the defects; otherwise we can just use the traditional lattice nlσm). To do so, we need a minimal enlargement $\mathbf{ET}'$ of $\mathbf{ET}$, such that $\mathbf{ET}'$ contains the $\mathbf{ET}$ in (136) as a subcategory, and $\mathbf{ET}'$ is ananaturally equivalent to a trivial category; moreover, $\mathbf{ET}'$ is the smallest category that satisfies these two

---

[143]A recent work [93] also considered forbidding defects in a lattice nlσm. The construction in [93] is by discretizing the target space $\mathcal{T}$ into a simplicial complex, so the target category is also a simplicial set. However, $\mathbf{ET}_0$ is thus not $\mathcal{T}$, but only some discrete points in $\mathcal{T}$, so the local dynamics of the continuous-valued d.o.f. is lost, and moreover the original continuous global symmetry on $\mathcal{T}$ cannot act on $\mathbf{ET}$ anymore. By comparison, the target category we constructed in (136) has the ordinary $\mathcal{T}$ d.o.f., with ordinary global symmetry action.

properties. The rationale behind these properties is the same as that explained below (91) and (92) through examples.

(Interestingly, the algebraic perspective in Section 5.5 constructs target categories that allow defects by default, and an extra step is needed if we want to forbid defects. While the geometrical perspective in this section constructs target categories that forbid defects by default, and an extra step is needed if we want to allow defects.)

This explains our basic idea of how higher category theory leads to a more systematic understanding of what it means to "discretize a continuum QFT". At this stage, the connection is only built at the level of field configurations in the path integral. In future works, it is important to also cast the path integral weight into this language.

6.2 Gauge theories

Now we attempt to suggest a reasonable systematic relation from continuum gauge theory to lattice gauge theory. Further work is needed to complete the understanding.

In the continuum, there are two ways to think about a gauge field configuration,

$$
\begin{array}{ccc}
\bar{\mathcal{P}}\mathcal{M} & & G \\
\Downarrow & \to & \Downarrow \\
\mathcal{M} & & *
\end{array}
\qquad\qquad \mathcal{M} \to |BG| \tag{137}
$$

where the first way, shown on the left, is the anafunctor description explained in Section 5.2, while the second way, shown on the right, makes use of the universal gauge connection on $|BG|$ [19]. The advantage of the first way is that the target category is finite dimensional and the anafunctor is readily the Wilson lines, and thus a field configuration in traditional lattice gauge theory is just a sampling

$$
\begin{array}{ccccc}
\mathcal{L}_1 & \hookrightarrow & \bar{\mathcal{P}}\mathcal{M} & & G \\
\Downarrow & & \Downarrow & \longrightarrow & \Downarrow \\
\mathcal{L}_0 & \hookrightarrow & \mathcal{M} & & *
\end{array}
\;. \tag{138}
$$

[144] The advantage of the second way is that a gauge theory can now be seen as a nlσm valued in $|BG|$, so that we can connect the problem to what we already know for nlσm, albeit there is a difference that $|BG|$ is in general infinite dimensional.

If we view a continuum gauge field in the second way, then we are almost done. Following

---

[144]Although the functor from the path groupoid $\bar{P}\mathcal{M}$ to BG is an anafunctor, the functor from the lattice to BG can be an ordinary functor, because the lattice is discrete.

the reasoning that led to (136), we have

$$
\begin{array}{ccccccc}
\mathcal{L}_d & \hookrightarrow & \bar{\Delta}_d\mathcal{M} & \to & \bar{\Delta}_d|BG| & & \mathbf{BEG}_d \\
\downarrow\cdots\downarrow & & \downarrow\cdots\downarrow & & \downarrow\cdots\downarrow & & \downarrow\cdots\downarrow \\
\cdots & \hookrightarrow & \cdots & \to & \cdots & & \cdots \\
\Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow \\
\mathcal{L}_2 & \hookrightarrow & \Delta_2\mathcal{M} & \to & \Delta_2|BG| & \xrightarrow[\text{what we care}]{\text{equiv up to}} & \mathbf{BEG}_2 \\
\Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow & & \Downarrow\Downarrow\Downarrow \\
\mathcal{L}_1 & \hookrightarrow & \Delta_1\mathcal{M} & \to & \Delta_1|BG| & & G \\
\Downarrow & & \Downarrow & & \Downarrow & & \Downarrow \\
\mathcal{L}_0 & \hookrightarrow & \mathcal{M} & \to & |BG| & & *
\end{array}
\tag{139}
$$

where the desired topologically refined target category on lattice is a finite dimensional Kan simplicial manifold $\mathbf{BEG}$ that is ananaturally equivalent (perhaps up to whatever topological information we care about) to the simplicial path d-groupoid of $|BG|$. But instead of the $\mathbf{ET}_0 = \mathcal{T}$ condition in (136), here we require $(\mathbf{BEG}_1 \rightrightarrows \mathbf{BEG}_0) = BG = (G \rightrightarrows *)$ to be obtained from $\Delta_1|BG|$ using the connection on the universal bundle (recall the discussions below (131)). This is because, unlike $\mathcal{T}$ in actual nlσm, $|BG|$ is already infinite dimensional in general, so we want to only keep the finite dimensional Wilson line information instead of $|BG|$ itself; also, we expect $\mathbf{BEG}$ to cover the target category of traditional lattice gauge theory, BEG, whose two lowest layers indeed form BG. Again, the target category constructed by (139) forbids topological defects; the way to include topological defects is the same as that discussed below (136).

