Modular forms and universality classes of topological matter

Kristian Stølevik Olsen

Thesis for the degree Master of Science
in Theoretical Physics.

Department of Physics
Faculty of Mathematics and Natural Sciences
University of Oslo

May 2017

Abstract

Investigations of available data yielding renormalization group flows for the quantum Hall system suggests that a strong holomorphic discrete symmetry is present in its parameter space. The geometry of this flow is intimately connected with the theory of automorphic forms for the corresponding symmetry groups. Up to conjugacy, these groups are subgroups of the modular group $\mathrm{SL}_2(\mathbb{Z})$. In this thesis we explore the connection between the theory of modular forms and universality classes of quantum Hall systems. We have studied the renormalization group flow of the Hall effect in a wide range of new materials, which fits remarkably well with the hypothesis of modular symmetry. This modular symmetry suggests the use of modular symmetric effective field theories to describe each universality class. We discuss connections between the topological understanding of the quantum Hall effect and the low-energy renormalization group fixed points of these modular symmetric theories. There is also a rich mathematical structure underlying these symmetries which is explored. In particular we discuss how the theory of modular forms can be presented clearly by considering it as a differential geometry tailored to fit the modular groups.

Acknowledgements

First and foremost I would like to thank my supervisor Carsten Andrew Lütken for his guidance and patience over the past two years. Without our long and insightful discussions this thesis would not have been possible. I would also like to thank Henrik S. Limseth for invaluable discussions as they relate to the topics of this thesis and physics in general.

The completion of this thesis would have been difficult were it not for the support of friends at the Department of Physics. A special thanks goes to Ask J. Markestad and Sean B. S. Miller for stimulating discussions and encouragement over the past years. I have also benefited greatly from conversations with Vidar Skogvold and Kine Ø. Hanssen, which have been constant sources of advise. Lastly I would like to thank the theory group for providing a great environment both socially and professionally.

Contents

Preface

One of the most remarkable facts about Nature is how simplicity and complexity is interwoven. The physical laws governing the behavior of particles on a microscopic level can be elegantly formulated on a single line on paper, as long as the right language is being used. As we consider larger and larger systems this simplicity turns into complexity where the behavior is hard to predict without the aid of numerics. Surprisingly, as the system grows larger still simplicity emerges. This simplicity emerges from the conspiracy of the systems microscopic constituents to produce new effective degrees of freedom in the large limit. If we are open to the idea that every system could in fact be emerging from some other system, we also have to be open to the idea that there is no such thing as a phenomena more fundamental than another. There is simply physics taking place on every scale of Nature.

In condensed matter the system under consideration is often some electron system in a complicated environment. Most important of the large-scale properties of these systems are the different phases in which the system can exist. The identification and classification of these phases of matter is a central problem in modern theoretical physics. As in any classification problem there are two main ingredients: a definition of the objects to be classified and a notion of equivalence. Different types of phases can come from different notions of equivalence. In recent decades a lot of effort has been put into the classification of topological phases. The classification most successful is perhaps that of topological insulators, which corresponds to topological classes of Hamiltonians describing gapped single-particle fermions.

The quantum Hall effect is an example of a 2-dimensional topological phase of matter, where the conductivity takes rational values in natural units. The integer effect [82] is in fact a topological insulator phase that has generalizations in every even spatial dimensions and is classified by a topological integer invariant [73]. While the classification of non-interacting topological insulators is more or less complete [43][69], there is still much work to be done regarding their interacting counterparts. The interacting version of the integer Hall effect is the fractional Hall effect [81], which can be understood trough for example effective topological quantum field theories. However, in spite of much work no model that predicts all Hall phases and characterize their transitions exist.

According to common lore, the large-distance properties of a phase is contained in an effective field theory obtained by integrating out high-energy degrees of freedom. This renormalization procedure can be used to probe universal features of a system. In general, two systems are said to lie in the same universality class if their low-energy (or equivalently large-distance) properties coincide. These universality classes are expected to be determined by general properties like spacetime dimension and symmetry groups. It is not expected that all microscopic details play any key role, as these short-distance phenomena are averaged over. For example, the quantum Hall effect has been observed in a wide range of materials which suggest that microscopic lattice structure is irrelevant as far as general properties of charge transport is concerned.

Experiments suggest that there is a strong modular symmetry in the scaling data of the Hall effect [54][55][63]. This has motivated the conjecture of a modular duality in the corresponding effective field theory [52][53][48][49]. The symmetries in the data corresponds to subgroups of the modular group $\mathrm{PSL}_2(\mathbb{Z})$, up to conjugacy. These emergent symmetries put so strong constraints of the effective theory that they in fact fix the renormalization group flow globally in the parameter space [49]. In this way, the different modular subgroups can be used to label different universality classes of quantum Hall effects as they encode both IR fixed points and critical points. This situation is somewhat akin to that in particle physics before the realization that particles could be classified according to symmetry groups. Here we are dealing with different types of quantum Hall effects in different 2-dimensional materials, which seemingly can be placed in a small handful of universality classes. The purpose of this thesis is, simply put, to explore these universality classes - the size of the classes, the connection to known models and their mathematical origin.

At the end of the day, we will have achieved nothing if our theoretical models have made no contact with reality. Hence, the most important way we study the universality classes of Hall effects is by studying scaling data from experiments in different materials. We observe renormalization group flows compatible with the modular groups in all cases, which strongly backs up the modular hypothesis. The results from this phenomenological study is presented in a joint paper (in preparation) with C. A. Lütken and H. S. Limseth appended at the end of this thesis.

There is also very rich mathematical structure underlying these modular symmetries that we want to explore. While there has always been a synergy between mathematics and physics, it has perhaps never been greater than in the past four decades. This has been to the benefit of both parties in surprising ways. Physics has always made use of mathematics as a tool, but it can also be a strong driving force for physical insight. In this way physics offers a natural setting in which many mathematical structures and ideas can be realized. Physical insight may then take the mathematics in an unexpected direction, just as the mathematical formulation of a physical problem can present new physical insight.

The mathematical toolbox of physicists have expanded rapidly in the last decades. A perfect example of this is the methods used to understand the topological phases of matter. Here mathematical ideas from differential geometry, topological K-theory lay the foundation for our theoretical understanding. These are tools that have been seen as a part of the abstract world of pure mathematics, but have now made their way into the real world through physics. The importance of these tools was also recognized in the 2016 Nobel prize in physics where the original work borrows ideas from the topology of vector bundles. This is a great example of the fact that there is really no limit to the mathematical ideas that a physical model can incorporate. At the end of the day, it is a matter of using the right tools for the job.

Symmetries are deeply connected with the mathematical theory of groups and their representations. In itself a group is an abstract concept that can be realized in many different ways depending on what objects the group is to act on. Since a modular symmetry is observed in the quantum Hall data, these groups must in a theoretical model be realized on the Hall parameter space. When acting on geometric objects on this parameter space, a connection to the mathematical theory of modular forms appears. In fact, under very general physical assumptions the experiments force the modular forms upon us.

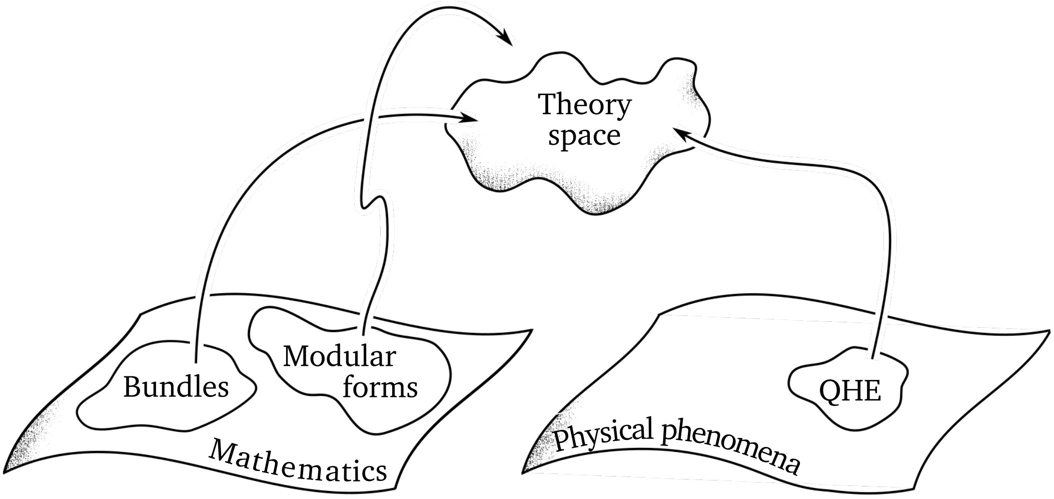


Figure 1: Generally a theory results from both physical and mathematical input. Different classes of physical phenomena may need different mathematical ideas to be modeled. The theoretical aspects of the quantum Hall system explored in this theses uses as mathematical input vector bundles and modular forms. While vector bundles have made their way into the physics toolbox, the theory of modular forms remains unknown to most physicists not familiar with string theory.

We explore both the mathematical theory of these modular forms as well as their appearance in physical models for the Hall effect. This is meant to complement the experimental analysis presented in the appended paper. Summarized, the goals of this thesis are the following.

(i) To formulate the theory of modular forms in a geometric language based on tensors and connections. While this is interesting in its own right, we also focus on a clear and pedagogical presentation of the subject.

(ii) To understand the interplay between universality and duality in mathematical modeling of Nature. In particular we study these ideas in the context of quantum field theory.

(iii) To connect the microscopic understanding of the different topological Hall phases with the phenomenological approach based on modular symmetries.

(iv) To explore the universality classes classified by modular symmetries by explicitly studying scaling data from quantum Hall experiments. This of course connects with the theoretical expectations in (iii) to some degree.

Outline

This thesis is divided into three parts that to a large extent can be read individually. Part I is devoted to the study of geometric structures and a geometric understanding of the theory of modular forms. The intention is to introduce modern tools relevant for the study of the Hall effect and its cousins. The reader familiar with these subjects is welcome to skip ahead to the next part. Most of our discussions will make use of geometry in one way or another. In particular we are going to need the ideas of fiber bundles and geometric structures. Most important of the structures will be the notions of Riemannian, complex and spin structures on a manifold. We therefore include a hopefully self-contained introduction to these topics in chapter 1. Using these geometric ideas we discuss modular forms in chapter 2. This chapter presents the theory of modular forms in an almost purely geometric manner, starting from the classification of complex tori with spin structure.

The main theoretical tool we use in modern physics is quantum field theory. Together with the ideas of universality classes and dualities, they offer a powerful approach to study both high and low-energy systems. Particularly important for our discussions will be the renormalization group flow, and how it behaves when discrete duality groups acts on the parameter space of a theory. Being the whole raison d'être of our discussions and results, we include a detailed discussion of quantum field theories, renormalization group flows and dualities in chapters 3, 4 and 5. These chapters comprise part II of the thesis.

Part III discuses the quantum Hall effects as topological phenomena and their relation to the modular subgroups. The two main classes of systems we consider are topological insulators and Dirac materials, where the low-energy excitations are Dirac fermions. We review the geometric and topological aspects of these phases of matter, with particular focus on the systems response to electric fields. These responses are the conductivities of the system, which coincide with the RG fixed points of the effective field theories with modular duality. To connect theory and experiment, we analyze charge transport data in 2-dimensional systems to check for modular symmetries. We observe the modular symmetry in a wide range of materials, which is a reflection of the universality of the Hall effect. This main result can be found in the aforementioned paper appended as the last part of this thesis.

Part I

Geometric structures and modular forms

1

Aspects of geometry

An interplay between mathematics and physics have always existed, where one side benefits from the connection to the other. A clear example of this would be Riemannian geometry and the theory of general relativity. After quantum mechanics developed in the 20s and 30s there was little interplay between mathematics and physics for almost four decades. In the 70s a geometric understanding of quantum field theories akin to the formulation of general relativity presented itself in the form of gauge theories. After this the connection between physical and mathematical theory has flourished to the benefit of both sides. Many topics long considered too abstract to have any physical relevance has been shown to be intimately connected with physical problems. This chapter is intended to be a self-contained introduction to several aspects of geometry with relevance in physics. In upcoming chapters we will use the geometric language presented here extensively.

1.1 Manifolds and smooth structure

A manifold is a topological space with local patches homeomorphic to Euclidian space. In some sense we define a more complicated object by demanding that locally it resembles a familiar object. This idea of local triviality will be a recurring theme in our discussions of geometry. Formally we say that a n-dimensional (real) manifold is a topological space M together with a collection of open subsets $\{U_i \subset M\}$ with homeomorphisms $\phi_i : U_i \to V_i \subset \mathbb{R}^n$ [60]. These homeomorphisms provide local coordinates on the manifold $\phi_i(p) = x^\mu = (x^1, ..., x^n)$. We demand that the open subsets U_i cover the manifold in the sense that $\cup_i U_i = M$. The tuple (U_i, ϕ_i) of a subset and the corresponding homeomorphism is called a chart, and the collection $\{(U_i, \phi_i)\}$ of charts is called an atlas $\mathcal{A}$. A manifold M may have a boundary, denoted ∂M. This is the case if there are point in M with local neighborhood homeomorphic to $\{(x^1, ..., x^n) | x^n \geq 0\}$. A manifold is said to be closed if it has no boundary $\partial M = 0$.

In the case where $U_j \cap U_i \neq \varnothing$, so that there are points that can be described by two different coordinate charts ϕ_j and ϕ_i, we define the transition function (also called coordinate transformation function) of these patches as

$$\psi_{ji} = \phi_j \circ \phi_i^{-1} : \mathbb{R}^n \to \mathbb{R}^n$$

and require these at least to be homeomorphisms. More intuitively, they deform the patches in Euclidian space into each other, telling us how the patches are glued together to construct M. Usually we impose stronger requirements on the transition functions, which turns out to be a nice way to give our manifold additional structure. We will study several additional structures later.

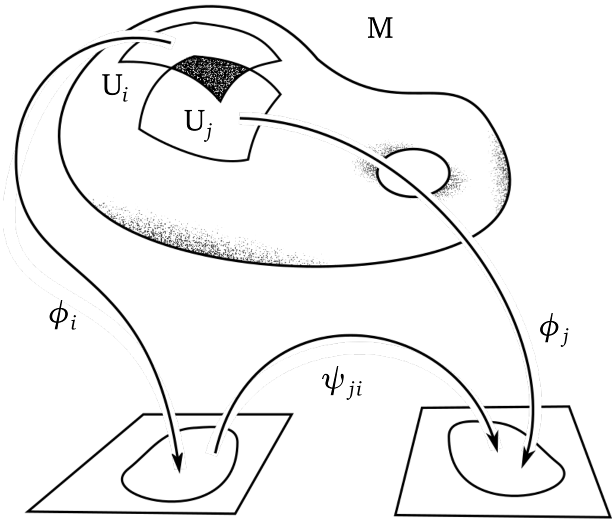


Figure 1.1: A manifold is covered in patches that provide local coordinates. Transition functions provide a way to go from one patch to the next.