It is currently unclear to us how to think about the problem of topological refinement if we, instead, begin with viewing a continuum gauge field in the first way in (137). It seems we can think of (134) but with $\mathcal{T} = G$ and the corresponding column delooped; the lowest two layers would indeed agree with the first way in (137). However, no matter which realization we use—the simplicial $\Delta_n G$, the cubical $\mathcal{P}^n G$, the weak $\mathcal{P}_n G$, or even the strict $\bar{\mathcal{P}}_n G$—the delooping is quite non-trivial, as we have seen through simpler examples (in which some Yang-Baxter equation constraint on the interchanger/braiding will come up) at the end of Section 4.2 and the end of Section 5.3.

The final target category we want for (139), $\mathbf{BEG}$, should indeed be a suitable delooping of $\mathbf{T}$ in (136) for $\mathcal{T} = G$. However, the actual construction in Section 4.2 did not directly use the delooping of the third column (the simplicial path d-groupoid of $\mathcal{T}$) in (136), despite some similarities in treatments. Instead, we used the continuum CS 3-form, and the Yang-Baxter equation issue due to delooping never really came up. Let us try to sketch a tentative answer to why. The continuum CS 3-form on a manifold $\mathcal{M}$ can be viewed as the pullback of the universal CS 3-form on $|BG|$. Therefore, it seems in Section 4.2, in constructing $\mathbf{BEG}$, "a suitable delooping of $\mathbf{T}$ for $\mathcal{T} = G$", we are already essentially using the perspective (139) to a certain extent, even though the infinite dimensional classifying space $|BG|$ did not explicitly come up. We should understand this better in future works.

7 Further Thoughts

This final section is for our further, scattered thoughts. We will begin with some near term problems. Then we will discuss some long term prospects.

Numerical implementation. Actual numerical implementation in the near future is definitely the primary aim of this paper. Our constructions in Section 4 for S^3 nlσm and $SU(N)$ gauge theory on lattice serve to introduce the key concepts that allow the topological operators to become well-defined. For actual numerical implementation, a more explicit proposal is presented in the subsequent work [10] (focusing on gauge theory). We emphasize that, given the principles stated in the present paper, the detailed implementation is not unique, and there may be better ways to practically construct the suitable (and, desirably, numerically optimized) path integral weights, especially the $W_2(e^{i\mathcal{W}}\mu^* + c.c.)$ in nlσm and $W_3(e^{i\mathcal{C}}\nu^* + c.c.)$ in gauge theory, either through some clever analytical method, or some automated optimization program such as some form of machine learning or so. While the actual implementation takes some extra efforts, the traditional fundamental obstacle to defining topological operators on the lattice should have been lifted by now with the key concepts we introduced.

Even aside of the purpose of explicitly defining the topological operators, it is still interesting to compare our construction to the traditional lattice QFT. In traditional lattice QFT, in order to better converge to the continuum limit, *Symanizk improvement* has been introduced [9, 53–57]. Roughly speaking, the Symanzik improvement introduced extra tuning parameters by going beyond nearest neighbor coupling; for gauge theory, this means to consider the gauge holonomy around more than one plaquette. By constrast, even without going beyond nearest neighbor coupling, our topological refinement introduces extra tuning parameters by weighing the higher morphisms in the target category, which roughly represent the interpolations of fields if we think of the lattice as being embedded in the continuum. It seems the extra weights introduced in the latter way are physically better interpretable. For the simplest example, consider the vortex fugacity weight introduced in the Villainized S^1 nlσm (12), which obviously controls the likelihood of vortices; this is important for setting up the renormalization analysis for the BKT transition [14, 52] (we will discuss more about renormalization later). Moreover, summing over the Villain integer variable m_l with non-trivial vortex fugacity weight will indeed generate beyond-nearest neighbor couplings between the traditional S^1 variables $e^{i\theta_v}$ (compared to (7) when the vortex fugacity weight is trivial), although the result cannot be expressed analytically. Similarly, integrating out the Berry connection field (along with its Dirac string field) with non-trivial Maxwell weight in the spinon-decomposed S^2 nlσm (38) will generate beyond-nearest neighbor coupling between the traditional S^2 variables. Based on this, we expect that, in general, the higher morphism weights from the topological refinement will (at least partly) play the role of Symanzik improvement, in a physically more interpretable manner; and since the topological operators are explicitly controlled, it is interesting to understand whether there is a relation to the numerical problem of topological freezing. These problems are in their own right worthwhile to be studied numerically.

Generalizations. There are some directions of generalization that worth working out.

1. Throughout this paper we have only been interested in those topological operators that are captured by the lowest non-trivial π_n, for $n \leq 3$. We should also consider cases with multiple types of topological operators of interest, captured by several non-trivial π_n's, since they might have non-trivial interplay. Physically relevant examples include S^2 nlσm with both π_2 and π_3 in consideration [112, 113] (rather than just π_2 in Section 2.4), $\mathbb{R}P^2$ nlσm with π_1, π_2 and π_3 in consideration (rather than just π_1 in Section 2.3); we will study these examples in subsequent works. For gauge theories, it is also important to consider non-abelian gauge groups such as $O(N)$ that have non-trivial π_0, π_1 before π_3 [19], and for these cases the general multiplicative bundle gerbes constructed in [91, 92] will be useful.

2. Throughout this paper our examples are either pure nlσm or pure gauge theory. We should also consider the topological operators when we couple lattice nlσm to lattice gauge field (background or dynamical), especially for those constructed in Section 4. As mentioned there, this is in particular important for manifesting the anomalies on lattice. (In Section 2, the anomalies in the known lattices models have been manifested, with details presented in the footnotes 12 and 35.)

3. Constructions for π_n topological operators for $n > 3$ seem to require some further efforts on the mathematical side. π_5 is particularly physically relevant for the 4d WZW term in the low energy nlσm of QCD [94, 95] (and also π_4 if the nlσm is the pion S^3). And there are other examples in strongly coupled theories in both high energy physics and condensed matter physics.