For now, we only need one kind of structure on our manifolds, namely smooth structure. This is achieved by demanding that the transition functions are smooth. If every transition function in an atlas $\mathcal{A}$ is smooth we call the atlas a differentiable or smooth atlas. The manifold in question is then called a differentiable manifold, and is said to be given a smooth structure. This smooth structure enables us to talk about differentiable maps, and the notion of diffeomorphisms of manifolds. More importantly, a smooth structure gives a *unambiguous* notion of smoothness of functions. We let M and N denote two manifolds, of dimension m and n respectively. We let $f : M \to N$ be a mapping between these. We let (U, ϕ) be some chart on M and (U', ϕ') a chart on N. The composed mapping

$$f^{\alpha}(x^{\mu}) = (\phi' \circ f \circ \phi^{-1})(x^{\mu}) \tag{1.1}$$

($\mu = 1, ..., m$, $\alpha = 1, ..., n$) is called the coordinate representation of f. The function f is then called smooth if the coordinate representation is smooth. If the point p lies in the intersection $\mathrm{U}_1 \cap \mathrm{U}_2$ we have two possible coordinate representations, as illustrated in the below figure.

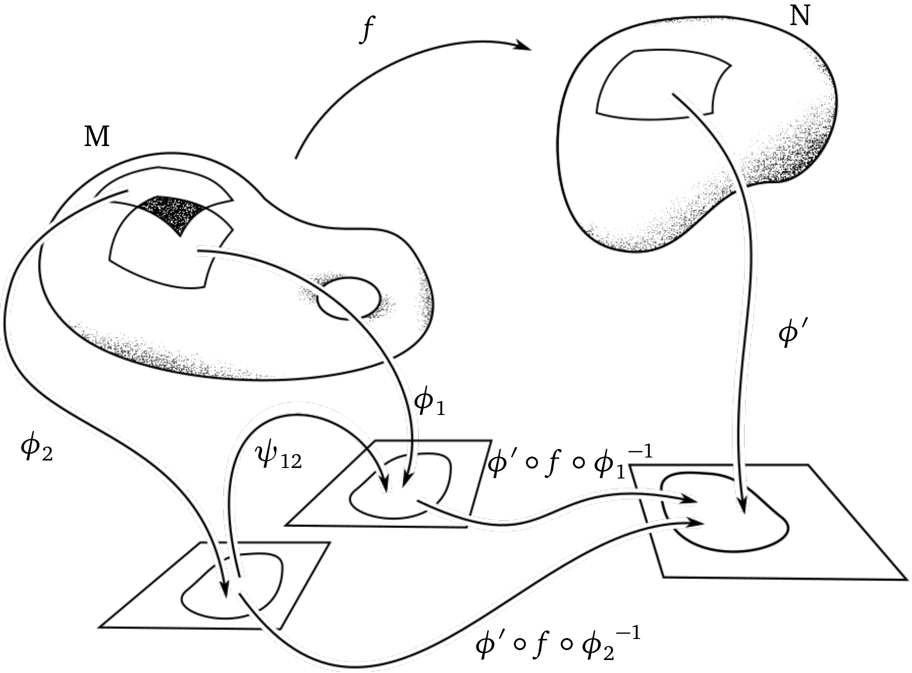

f
N
M
ϕ'
ϕ_1
ϕ_2
ψ_{12}
$\phi' \circ f \circ \phi_1{}^{-1}$
$\phi' \circ f \circ \phi_2{}^{-1}$

Figure 1.2

In the chart of U_2 our mapping f takes the form $\phi' \circ f \circ \phi_2^{-1}$. By inserting a identity map $\phi_1^{-1} \circ \phi_1$ we get the representation $\phi' \circ f \circ \phi_1^{-1} \circ (\phi_1 \circ \phi_2^{-1})$ which we recognize as the coordinate representation in U_1 composed with the transition function $(\phi_1 \circ \phi_2^{-1})$. Since, for differentiable manifolds, the transition functions are smooth we see that the coordinate representation of f in U_1 is differentiable only if the representation in U_2 is as well. Hence there are no ambiguities when it comes to the notion of differentiability of a function, given a smooth structure.

Two manifolds related by a smooth map with a smooth inverse are called diffeomorphic [60]. This may be seen as an equivalence relation stronger than homeomorphisms, in the sense that we require now not only continuity but also smoothness when we deform our spaces. The set of smooth manifolds together with smooth maps between them form the category Mfd.

Before we move on, we have some remarks.

(i) The smooth structure is strictly speaking defined to be a equivalence class of smooth atlases. However, since a single representative from this class is sufficient to construct the whole class (given the equivalence relation), one rarely needs to construct more that one atlas when it comes to practical calculations.

(ii) The study of geometry, at least from the perspective in this thesis, may be regarded as the study of the underlying topological space given some extra structure. "Structure" may intuitively be thought of as a sort of help to access more information than before. For example, it may be that we are able to discuss distances or angles which we so far have not been able to define. For now we have only met two sort of structures. First, we have covered our topological space with charts, enabling us to lay down coordinates. Secondly, the smooth structure enabled us to discuss smoothness. We will meet a lot of additional structure later on.

(iii) By composing maps we can construct coordinates for objects on the manifold. For example if we have a curve $\gamma : \mathbb{R} \to \mathrm{M}$ parametrized by $\lambda \in \mathbb{R}$ there is a natural coordinate for the point $\gamma(\lambda)$, namely $x^{\mu}(\lambda) = (\phi \circ \gamma)^{\mu}$. Sometimes one sees the notation γ^{μ} to remind us that this is a local coordinate expression for the points along the curve. Similarly we could place higher dimensional objects on M and find their local coordinates.

Given two manifolds M and N there is a simple way to construct a third, called the product manifold $\mathrm{M} \times \mathrm{N}$. Suppose M has an atlas $\{(u_i, \phi_i)\}$ and N the atlas $\{(v_i, \psi_i)\}$. Then, by considering the Cartesian products $u_i \times v_j$ we can construct the atlas $\{(u_i \times v_j, (\phi_i, \psi_j))\}$ on the product manifold. In the same way as manifolds were modeled on Euclidian space, we will encounter manifolds modeled on these product spaces shortly.

The cobordisms are a special class of manifolds. For two closed n-manifold N, N' a cobordism is a (n+1)-manifold M with $\partial \mathrm{M} = \mathrm{N} \sqcup \mathrm{N}'$. In this case N and N' are called cobordant ("jointly bounding"). This defines as equivalence relation of manifolds $\mathrm{N} \sim \mathrm{N}'$. Clearly every closed manifold M is cobordant to itself since it sits of the boundary of the manifold $\mathrm{M} \times [0,1]$. With this notion of equivalence of manifolds we can in fact realize the cobordisms as a category $nCob$. Here the objects are the manifolds N, N' while the morphisms are cobordisms M between objects.

1.2 Tensors

Having defined what a manifold is and some basic properties, we are ready to discuss tensors. This starts with the construction of the tangent spaces and their duals. The construction of the tangent spaces may be seen as a way of locally making sense of the notion of a vector on a manifold. As we will see this construction depends only on the smooth structure, but may be enhanced by adding further structure. For example, given a Riemannian structure one can canonically promote the tangent spaces to inner product spaces.

The tangent space $\mathrm{T}_p\mathrm{M}$ at a point p is identified with directional derivative operators along curves on the manifold passing trough p [60]. To the end of making sense of this let us consider a curve $x^\mu(\lambda)$ on our manifold that passes trough some point p, and a second curve $x^\mu(\xi)$ also passes trough p. Obviously the set of operators $\frac{d}{d\lambda}, \frac{d}{d\xi}$ etc will be closed under scalar multiplication and addition, and they are manifestly linear operators. In other words, they constitute a vector space. We consider the point p and a chart (U, ϕ) containing p which gives us n local coordinates x^μ.

Let $f : \mathrm{M} \to \mathbb{R}$ be a real valued function on M. With a slight abuse of notation we denote

$$\frac{df}{d\lambda} = \frac{d}{d\lambda}(f \circ \gamma). \tag{1.2}$$

Using the chain rule we have

$$\frac{d}{d\lambda}f = \frac{d(\phi \circ \gamma)^\mu}{d\lambda}\frac{\partial(f \circ \gamma)}{\partial(\phi \circ \gamma)^\mu} = \frac{dx^\mu(\lambda)}{d\lambda}\partial_\mu f \tag{1.3}$$

where the summation convention is implied. Thus, we identify the operators

$$\frac{d}{d\lambda} = \frac{dx^\mu}{d\lambda}\partial_\mu. \tag{1.4}$$

In other words, as a vector every directional derivative along some curve can be decomposed into a linear combination of the partial derivatives acting on f, with vector components $dx^\mu/d\lambda$. This basis $\{\partial_\mu\}$ is sometimes called the coordinate basis or standard basis for $\mathrm{T}_p\mathrm{M}$. In principle, we can perform transformations to another basis if this is more convenient.

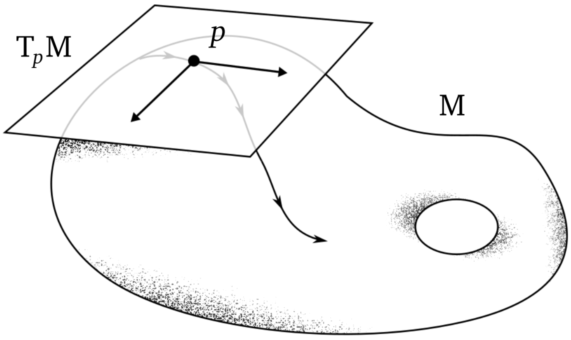


Figure 1.3: In the case of two dimensions one may visualize the tangent space as a actual space tangential to the surface. Note however that the definition of T_pM does not depend on a larger surrounding space.

We note that under some coordinate change $x^\mu \to \xi^\mu$ the basis vectors transform

$$\partial_\mu = \frac{\partial}{\partial x^\mu} \to \frac{\partial}{\partial \xi^\mu} = \frac{\partial x^\rho}{\partial \xi^\mu}\frac{\partial}{\partial x^\rho}. \tag{1.5}$$

Some vector $v = v^\mu \partial_\mu$ thus transforms to $v^\mu \partial_\mu \to (v')^\mu \partial_\mu{}' = (v')^\mu \frac{\partial x^\rho}{\partial \xi^\mu}\partial_\rho$. Thus we identify $(v')^\mu = \frac{\partial \xi^\mu}{\partial x^\rho} v^\rho$, which transforms inversely to the basis vectors in order to maintain invariance of the vector v. This is what we mean by a vector transformation. Note also that the vector components may be calculated by the action of the vector on the coordinates $v(x^\rho) = v^\mu \partial_\mu x^\rho = v^\rho$.

For two vectors $X = X^\mu \partial_\mu$ and $Y = Y^\mu \partial_\mu$ we define the Lie bracket [60]

$$[\cdot,\cdot] : T_pM \times T_pM \to T_pM$$

by the action on a function f:

$$[X,Y]f \equiv X(Y(f)) - Y(X(f)).$$

Writing this in the local coordinates we find

$$\begin{aligned}[X,Y]f &= x^\mu \partial_\mu (Y^\rho \partial_\rho f) - Y^\mu \partial_\mu (X^\rho \partial_\rho f) \\ &= (X^\mu \partial_\mu Y^\rho - Y^\mu \partial_\mu X^\rho)\partial_\rho f\end{aligned}$$

so we identify the vector components $[X,Y]^\rho = X^\mu \partial_\mu Y^\rho - Y^\mu \partial_\mu X^\rho$. For the moment the Lie bracket is of little importance, but later it will simplify our discussions.

As in linear algebra we may now construct the dual space T_p^*M to the tangent space at p as the space of linear maps $\omega : T_pM \to \mathbb{R}$. We call this the cotangent space [60]. A dual vector $\omega \in \mathrm{T}_p^*\mathrm{M}$ is also called a one-form or cotangent vector. The gradient of a function f is the canonical example of a one-form. We denote the gradient by df, and it is defined by the action

$$df\left(\frac{d}{d\lambda}\right) = \frac{df}{d\lambda}. \tag{1.6}$$

Just as the partial derivatives of the n coordinates x^μ provided a natural basis for $\mathrm{T}_p\mathrm{M}$, the gradients dx^μ will provide us with a natural basis for the cotangent space, since

$$dx^\mu(\partial_\nu) = \delta^\mu_\nu. \tag{1.7}$$

Just as we expanded an element $v \in \mathrm{T}_p\mathrm{M}$ as $v = v^\mu \partial_\mu$ we expand a general one-form as $\omega = \omega_\mu dx^\mu$. We may now define the inner product as a map $\langle \cdot, \cdot \rangle$: $\mathrm{T}_p^*\mathrm{M} \times \mathrm{T}_p\mathrm{M} \to \mathbb{R}$ by

$$\langle \omega, v \rangle = \omega_\mu v^\nu dx^\mu(\partial_\nu) = \omega_\mu v^\nu \delta^\mu_\nu = \omega_\mu v^\mu. \tag{1.8}$$

We note that a inner product is not defined for two vectors, but a one-form and a vector. We will shortly discuss these inner products in more detail. With the definition of one-forms and cotangent spaces in mind we can now move on to describing general tensors. A (p,q)-tensor or equivalently a tensor of type (p,q) is a multilinear map [60]

$$\mathrm{T} : \bigotimes^p \mathrm{T}_p^*\mathrm{M} \bigotimes^q \mathrm{T}_p\mathrm{M} \to \mathbb{R} \tag{1.9}$$

taking the form

$$\mathrm{T} = \mathrm{T}^{\mu_1 \dots \mu_p}{}_{\nu_1 \dots \nu_q} \partial_{\mu_1} \otimes \dots \otimes \partial_{\mu_p} \otimes dx^{\nu_1} \otimes \dots \otimes dx^{\nu_q}. \tag{1.10}$$

Its action on vectors and one-forms v_i, ω_i is

$$\mathrm{T}(\omega_1, \dots, \omega_p; v_1, \dots, v_q) = \mathrm{T}^{\mu_1 \dots \mu_p}{}_{\nu_1 \dots \nu_q} \omega_{1\mu_1} \dots \omega_{p_{\mu_p}} v_1^{\nu_1} \dots v_q^{\nu_q}. \tag{1.11}$$

We note that a (p,q)-tensor is defined at a point $p \in \mathrm{M}$ since it is built from elements on $\mathrm{T}_p\mathrm{M}$ and $\mathrm{T}_p^*\mathrm{M}$. To discuss for example vector fields, as objects extending globally on a manifold, we need the notion of a fiber bundle.