More general observables and representations of weak higher groups. Consider our topologically refined $SU(N)$ lattice gauge theory for example. At the end of Section 5.5, we explained that the target category is a weak Lie 4-group, realized as a Kan simplicial manifold with single object. Mathematically, there should exist a suitable notion of "representations of the weak Lie 4-group", which should be worked out explicitly.

Physically, this corresponds to answering the following question. Suppose the Yang-Mills theory lives on a spacetime of dimension $d \geq 4$. We know there is a class of observables living on 1d submanifolds, the Wilson lines, characterized by representations of G, where G is the 1-morphisms of the Lie 4-group. There is a class of observables living on (oriented) 3d submanifolds, the CS terms, characterized by the integer CS levels, which are representations of $U(1)$, where $U(1)$ is the new d.o.f. in the 3-morphisms of the Lie 4-group. There is a class of observables living on (oriented) 4d submanifolds, the topological theta terms, characterized by the theta angles, which are representations of $\mathbb{Z}$, where $\mathbb{Z}$ is the new d.o.f. in the 4-morphisms of the 4-group. But can we also characterized some observables living on 2d submanifolds? The new d.o.f. in the 2-morphisms of the weak 4-group do not form a group in the ordinary sense, so they do not have representation in the ordinary sense, but since the whole structure forms a weak 4-group, it is reasonable to anticipate that we can organize observables living on submanifolds from 1d to 4d into some notion of representation of the weak 4-group.

Similarly, we should also ask, for a nlσm that lives in $d \geq 3$, on 0d submainifolds there are the order parameters, on 2d submanifolds there are WZW levels, on 3d submanifolds there

121

are topological theta terms, then how shall we characterize some observables living on 1d submanifolds, so that all these observables together form a coherent categorical structure?

Hamiltonian formalism. It is natural to ask if the topologically refined lattice constructions we introduced on the Euclidean spacetime lattice have corresponding versions on the spatial lattice in the Hamiltonian formalism. While we expect there to be, it takes further efforts to work out the details. In particular, for ordinary group valued operators, their canonical operators are characterized by the representations, so now the weak higher group representation problem described above might become particularly relevant.

There is an extra issue to be noted as the d.o.f. of interest are continuous-valued—even when the d.o.f. are ordinary groups, such as in Villainization. We emphasized that the d.o.f. in the target category in general do not factorize; on the other hand, a lot of times in the Hamiltonian formalism it is desired that the physical Hilbert space factorizes locally on the spatial lattice. If we indeed demand so, there is a familiar treatment when the d.o.f. are discrete-valued [21, 22]: We can let the physical Hilbert space be an enlarged, locally factorized one, and then have energy penalty terms in the Hamiltonian, such that a low energy subspace is exactly the desired, non-factorized Hilbert space, and moreover all higher energy states have a finite gap above this low energy subspace. However, when the d.o.f. are continuous-valued, under the same treatment there is no such gap because the energies of the states vary continuously (unless we use a Hamiltonian with discontinuous matrix elements, but such unphysical treatment will lead to other problems). Suitably modified treatment has been developed in the case of Villainized $U(1)$ gauge theory [74, 75], in order to ensure the emergence of a low energy subspace with the desired non-factorized properties, meanwhile having a finite gap separated from the higher energy states. We expect similar issue occurs for more general target categories, if we want a locally factorized physical Hilbert space.

The above are more or less well-defined problems that we believed can be solved in the near future. In the below, we sketch some directions that we believe worth explorations for the long term. The discussions below are highly speculative at this point.

Renormalization. As we have seen, a field configuration in a lattice QFT is a functor from the lattice to a target category, where the latter is constructed based on the target space of the desired continuum QFT, either from the more algebraic perspective described in Section 5.5, or the more geometric perspective described in Section 6. The path integral is to integrate over the space of all such functors; at least in the examples that we have seen, the measure to use for the integral is obvious, due to the global symmetry or gauge group. On the other hand, the integrand, i.e. the path integral weight, still awaits to be casted in this categorical language. Of course, the weight is a suitably constructed map from the space of field configuration functors to the non-zero complex numbers $\mathbb{C}_*$. But what is meant by "suitable" needs to be clarified.

Locality is a crucial requirement. In the constructions we presented, the weight is a product of factors contributed by individual vertices, links, plaquettes, and so on, therefore a map from the space of n-morphisms of the target category, for each n, to $\mathbb{C}_*$ is involved. But more general weights are also legitimate—those short ranged but beyond nearest neigh-

bor couplings (we have mentioned this when discussing numerical implementation at the beginning of this section). So we need to find a concise way to convey the requirement of locality in the weight assignment.

Another layer of the problem is that there are two kinds of weight contributions: the "non-topological" ones which contributes a positive magnitude, and the "topological" ones which contributes a $U(1)$ phase, such as the topological theta terms, Berry phase, WZW phase, and CS phase. As required by reflection positivity [114,115]—the Euclidean version of unitarity—under orientation reversal of the spacetime, [145] the positive magnitude contributions must be invariant, while the phase contributions must become complex conjugation. But there are further distinctions between the two kinds of contributions: The "non-topological" weights seem to be locally well-defined "outright", whilst the "topological" weights (such as the Berry phase, WZW phase, CS phase) may not be well-defined on individual lattice cells or, more generally, regions with boundaries, in the sense that there will be dependence on some notion of "gauge" on the boundary conditions, and related to this the $U(1)$ phase contribution from such a region in general takes value from a non-trivial $U(1)$ bundle over the space of boundary conditions. (For $U(1)$ CS-Maxwell, see [66] for details.) We need a concise way to capture these aspects into a complete categorical definition of lattice QFT with generic dynamics.

Suppose the above can be achieved in the foreseeable future. Then we can try to formulate renormalization in the categorical language. It can envisioned that there should be a category of lattice QFTs, whose objects contain information about the (topologically refined) target category and the path integral weight assignment. The coarse graining of lattice can certainly be realized in terms of inclusion functors between lattices (with the IR limit being some notion of skeletal lattice). And we want the coarse graining inclusion functor, as morphisms in the category of lattices (discrete Kan complexes), to induce certain "renormalization morphisms" in the category of lattice QFTs.