1.3 The theory of fiber bundles

Just as in the case of a manifold, a fiber bundle will be constructed by locally reducing it to a known object. Roughly speaking a fiber bundle is a generalization of the product space in the same way that manifolds generalize Euclidian space.

In this section we briefly explore such bundles. This section is inspired by [37] and [40].

To construct a fiber bundle, we pick a fiber F and the base manifold M. The fiber bundle can then be thought of as a continuous collection of the fibers F parametrized by the base M. Roughly speaking, one attaches a copy of the fiber at every point of the base. The fibers may be manifolds, vector spaces, et cetera. When the fibers are vector spaces, the fiber bundle is called a vector bundle. The new geometric object achieved in this manner is called the total space and is often denoted $\mathcal{E}$. From this total space one has a map $\pi : \mathcal{E} \to \mathrm{M}$ called a projection that intuitively collapses each fiber to its corresponding point, and will help us move between the base space and the total space. The preimage of the projection is isomorphic to the fibers of the bundle, i.e. $\pi^{-1}(x) = \mathrm{F}$.

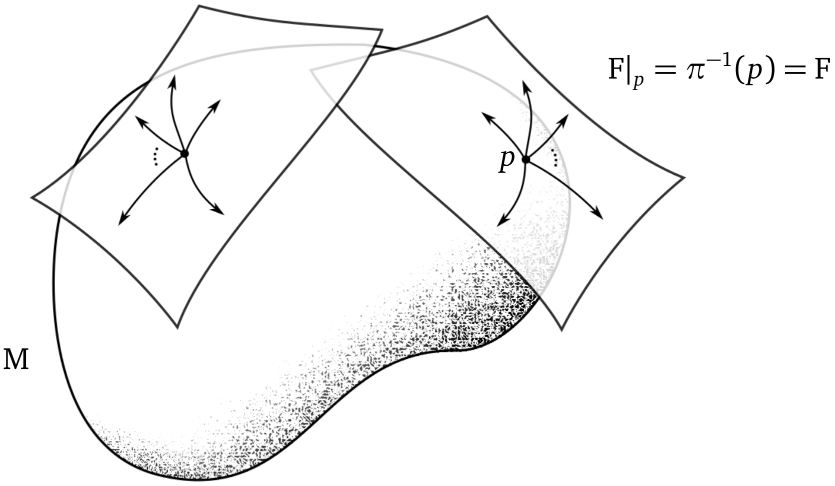


Figure 1.4: Schematic picture of a bundle. Over every point on M we have a fiber, here depicted as a n-dimensional vector space. The projection collapses these vector spaces to the point over which they are defined, and the inverse projection over a point gives back the fiber.

The simplest example of a fiber bundle is a product space $\mathcal{E} = \mathrm{M} \times \mathrm{F}$. On a point $(p, f) \in \mathrm{M} \times \mathrm{F}$ the projection map simply acts by $\pi((p, f)) = p$. Fiber bundles where the total space can be written as a product space is called trivial. We will meet different criteria for triviality soon. The construction of a more general fiber bundle then proceeds analogously to the construction of a manifold: Just as manifolds locally look like Euclidian space, fiber bundles locally look like product spaces. Thus we can define a fiber bundle to be the collection of data $(\mathrm{M}, \mathrm{F}, \pi, \mathcal{E})$, where if given a subset $U \subset M$ we have a isomorphism $\rho : \pi^{-1}(U) \simeq U \times F$. These isomorphisms are analogous to the charts of a manifold, "trivializing" the bundle locally. In the following we let $U_\alpha, U_\beta, U_\gamma$ denote three overlapping subsets of M. The isomorphisms $\rho_\alpha : \pi^{-1}(U_\alpha) \to U_\alpha \times F$ we write $\rho_\alpha = (\pi, \phi_\alpha)$ where now $\phi_\alpha : \pi^{-1}(U_\alpha) \to F$. This way we don't have to drag the base point around. Just as we needed transition functions to glue together overlapping charts on a manifold, we need to know how to glue together elements of the fibers on the intersections. To this end we define the transition map

$$\begin{aligned} \psi_{\beta\alpha} = \phi_\beta \circ \phi_\alpha^{-1} &: F \to F \\ &: f \to \psi_{\beta\alpha}(p)(f) \end{aligned}$$

where F is the fiber over $p \in M$. We want this map to have the following properties [40][60]

$$\psi_{\alpha\beta}(p) = {\psi_{\beta\alpha}}^{-1}(p)$$
$$\psi_{\alpha\beta}(p)\psi_{\beta\gamma}(p)\psi_{\gamma\alpha}(p) = \mathrm{id}_M$$

where the point p lies in the intersection of the subsets. The latter of these will sometimes be referred to as the cocycle condition. We will see that these conditions can be seen as consistency conditions for objects living in the fibers. If one is given the charts $\{U_\alpha\}$ covering M and the transition functions of the bundle, one can reconstruct the bundle by gluing together the local trivializations

$$\mathcal{E} = \bigcup_\alpha U_\alpha \times F/\sim$$

where we identify points $(p, f) \sim (p, \psi_{\beta\alpha}(p) f)$. The set of transition functions $\{\psi_{\alpha\beta}\}$ constitute a group called the structure group G of the bundle [40]. Note that the transition functions can't be arbitrary diffeomorphisms $\psi_{\beta\alpha} : U_\alpha \cap U_\beta \to \mathrm{Diff}(F)$, but has to respect the structure of the fibers. So, for example, if the fibers are finite dimensional vector spaces the transition functions take values in $GL_n(\mathbb{R})$, and if we have the structure of a inner product they have to be orthogonal $\psi_{\beta\alpha}(p) \in O(n)$. Whatever structure we have put on our fibers the transition functions must conserve when we "glue together" the entire bundle [40].

A section of a bundle is a map $s : M \to \mathcal{E}$, associating to every point in the base manifold a element of the fiber, such that the composed map $\pi \circ s$ is the identity map on the base. The space of sections is typically denoted $\Gamma(\mathcal{E})$. A local section is similarly a map $s_\alpha : U_\alpha \to \mathcal{E}$, which we often think of in a trivialization so that the local section is really a map from U_α to the fibers. Given a section s_α over U_α and s_β over U_β the transition functions map $s_\alpha = \psi_{\alpha\beta} s_\beta$. On the triple intersections we can write $s_\alpha = \psi_{\alpha\beta} s_\beta$, $s_\beta = \psi_{\beta\gamma} s_\gamma$, $s_\gamma = \psi_{\gamma\alpha} s_\alpha$ which implies the cocycle condition as a consistency rule.

As an example of a section we can consider the trivial bundle $\mathcal{E} = \mathrm{M} \times \mathrm{N}$. Here we have chosen as fibers a manifold N, for example the sphere or a torus. A section will in this case map

$$s : p \to (p, f(p))$$

Omitting the base point p we see that sections of a trivial bundle may be interpreted as N-valued functions of the base manifold $f : \mathrm{M} \to \mathrm{N}$. In this trivial case, we will sometimes use the notation $\Gamma(\mathrm{M} \times \mathrm{N}) = \mathrm{Maps}(\mathrm{M}, \mathrm{N})$.

As a more non-trivial example we can consider the Möbius bundle. Locally this bundle looks like the product space $\mathrm{S}^1 \times \mathbb{R}$, but globally has a 180 degree twist. In some sense it is a twisted version of the cylinder. We let U_1 and U_2 be subsets of the circle that covers slightly more that 180 degrees. On the overlap we have the structure group $\mathbb{Z}_2$ acting on the points (θ, r) as

$$\psi_{21} : (\theta, r) \to (\theta, -r).$$

If we in stead chose the structure group to be trivial, we would have a trivial bundle $\mathrm{S}^1 \times \mathbb{R}$, i.e. the cylinder. Both the cylinder and the Möbius bundle are examples of line bundles, e.g. vector bundles with one dimensional fibers.

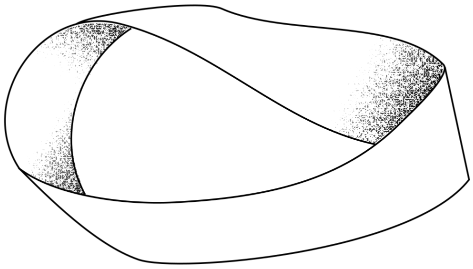


Figure 1.5: The Möbius bundle as a twisted version of a cylinder.

It is also interesting to study the sections of the Möbius bundle. Let $s : \mathrm{S}^1 \to$ Möb be a global section. After one circulation of the band, the transition function would have sent the section to its negative, in other words

$$s(\theta) = -s(\theta + 2\pi).$$

Hence the only well defined global section of the Möbius bundle is the zero section $\theta \to (\theta, 0)$.

By now it should be clear that the transitions functions holds some information regarding the topology of the bundle, in the sense that they give the rules as how to twist the fibers along the base. As in the case of the Möbius bundle, the global non-triviality, e.g. departure from product space structure, is captured by the transition functions. In particular, let $\mathcal{E}$ and $\tilde{\mathcal{E}}$ be two bundles over M with the same fibers F and structure group G. Over U_α we then have the homeomorphisms ϕ_α and $\tilde{\phi}_\alpha$ which enables us to define the composed homeomorphism

$$h_\alpha = \phi_\alpha \circ \tilde{\phi}_\alpha^{-1} : \mathrm{U}_\alpha \times \mathrm{F} \to \mathrm{U}_\alpha \times \mathrm{F}$$

that deforms the local trivializations into each other. Then, by inserting the identity transformation we see that

$$\tilde{\psi}_{\alpha\beta} = \tilde{\phi}_\alpha \circ \tilde{\phi}_\beta^{-1} = \tilde{\phi}_\alpha \circ \phi_\alpha^{-1} \circ \phi_\alpha \circ \phi_\beta^{-1} \circ \phi_\beta \circ \tilde{\phi}_\beta^{-1} = h_\alpha^{-1} \circ \psi_{\alpha\beta} \circ h_\beta.$$

Since the h's are local homeomorphisms, the non-triviality of the transition functions $\tilde{\psi}$ is in some sense inherited from that of ψ, and the bundles are considered topologically the same [9]. In physical applications this equivalence is often a gauge freedom.

For trivial bundles, it is clear that the structure group should be trivial, in the sense that the total space has no twists or turns. In the case of the Möbius band we saw that the twist meant that a global section was not well defined, unless it was the zero section. Conversely, a bundle is trivial if one can find a global section such that every fiber has a basis.

The prototypical example of a bundle is the tangent bundle and cotangent bundle. The tangent bundle TM over M is a vector bundle with typical fiber the tangent space $\mathrm{T}_p\mathrm{M}$. The projection map simply sends $\mathrm{T}_p\mathrm{M} \to p$. Sections of this bundle attaches to every point of the manifold a vector, i.e. creates a vector field. The local coordinates on the base manifold now also provide a local trivialization of the bundle [37], as we can locally express a vector field as $v = v^\mu \partial_\mu$ in the coordinate basis $\{\partial_\mu\}$. As we change patch on the base manifold we known that the basis vectors transform

$$\frac{\partial}{\partial x^\mu} \to \frac{\partial y^\nu}{\partial x^\mu}\frac{\partial}{\partial y^\nu} \equiv \psi^{\nu}_{\ \mu}\frac{\partial}{\partial y^\nu}$$

and hence also define the transition maps for the fibers. These are the general linear transformations $\mathrm{GL}_n(\mathbb{R})$. As a consequence of the fact that $dx^\mu(\partial_\nu) = \delta^\mu_\nu$ one can easily check that the cotangent bundle $\mathrm{T}^*\mathrm{M}$ has the inverse transformations as structure group.

Equipped with the tangent and cotangent bundle we can construct a myriad of bundles using vector space operations [37][40]. Given a vector bundle with total space $\mathcal{E}$ we can for example construct the bundles $\mathcal{E} \oplus \mathcal{E}$, $\mathcal{E} \otimes \mathcal{E}$ and continue this as many times we wish. These operations are defined by their operation on the fibers, which are the usual notions of direct sums and tensor products known from linear algebra. For example we can construct a tensor bundle as a vector bundle with typical fiber

$$\bigotimes^{p} \mathrm{T}_p^* \mathrm{M} \bigotimes^{q} \mathrm{T}_p \mathrm{M}.$$

We denote this bundle by $\mathrm{T}^{(p,q)}\mathrm{M}$. Clearly this bundle has as sections tensor fields, locally on the form of (p,q)-tensors.

Before we move on to discuss differential forms, some remarks are in order.

(i) In many physical applications, some of which we will encounter, the local framework provided by the tangent (cotangent) spaces are not sufficient. Bundles are the framework for dealing with such inherently global aspects. This being said, most calculations can, and will, be done locally.

(ii) Fiber bundles are also good tools for adding extra degrees of freedom. The manifold itself may for example represent the translational degrees of freedom of a particle. A fiber bundle over this space then provides a certain internal degree of freedom.

(iii) Note that it may be easy to mix up the transition maps on the manifold providing local change of coordinates and the transition function on the bundle as they are both denoted ψ. The transition maps on the level of the bundle tells us how to properly glue together the fibers just as the transition maps on the level of the manifold tells us how to perform coordinate changes.

(iv) Bundles where the fibers are one dimensional are called line bundles, as briefly mentioned in the discussion on the Möbius bundle. The standard symbol for these bundles are $\pi : \mathcal{L} \to \mathrm{M}$, and locally $\pi^{-1}(x) = \mathcal{L}_x$. For example, a trivial real line bundle is of the form $\mathrm{M} \times \mathbb{R}$, and the sections $s : x \to (x, f(x))$ are nothing but real valued functions.