Perhaps a better way to realize the coarse graining of lattice is by general anafunctor, rather than ordinary inclusion functor, despite that according to Section 5.2 there seem to be no necessity to use anafunctor when dealing with discrete spaces. The reason is, over the past two decades, it has become increasingly clear that a good way to think about renormalization is to think about an AdS_{d+1} spacetime, with the extra "radial direction" representing the renormalization scale [116]. Then, by (88), naturally the lattice links, plaquettes and so on connecting two consecutive radial slices constitute the span of the anafunctor for one step of coarse graining. (It is furthermore illuminating to think of the lattice AdS_{d+1} as a double category—recall footnote 94.) Then, the problem becomes how this perspective of coarse graining a lattice is lifted to the level of renormalizing a lattice QFT.

Relation to categories involved in topological quantum field theory. In the long term, if we have a good categorical understanding of what renormalization is, we can then discuss what a renormalization fixed point means. Hopefully, we can see how the familiar categorical description of the IR fixed point emerges from a description of generic QFT after

---

[145]It is understood that, if there are extra background structures on the spacetime, such as background gauge field, branching structure, etc., involved in defining the theory, these background structures are also transformed under the orientation reversal.

renormalization.

Even at the present stage, it may be a good idea to begin pondering the difference between the categories familiar in the topological QFT (TQFT) context versus the categories involved in the present work for QFT with generic dynamics. They seem to belong to different branches of the development of category theory. The categories that familiarly describe the IR physics are equipped with a long list of extra structures and requirements (see e.g. [106]), in order to reproduce all the desired nice physical properties that an IR fixed point should have; moreover, the systematically well-studied ones involve discrete-valued d.o.f. only, while continuous-valued d.o.f. still poses a crucial challenge. In comparison, the categories we used in the present work are much "simpler" in definition, with less structures and requirements; moreover, they can, and are primarily designed to, describe continuous-valued d.o.f. with homotopy properties that are of interest.

While here we cannot launch a full-scale analysis of the problem, we want to bring up the aspect that we think is crucial. The categories used in TQFT are equipped with—more precisely, enriched by—complex linear spaces, which will play the role of the quantum Hilbert spaces. In comparison, the categories we used in this paper have no built-in linear structure, but we know our constructions do have the quantum mechanical linearity, simply because, in the end, we are constructing well-defined path integrals. It seems the former case requires the linear space structure to be built-in because for TQFT there is no distinction between the UV and the IR, so the same category is to be used to describe both the UV d.o.f. (say, in constructing a lattice model) and the IR states. While in the latter case, the UV d.o.f. do not have to *a priori* incorporate anything about the IR state.

Based on this idea, it seems we can bridge these two kinds of categories by dropping the built-in linear structure in the former. But the linear structure enrichment plays some crucial role, so one may wonder how that role is otherwise fulfilled now that we dropped the linear structure. In particular, let us consider the notion of fusion in a fusion category. The linear structure allows one to define the notions of simple objects (interpreted as simple anyons) and direct sums, such that every object is some finite direct sum of simple objects; when we fuse two simple objects, the result is in general a non-simple object, a key feature of non-abelian topological order. How do we reproduce this if we give up the linear structure? The answer is to simply phrase the above is a more intuitive language—what we will usually say is, when two (simple) anyons fuse, there can be multiple possible fusion channels, giving rise to multiple possible results of (simple) anyon; there is no mention of non-simple anyon. But this is literally what a simplicial set does. That is, the role of the linear space in the fusion process can be played by the non-unique composition (of 1-morphisms) in simplicial weak categories. This is indeed how, as we mentioned in Section 5.5, the Turaev-Viro theory [20] and its generalizations (see e.g. [41]), whose construction is traditionally phrased in terms of unitary fusion categories, can be alternatively (and naturally) viewed, in our language, as simply having a simplicial set as target category, without having to mention any *a priori* built-in linear structure. [146]

---

[146]The basic d.o.f. in the Turaev-Viro model are link variables, and they are simple objects from a unitary fusion category, which describes the simple anyons on a gapped (1+1)d boundary of the (2+1)d system [117] (and the bulk anyons are described by the Drinfeld center of this unitary fusion category, see footnote 104). In our construction, the 1-morphisms that are being composed non-uniquely are indeed link variables.

Interestingly, if we really want to, we can still catch some reminiscence of the linear enrichment. Recall we said in Section 5.5 that when applying (123) to discrete BG for Dijkgraaf-Witten theory, or replacing BG with more general discrete simplicial set for Tureav-Viro theory, the Λ in the span has an independent $U(1)$ d.o.f.—which is often ignored in the lattice models. We can recognize this $U(1)$ in the fusion category language as the phase of the $\mathbb{C}$ in the linear enrichment.

When the d.o.f. are continuous rather than discrete, the familiar categorical paradigm for TQFT runs into difficulties at the definition level, and the difficulties indeed have to do with the built-in linear structure. On the other hand, as we have seen throughout this paper, the use of simplicial weak category without reference to the built-in linear structure enrichment can work for both discrete-valued and continuous-valued d.o.f.. (When the d.o.f. become continuous, at least one important change is that the extra $U(1)$ d.o.f. in Λ in general forms a non-trivial $U(1)$ bundle over the space of the continuous d.o.f..) Therefore, we anticipate that, exploring along this line of ideas might be a fruitful route towards a future unification of the use of categories in QFT, regardless of whether the QFT is IR TQFT or generic dynamical QFT, and whether the d.o.f. are discrete or continuous-valued.