1.4 Forms and de Rham cohomology

Differential forms of rank r arise as antisymmetric sections of the bundle $\mathrm{T}^{(0,r)}\mathrm{M}$ [60]. The bundle of such tensor fields is denoted $\wedge^r \mathrm{T}^*\mathrm{M}$. The local tensors are also referred to as r-forms for short. We denote the space of such tensors by $\Gamma(\wedge^r \mathrm{T}^*\mathrm{M}) = \Omega^r(\mathrm{M})$. The wedge product is defined as a map $\wedge : \Omega^r \times \Omega^s \to \Omega^{r+s}$ trough the action [60]

$$\omega \wedge \eta(v_1, ..., v_{r+s}) = \frac{1}{r!s!} \sum_{\mathrm{P}} sgn(\mathrm{P})\omega(v_{\mathrm{P}(1)}, ..., v_{\mathrm{P}(r)})\eta(v_{\mathrm{P}(r+1)}, ..., v_{\mathrm{P}(r+s)})$$

where ω is a r-form and η is a s-form. Given r one-forms we can construct the r-form [60]

$$dx^{\mu_1} \wedge dx^{\mu_2} \wedge \ldots \wedge dx^{\mu_r} = \sum_{\mathrm{P} \in \mathrm{S}_r} sgn(\mathrm{P}) dx^{\mu_{\mathrm{P}(1)}} \otimes \ldots \otimes dx^{\mu_{\mathrm{P}(r)}}. \tag{1.12}$$

These act as a basis for $\Omega^r(\mathrm{M})$, and a general element may be expanded as

$$\omega = \frac{1}{r!} \omega_{\mu_1 \ldots \mu_r} dx^{\mu_1} \wedge \ldots \wedge dx^{\mu_r}. \tag{1.13}$$

The factorial factor here is optional. It is clear from the basis elements that we choose r out of $\dim(\mathrm{M}) = n$ possible one-forms, making the dimension of $\Omega^r(\mathrm{M})$ equal $\binom{n}{r} = n!/(n-r)!r!$. Thus on a 4-manifold, for example, there will be one 0-form, four one-forms, six 2-forms,four 3-forms, and one 4-form. Also, on a 2-manifold there will be only one linearly independent antisymmetric rank 2 tensor, which we may think of simply as an antisymmetric 2×2 matrix. Clearly from the above definitions we have $\Omega^0(\mathrm{M})$ as the set of sections of the trivial real line bundle, i.e. space of smooth functions on M. Note also that the space of n-forms on a n-manifold is one dimensional, and hence $\wedge^n \mathrm{T}^*\mathrm{M}$ is a line bundle. This line bundle is called the canonical bundle.

Recall that from a function $f : \mathrm{M} \rightarrow \mathbb{R}$ we can create the one-form $df = \frac{\partial f}{\partial x^\mu} dx^\mu$. As we realized that real valued functions are members of $\Omega^0(\mathrm{M})$ we can abstractly view the transition from f to df as a map

$$\begin{aligned} d : \Omega^0(\mathrm{M}) &\rightarrow \Omega^1(\mathrm{M}) \qquad (1.14)\\ : f &\rightarrow df. \qquad (1.15) \end{aligned}$$

In this way there is a natural relation between the sections of the bundles $\wedge^0 \mathrm{T}^*\mathrm{M}$ and $\wedge^1 \mathrm{T}^*\mathrm{M}$. A question that naturally arises is then if this behavior continues. The suitable generalization should be a map $d : \Omega^r(\mathrm{M}) \rightarrow \Omega^{r+1}(\mathrm{M})$, which we define trough the formula [60]

$$d\omega \equiv \frac{1}{r!} \left(\frac{\partial}{\partial x^\nu} \omega_{\mu_1 \ldots \mu_r} \right) dx^\nu \wedge dx^{\mu_1} \wedge \ldots \wedge dx^{\mu_r}. \tag{1.16}$$

Here, as above, the factorial is a optional normalization factor. This sort of differentiation on forms is called an exterior derivative. The composed operation $d^2 = d \circ d$ will then be a map $\Omega^r(\mathrm{M}) \rightarrow \Omega^{r+1}(\mathrm{M}) \rightarrow \Omega^{r+2}(\mathrm{M})$ which takes the form

$$\begin{aligned} d^2\omega &= \frac{1}{r!} \left(\frac{\partial}{\partial x^\rho} \frac{\partial}{\partial x^\nu} \omega_{\mu_1 \ldots \mu_r} \right) dx^\rho \wedge dx^\nu \wedge dx^{\mu_1} \wedge \ldots \wedge dx^{\mu_r} \qquad (1.17)\\ &= -\frac{1}{r!} \left(\frac{\partial}{\partial x^\rho} \frac{\partial}{\partial x^\nu} \omega_{\mu_1 \ldots \mu_r} \right) dx^\nu \wedge dx^\rho \wedge dx^{\mu_1} \wedge \ldots \wedge dx^{\mu_r}. \qquad (1.18) \end{aligned}$$

However, since the partials commute we may simply rename indices to realize that $d^2\omega = -d^2\omega$, i.e. $d^2 = 0$. This leads to an interesting classification of forms as follows. A r-form is said to be closed if $d\omega = 0$ and exact if we can write $\omega = d\tilde{\omega}$, where $\tilde{\omega}$ is a $(r-1)$-form. Clearly all exact r-forms are also closed as $d^2 = 0$, but the converse is not necessarily true.

A lemma by Poincare gives us a situation where closed forms are exact. On a small neighborhood U of a manifold, any closed r-form is exact if U is contractable to a point [60]. To better understand the nontrivial solutions to $d\omega = 0$, we want to study the closed forms that are not exact. Consider the sequence of maps

$$\Omega^0(\mathrm{M}) \xrightarrow{d} \Omega^1(\mathrm{M}) \xrightarrow{d} \Omega^2(\mathrm{M}) \xrightarrow{d} \ldots \xrightarrow{d} \Omega^n(\mathrm{M}) \xrightarrow{d} 0.$$

We write 0 at the end of the sequence as is has to terminate: there are no forms of higher rank than the dimension of the manifold. At any stage in this sequence the image of the incoming map has to lie in the kernel of the outgoing map, since $d^2 = 0$, i.e

$$\mathrm{Im}(d) \subset \mathrm{Ker}(d).$$

Forms in the image are what we called exact, while those in the kernel are closed. To better understand closed forms one constructs the de Rham Cohomology as the quotient [60]

$$\mathrm{H}^r(\mathrm{M}) = \mathrm{Ker}(d)/\mathrm{Im}(d)|_{\Omega^r}$$

at some point of the sequence. Often one introduces even more notation and writes $\mathrm{Ker}(d)|_{\Omega^r} = \mathrm{C}^r(\mathrm{M})$ and $\mathrm{Im}(d)|_{\Omega^r} = \mathrm{D}^r(\mathrm{M})$. We may think of this as the study of equivalence classes, with

$$\mathrm{H}^r(\mathrm{M}) = \mathrm{C}^r/\mathrm{D}^r = \mathrm{C}^r/\sim$$

$$\omega \sim \sigma \;\text{ iff }\; \omega - \sigma \in \mathrm{D}^r$$

i.e. two elements of the closed forms C^r are cohomologically equivalent if they differ by an exact form. A figure summarizing this discussion is added below.

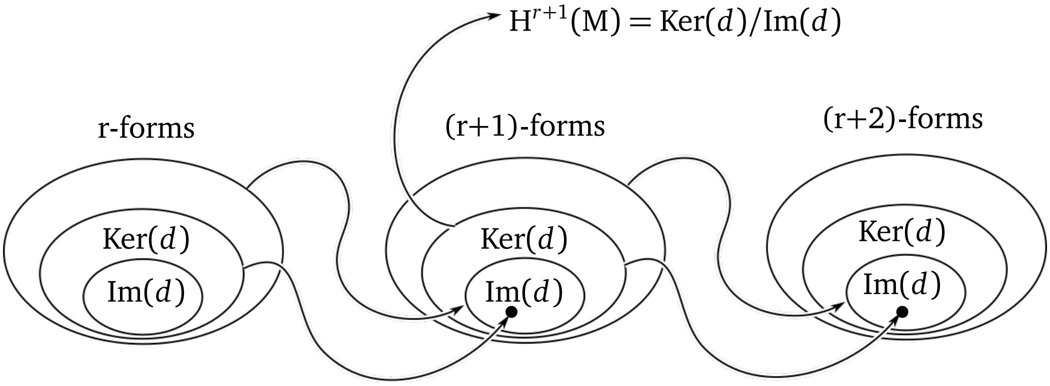


Figure 1.6: Schematics of the de Rham complex. All r-forms are mapped to the image of d, while the kernel is further mapped to 0 (symbolized by the dot). The de Rham cohomology group studies elements of Ker(d) modulo exact forms.