Constructive quantum field theory. An ultimate question about a lattice QFT is whether some suitable notion of continuum limit exists. Numerically there are good evidences for the convergence in lattice QCD, but one may wonder whether this can be shown analytically. In fact, this is one possible route towards the program of *constructive QFT*, i.e. towards constructively defining what a continuum QFT is. This route has some crucial advantages compared to other possible routes, aside from being more intuitive: Most importantly, reflection positivity (see our discussion about renormalization above) is built-in as long as the lattice QFT itself is a legitimate path integral; moreover, if dynamical gauge field is involved, the gauge redundancy (for compact Lie group) requires literally no treatment at the fundamental level [1] (though if one wants to one can still fix the gauge).

Remarkable partial results have been achieved by Balaban in this regard. Through highly technical analyses, Balaban showed that, in 3d [118] and 4d [119,120], given a finite size four-torus Euclidean spacetime, as the lattice spacing decreases towards zero, Wilson's lattice Yang-Mills theory [1] is renormalized such that the value of the partition function remains stable within a finite range. This program is, unfortunately, almost not being carried on since then, perhaps due to its highly involved technicality.

It is natural to ask how our topological refinement of Wilson's traditional lattice gauge theory affects the analyses in this program. This is a technically very difficulty yet important question in the long term.

Since the topological refinement introduces new higher morphisms d.o.f. on the lattice and new weight factors for them, the renormalization flow is affected. The optimistic hope is, now that the topologically refined lattice QFT has a more systematic relation to the desired continuum QFT (Section 6), and moreover the non-perturbative topological operators such as instantons have become well-defined and explicitly controllable in the path integral, the renormalization towards the continuum may also be under better control.

Another optimistic hope is, now that we have a categorical understanding (at a preliminary level for now) of what a lattice QFT is in relation to a desired continuum QFT, and

in the future such an understanding may hopefully be extended to cover renormalization, eventually we may hope for a reorganization of the now highly involved analyses towards the continuum limit. Even though the essential technicality most likely will not be eliminated, a more systematic reorganization, if possible, may help with the progress on the analyses and the physical understanding of it.

References

[1] K.G. Wilson, *Confinement of Quarks*, *Phys. Rev. D* **10** (1974) 2445.

[2] M. Creutz, *Confinement and the Critical Dimensionality of Space-Time*, *Phys. Rev. Lett.* **43** (1979) 553.

[3] K.G. Wilson, *Monte Carlo Calculations for the Lattice Gauge Theory*, *NATO Sci. Ser. B* **59** (1980) 363.

[4] A.A. Belavin, A.M. Polyakov, A.S. Schwartz and Y.S. Tyupkin, *Pseudoparticle Solutions of the Yang-Mills Equations*, *Phys. Lett. B* **59** (1975) 85.

[5] G. 't Hooft, *Computation of the Quantum Effects Due to a Four-Dimensional Pseudoparticle*, *Phys. Rev. D* **14** (1976) 3432.

[6] G. 't Hooft, *How Instantons Solve the U(1) Problem*, *Phys. Rept.* **142** (1986) 357.

[7] T. Schäfer and E.V. Shuryak, *Instantons in QCD*, *Rev. Mod. Phys.* **70** (1998) 323 [`hep-ph/9610451`].

[8] M. Lüscher, *Topology of Lattice Gauge Fields*, *Commun. Math. Phys.* **85** (1982) 39.

[9] C. Alexandrou, A. Athenodorou, K. Cichy, A. Dromard, E. Garcia-Ramos, K. Jansen et al., *Comparison of topological charge definitions in Lattice QCD*, *Eur. Phys. J. C* **80** (2020) 424 [`1708.00696`].

[10] P. Zhang and J.-Y. Chen, *An Explicit Categorical Construction of Instanton Density in Lattice Yang-Mills Theory*, `2411.07195`.

[11] V.L. Berezinsky, *Destruction of long range order in one-dimensional and two-dimensional systems having a continuous symmetry group. I. Classical systems*, *Sov. Phys. JETP* **32** (1971), 493.

[12] V.L. Berezinsky, *Destruction of long range order in one-dimensional and two-dimensional systems having a continuous symmetry group. II. Quantum systems*, *Sov. Phys. JETP* **34** (1972), 610.

[13] J.M. Kosterlitz and D.J. Thouless, *Ordering, metastability and phase transitions in two-dimensional systems*, *J. Phys. C* **6** (1973) 1181.

[14] J.V. José, L.P. Kadanoff, S. Kirkpatrick and D.R. Nelson, *Renormalization, vortices, and symmetry-breaking perturbations in the two-dimensional planar model*, *Phys. Rev. B* **16** (1977) 1217.

[15] D.J. Gross and I.R. Klebanov, *One-Dimensional String Theory on a Circle*, *Nucl. Phys. B* **344** (1990) 475.

[16] M.E. Peskin, *Mandelstam 't Hooft Duality in Abelian Lattice Models*, *Annals Phys.* **113** (1978) 122.

[17] M.B. Einhorn and R. Savit, *Topological Excitations in the Abelian Higgs Model*, *Phys. Rev. D* **17** (1978) 2583.

[18] M.B. Einhorn and R. Savit, *Phase Transitions and Confinement in the Abelian Higgs Model*, *Phys. Rev. D* **19** (1979) 1198.

[19] R. Dijkgraaf and E. Witten, *Topological Gauge Theories and Group Cohomology*, *Commun. Math. Phys.* **129** (1990) 393.

[20] V.G. Turaev and O.Y. Viro, *State sum invariants of 3 manifolds and quantum 6j symbols*, *Topology* **31** (1992) 865.

[21] A.Y. Kitaev, *Fault tolerant quantum computation by anyons*, *Annals Phys.* **303** (2003) 2 [`quant-ph/9707021`].

[22] M.A. Levin and X.-G. Wen, *String-net condensation: A physical mechanism for topological phases*, *Phys. Rev. B* **71** (2005) 045110 [`cond-mat/0404617`].

[23] X. Chen, Z.-C. Gu, Z.-X. Liu and X.-G. Wen, *Symmetry protected topological orders and the group cohomology of their symmetry group*, *Phys. Rev.* **B87** (2013) 155114 [`1106.4772`].

[24] S. Elitzur, *Impossibility of Spontaneously Breaking Local Symmetries*, *Phys. Rev. D* **12** (1975) 3978.

[25] Y. Iwasaki and T. Yoshie, *Instantons and Topological Charge in Lattice Gauge Theory*, *Phys. Lett. B* **131** (1983) 159.

[26] M. Lüscher, *Properties and uses of the Wilson flow in lattice QCD*, *JHEP* **08** (2010) 071 [`1006.4518`].

[27] S. Itoh, Y. Iwasaki and T. Yoshie, *The U(1) Problem and Topological Excitations on a Lattice*, *Phys. Rev. D* **36** (1987) 527.

[28] P. Hasenfratz, V. Laliena and F. Niedermayer, *The Index theorem in QCD with a finite cutoff*, *Phys. Lett. B* **427** (1998) 125 [`hep-lat/9801021`].

[29] M. Lüscher and F. Palombi, *Universality of the topological susceptibility in the SU(3) gauge theory*, *JHEP* **09** (2010) 110 [`1008.0732`].

[30] J. Villain, *Theory of one-dimensional and two-dimensional magnets with an easy magnetization plane. II. The planar, classical, two-dimensional magnet*, *Journal de Physique* **36** (1975) 581.

[31] G. Mack and V.B. Petkova, *Z2 Monopoles in the Standard SU(2) Lattice Gauge Theory Model*, *Z. Phys. C* **12** (1982) 177.

[32] S. Gukov and A. Kapustin, *Topological Quantum Field Theory, Nonlocal Operators, and Gapped Phases of Gauge Theories*, `1307.4793`.

[33] A. Kapustin and R. Thorngren, *Topological Field Theory on a Lattice, Discrete Theta-Angles and Confinement*, *Adv. Theor. Math. Phys.* **18** (2014) 1233 [`1308.2926`].

[34] A. Kapustin and R. Thorngren, *Higher symmetry and gapped phases of gauge theories*, `1309.4721`.

[35] C.J. Schommer-Pries, *Central extensions of smooth 2–groups and a finite-dimensional string 2–group*, *Geometry & Topology* **15** (2011) 609 [`0911.2483`].

[36] A.L. Carey, S. Johnson, M.K. Murray, D. Stevenson and B.-L. Wang, *Bundle gerbes for Chern-Simons and Wess-Zumino-Witten theories*, *Commun. Math. Phys.* **259** (2005) 577 [`math/0410013`].

[37] M.K. Murray, *Bundle gerbes*, *J. Lond. Math. Soc.* **54** (1996) 403 [`dg-ga/9407015`].

[38] J.C. Baez, D. Stevenson, A.S. Crans and U. Schreiber, *From loop groups to 2-groups*, `math/0504123`.

[39] A. Grothendieck, *Pursuing Stacks*, unpublished (1983) [`2111.01000`].

[40] J.C. Baez, *An introduction to n-categories*, in *International Conference on Category Theory and Computer Science*, pp. 1–33, Springer, 1997 [`q-alg/9705009`].

[41] M. Barkeshli, P. Bonderson, M. Cheng and Z. Wang, *Symmetry Fractionalization, Defects, and Gauging of Topological Phases*, *Phys. Rev. B* **100** (2019) 115147 [`1410.4540`].

[42] C. Córdova, T.T. Dumitrescu and K. Intriligator, *Exploring 2-Group Global Symmetries*, *JHEP* **02** (2019) 184 [`1802.04790`].

[43] F. Benini, C. Córdova and P.-S. Hsin, *On 2-Group Global Symmetries and their Anomalies*, *JHEP* **03** (2019) 118 [`1803.09336`].

[44] J.C. Baez, *Higher Yang-Mills theory*, `hep-th/0206130`.

[45] H. Pfeiffer, *Higher gauge theory and a nonAbelian generalization of 2-form electrodynamics*, *Annals Phys.* **308** (2003) 447 [`hep-th/0304074`].

[46] J.C. Baez and J. Huerta, *An Invitation to Higher Gauge Theory*, *Gen. Rel. Grav.* **43** (2011) 2335 [`1003.4485`].

[47] K. Shiozaki, *A discrete formulation for three-dimensional winding number*, `2403.05291`.

[48] S. Ohyama and S. Ryu, *Higher structures in matrix product states*, *Phys. Rev. B* **109** (2024) 115152 [`2304.05356`].

[49] M. Qi, D.T. Stephen, X. Wen, D. Spiegel, M.J. Pflaum, A. Beaudry et al., *Charting the space of ground states with tensor networks*, `2305.07700`.

[50] A.Y. Kitaev, *On the Classification of Short-Range Entangled States*, unpublished, video of talk at Simons Center for Geometry and Physics, Program: Topological Phases of Matter (2013).

[51] A. Kapustin and L. Spodyneiko, *Higher-dimensional generalizations of Berry curvature*, *Phys. Rev. B* **101** (2020) 235130 [`2001.03454`].

[52] J.B. Kogut, *An Introduction to Lattice Gauge Theory and Spin Systems*, *Rev. Mod. Phys.* **51** (1979) 659.

[53] K. Symanzik, *Continuum Limit and Improved Action in Lattice Theories. 1. Principles and φ^4 Theory*, *Nucl. Phys. B* **226** (1983) 187.

[54] K. Symanzik, *Continuum Limit and Improved Action in Lattice Theories. 2. O(N) Nonlinear Sigma Model in Perturbation Theory*, *Nucl. Phys. B* **226** (1983) 205.

[55] P. Weisz, *Continuum Limit Improved Lattice Action for Pure Yang-Mills Theory. 1.*, *Nucl. Phys. B* **212** (1983) 1.

[56] P. Weisz and R. Wohlert, *Continuum Limit Improved Lattice Action for Pure Yang-Mills Theory. 2.*, *Nucl. Phys. B* **236** (1984) 397.

[57] G. Curci, P. Menotti and G. Paffuti, *Symanzik's Improved Lagrangian for Lattice Gauge Theory*, *Phys. Lett. B* **130** (1983) 205.

[58] C. Córdova, D.S. Freed, H.T. Lam and N. Seiberg, *Anomalies in the Space of Coupling Constants and Their Dynamical Applications I*, *SciPost Phys.* **8** (2020) 001 [`1905.09315`].

[59] T. Sulejmanpasic and C. Gattringer, *Abelian gauge theories on the lattice: θ-Terms and compact gauge theory with(out) monopoles*, *Nucl. Phys. B* **943** (2019) 114616 [`1901.02637`].

[60] J.-Y. Chen, *Abelian Topological Order on Lattice Enriched with Electromagnetic Background*, *Commun. Math. Phys.* **381** (2021) 293 [`1902.06756`].

[61] P. Gorantla, H.T. Lam, N. Seiberg and S.-H. Shao, *A modified Villain formulation of fractons and other exotic theories*, *J. Math. Phys.* **62** (2021) 102301 [`2103.01257`].

[62] A.M. Polyakov, *Compact Gauge Fields and the Infrared Catastrophe*, *Phys. Lett. B* **59** (1975) 82.

[63] M.B. Hastings and X.-G. Wen, *Quasiadiabatic continuation of quantum states: The stability of topological ground-state degeneracy and emergent gauge invariance*, *Phys. Rev. B* **72** (2005) 045141 [`cond-mat/0503554`].

[64] T. Jacobson and T. Sulejmanpasic, *Modified Villain formulation of Abelian Chern-Simons theory*, *Phys. Rev. D* **107** (2023) 125017 [`2303.06160`].

[65] D. Gaiotto and A. Kapustin, *Spin TQFTs and fermionic phases of matter*, *Int. J. Mod. Phys.* **A31** (2016) 1645044 [`1505.05856`].

[66] Z.-A. Xu and J.-Y. Chen, *Lattice Chern-Simons-Maxwell Theory and its Chirality*, `2410.11034`.

[67] C. Peng, M.C. Diamantini, L. Funcke, S.M.A. Hassan, K. Jansen, S. Kühn et al., *Hamiltonian Lattice Formulation of Compact Maxwell-Chern-Simons Theory*, `2407.20225`.

[68] F. Berruto, M.C. Diamantini and P. Sodano, *On pure lattice Chern-Simons gauge theories*, *Phys. Lett. B* **487** (2000) 366 [`hep-th/0004203`].

[69] T. Jacobson and T. Sulejmanpasic, *Canonical quantization of lattice Chern-Simons theory*, `2401.09597`.

[70] D. Bar-Natan and E. Witten, *Perturbative expansion of Chern-Simons theory with noncompact gauge group*, *Commun. Math. Phys.* **141** (1991) 423.

[71] E. Witten, *Quantum Field Theory and the Jones Polynomial*, *Commun. Math. Phys.* **121** (1989) 351.

[72] A.M. Polyakov, *Fermi-Bose Transmutations Induced by Gauge Fields*, *Mod. Phys. Lett.* **A3** (1988) 325.

[73] T.H. Hansson, A. Karlhede and M. Rocek, *On Wilson Loops in Abelian Chern-Simons Theories*, *Phys. Lett.* **B225** (1989) 92.

[74] Z. Han and J.-Y. Chen, *Solvable lattice Hamiltonians with fractional Hall conductivity*, *Phys. Rev. B* **105** (2022) 155130 [2107.02817].

[75] Z. Han and J.-Y. Chen, *Fractional Hall conductivity and spin-c structure in solvable lattice Hamiltonians*, *JHEP* **02** (2023) 130 [2208.13785].

[76] M. Nguyen, T. Sulejmanpasic and M. Ünsal, *Phases of theories with $\mathbb{Z}_N$ 1-form symmetry and the roles of center vortices and magnetic monopoles*, 2401.04800.

[77] T.D. Ellison, Y.-A. Chen, A. Dua, W. Shirley, N. Tantivasadakarn and D.J. Williamson, *Pauli Stabilizer Models of Twisted Quantum Doubles*, *PRX Quantum* **3** (2022) 010353 [2112.11394].

[78] E. Rabinovici and S. Samuel, *The CP^{N-1} Model: A Strong Coupling Lattice Approach*, *Phys. Lett. B* **101** (1981) 323.

[79] P. Di Vecchia, A. Holtkamp, R. Musto, F. Nicodemi and R. Pettorino, *Lattice CP^{N-1} Models and Their Large N Behavior*, *Nucl. Phys. B* **190** (1981) 719.

[80] S. Sachdev and R. Jalabert, *Effective lattice models for two-dimensional quantum antiferromagnets*, *Mod. Phys. Lett.* **B4** (1990) 1043.

[81] S. Sachdev and K. Park, *Ground states of quantum antiferromagnets in two dimensions*, *Annals Phys.* **298** (2002) 58 [cond-mat/0108214].

[82] G. 't Hooft, *Magnetic Monopoles in Unified Gauge Theories*, *Nucl. Phys. B* **79** (1974) 276.

[83] A.M. Polyakov, *Particle Spectrum in Quantum Field Theory*, *JETP Lett.* **20** (1974) 194.

[84] F.D.M. Haldane, *Continuum dynamics of the 1-D Heisenberg antiferromagnetic identification with the O(3) nonlinear sigma model*, *Phys. Lett. A* **93** (1983) 464.

[85] J. Wang, X.-G. Wen and E. Witten, *Symmetric Gapped Interfaces of SPT and SET States: Systematic Constructions*, *Phys. Rev. X* **8** (2018) 031048 [1705.06728].

[86] S.D. Pace, *Emergent generalized symmetries in ordered phases*, 2308.05730.

[87] T.H.R. Skyrme, *A Unified Field Theory of Mesons and Baryons*, *Nucl. Phys.* **31** (1962) 556.

[88] J.-Y. Chen, *Static Magnetic Response of Non-Fermi Liquid Density*, *Phys. Rev. Lett.* **119** (2017) 096601 [1704.04241].

[89] K. Gawędzki and N. Reis, *WZW branes and gerbes*, *Rev. Math. Phys.* **14** (2002) 1281 [`hep-th/0205233`].

[90] K. Waldorf, *A construction of string 2-group models using a transgression-regression technique*, *Analysis, geometry and quantum field theory* **584** (2012) 99 [`1201.5052`].

[91] E. Meinrenken, *The basic gerbe over a compact simple Lie group*, `math/0209194`.

[92] K. Gawedzki and N. Reis, *Basic gerbe over nonsimply connected compact groups*, *J. Geom. Phys.* **50** (2004) 28 [`math/0307010`].

[93] S.D. Pace, C. Zhu, A. Beaudry and X.-G. Wen, *Generalized symmetries in singularity-free nonlinear σ-models and their disordered phases*, `2310.08554`.

[94] E. Witten, *Global Aspects of Current Algebra*, *Nucl. Phys. B* **223** (1983) 422.

[95] Y. Lee, K. Ohmori and Y. Tachikawa, *Revisiting Wess-Zumino-Witten terms*, *SciPost Phys.* **10** (2021) 061 [`2009.00033`].

[96] N. Seiberg, *Analytic Study of θ Vacua on the Lattice*, *Phys. Lett. B* **148** (1984) 456.

[97] M. Abe, O. Morikawa, S. Onoda, H. Suzuki and Y. Tanizaki, *Topology of SU(N) lattice gauge theories coupled with $\mathbb{Z}_N$ 2-form gauge fields*, *JHEP* **08** (2023) 118 [`2303.10977`].

[98] J.F. Martins and R. Picken, *The fundamental Gray 3-groupoid of a smooth manifold and local 3-dimensional holonomy based on a 2-crossed module*, *Differential Geometry and its Applications* **29** (2011) 179 [`0907.2566`].

[99] T. Porter, *The Crossed Menagerie: an introduction to crossed gadgetry and cohomology in algebra and topology*, lecture notes being regularly updated since 2007.

[100] J.C. Baez and A.D. Lauda, *Higher-dimensional algebra V: 2-groups*, `math/0307200`.

[101] J. Thierry-Mieg, *Geometrical reinterpretation of Faddeev-Popov ghost particles and BRS transformations*, *J. Math. Phys.* **21** (1980) 2834.

[102] R. Brown and P.J. Higgins, *The equivalence of ∞-groupoids and crossed complexes*, *Cahiers de topologie et géométrie différentielle* **22** (1981) 371.

[103] R. Brown and P.J. Higgins, *Colimit theorems for relative homotopy groups*, *Journal of Pure and Applied Algebra* **22** (1981) 11.

[104] H.X. Sinh, *Gr-catégories*, Université Paris VII doctoral thesis (1975).

[105] R. Gordon, A.J. Power and R. Street, *Coherence for tricategories*, *American Mathematical Soc.* **117** (1995) .

[106] L. Kong and Z.-H. Zhang, *An invitation to topological orders and category theory*, `2205.05565`.

[107] J.-L. Brylinski, *Loop spaces, characteristic classes and geometric quantization*, Springer (1993).

[108] M.K. Murray and D. Stevenson, *Bundle gerbes: Stable isomorphism and local theory*, *J. Lond. Math. Soc.* **62** (2000) 925 [`math/9908135`].

[109] G. Segal, *Cohomology of topological groups*, in *Symposia Mathematica*, vol. 4, pp. 377–387, Academic Press London, 1970.

[110] J.-L. Brylinski, *Differentiable cohomology of gauge groups*, `math/0011069`.

[111] A. Kapustin and N. Sopenko, *Anomalous symmetries of quantum spin chains and a generalization of the Lieb-Schultz-Mattis theorem*, `2401.02533`.

[112] J. Gladikowski and M. Hellmund, *Static solitons with nonzero Hopf number*, *Phys. Rev. D* **56** (1997) 5194 [`hep-th/9609035`].

[113] L.D. Faddeev and A.J. Niemi, *Knots and particles*, *Nature* **387** (1997) 58 [`hep-th/9610193`].

[114] K. Osterwalder and R. Schrader, *Axioms for Euclidean Green's functions*, *Commun. Math. Phys.* **31** (1973) 83.

[115] K. Osterwalder and R. Schrader, *Axioms for Euclidean Green's functions II*, *Commun. in Math. Phys.* **42** (1975) 281.

[116] B. Swingle, *Entanglement Renormalization and Holography*, *Phys. Rev. D* **86** (2012) 065007 [`0905.1317`].

[117] A. Kitaev and L. Kong, *Models for Gapped Boundaries and Domain Walls*, *Commun. Math. Phys.* **313** (2012) 351 [`1104.5047`].

[118] T. Balaban, *Ultraviolet Stability of Three-Dimensional Lattice Pure Gauge Field Theories*, *Commun. Math. Phys.* **102** (1985) 255.

[119] T. Balaban, *Large Field Renormalization. 1: The Basic Step of the R Operation*, *Commun. Math. Phys.* **122** (1989) 175.

[120] T. Balaban, *Large Field Renormalization. 2: Localization, Exponentiation, and Bounds for the R Operation*, *Commun. Math. Phys.* **122** (1989) 355.