Integration can also be defined now that forms have been introduced. Abstractly, integration may be seen as a map

$$\int : \Lambda^p T^* M \to \mathbb{R}.$$

From above we known that the vector space of top dimensional forms have dimension one, i.e. its local sections are of the form

$$\omega = h(x) dx^1 \wedge ... \wedge dx^n.$$

Once Riemannian structure has been introduced, we will see that there is a natural choice of the prefactor $h(x)$ so that ω will be called a volume form $d\text{vol}_M$, and integration of a function $F : M \to \mathbb{R}$ is defined by integration of the form $f(x) d\text{vol}_M$, which reproduces the familiar notion of integration. A nice connection between integration and cohomology is presented by Stokes theorem. We know that given a $(n-1)$-form η we can make a form of one higher degree by $\omega = d\eta$. Integration of such forms are captured in the Stokes theorem [58][2]

$$\int_T d\eta = \int_{\partial T} \eta$$

where T is a subset of M and ∂T its boundary. As a consequence, we can in the case $\partial M = 0$ change a form ω to $\omega + d\theta$ while its integral remains unchanged. In this sense the integral associated a number to the whole cohomology class.

1.5 The pullback

Having discussed fiber bundles, and in particular the tensor bundles and forms, we now discuss some operations that can be performed on these. The idea is that a map between manifolds can be used to push forward and pull back geometric information. This will be crucial in our later study of the nonlinear sigma models. Let M_1, M_2, M_3 be manifolds and consider the maps

$$\phi : M_1 \to M_2,$$

$$f : M_2 \to M_3.$$

Then we define the pullback of f under ϕ as the map

$$\phi^* f \equiv f \circ \phi : M_1 \to M_3$$

which is useful for example if we have a function on M_2 that we want to pull back to a function on M_1 with the aid of some map ϕ. Of course this is nothing but new notation for the composition of maps. In field theoretical applications the map ϕ is often the field itself. To study the pullback of other more interesting geometrical objects, we first need the notion of the pushforward of a vector.

From earlier discussions we know that the role of vectors is to differentiate functions $f : N \to \mathbb{R}$. To see what a map $\phi : M \to N$ lets us do to vectors we study the sequence

$$M \xrightarrow{\phi} N \xrightarrow{f} \mathbb{R}.$$

We know how to pull the map $f : N \to \mathbb{R}$ back to a map $\phi^* f : M \to \mathbb{R}$, and we define pushforward as the map [2]

$$\phi_* : T_p M \to T_{\phi(p)} N$$

$$: V \to \phi_* V$$

where the defining property is that $(\phi_* V)(f) = V(\phi^* f)$. In other words, the pushed forward vector yields the same value when acting on f as the original vector does when acting on the pullback $\phi^* f$.

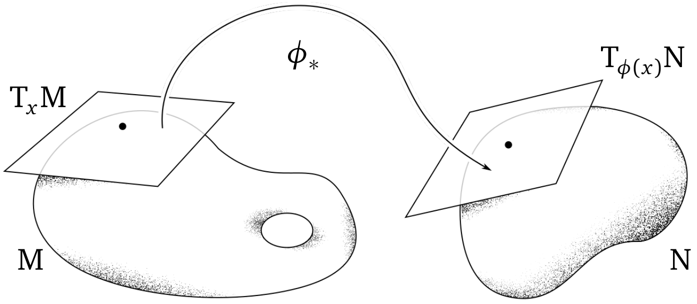


Figure 1.7: The pushforward map provide a way to induce new vector fields from old, given a map between manifolds.

To see what happens at the level of coordinates, we write out the vector $\phi_* \mathrm{V}$ in the basis of $\mathrm{T}_{\phi(p)}\mathrm{N}$ as follows:

$$\phi_* \mathrm{V} = \left[\phi_*(\mathrm{V}^i \frac{\partial}{\partial x^i})\right]^j \frac{\partial}{\partial (y^j \circ \phi)}$$

where x and y are coordinates on M and N respectively. We here note that the coordinates of interest on N are in some sense "parametrized" by the map ϕ, and we will from here use the notation $(y^j \circ \phi) = \phi^j$ [1]. As we noted in the discussion on tangent spaces, the components of vectors may be found by the action of the vector on the coordinates y^i, meaning in the present case that

$$\left[\phi_*(\mathrm{V}^i \frac{\partial}{\partial x^i})\right]^j = \left[\phi_*(\mathrm{V}^i \frac{\partial}{\partial x^i})\right](y^j) = \mathrm{V}^i \frac{\partial (y^j \circ \phi)}{\partial x^i} = \mathrm{V}^i \frac{\partial \phi^j}{\partial x^i}$$

where in the second equality we used the defining property of the pushforward.

The pushforward is in fact related to the differential of the map f. To see this, note that we may write the pushforward explicitly as the tensor product [40]

$$\phi_* = \frac{\partial \phi^\alpha}{\partial x^i} dx^i \otimes \frac{\partial}{\partial \phi^\alpha}.$$

This we can do as it clearly has the appropriate action on vectors:

$$\begin{aligned} \phi_* \mathrm{V} &= \phi_* \mathrm{V}^i \frac{\partial}{\partial x^i} \\ &= \frac{\partial \phi^j}{\partial x^i} dx^i (\mathrm{V}^a \frac{\partial}{\partial x^a}) \frac{\partial}{\partial \phi^j} \\ &= \mathrm{V}^a \frac{\partial \phi^j}{\partial x^i} \delta^i_{\ a} \frac{\partial}{\partial \phi^j} \\ &= \mathrm{V}^i \frac{\partial \phi^j}{\partial x^i} \frac{\partial}{\partial \phi^j}. \end{aligned}$$

By looking at the explicit form on the pushforward, we see that in the case $\mathrm{N} = \mathbb{R}$ it reduces to the well known differential of f. Hence we will also refer to φ_* as the differential df of the map $f : \mathrm{M} \to \mathrm{N}$.

We are now ready to discuss the pullback of more interesting objects. If $\pi : \mathcal{E} \to \mathrm{N}$ is a vector bundle over N, it may be pulled back to a vector bundle $\varphi^* \pi$:

[1] To be picky, what we mean is the following. If at $\phi(p)$ there is a chart ψ the coordinates of $\phi(p)$ is given by the components of the composed map $(\psi \circ \phi)(p)$, i.e. $(\psi \circ \phi)^i = (y^i \circ \phi)$. As we are interested not in general coordinates y^i on N but rather at the point $\phi(p)$ we denote the coordinates ϕ^i.

$\varphi^* \mathcal{E} \to M$ where by definition the fibers over $x \in M$ are equal to the fibers over N at $\varphi(x)$ [37]. A we may have realized by now, the notion of a pullback is, at least conceptually, pretty much the same as composition of maps. One can think of it as an elaborate precomposition scheme for non-function objects. The pullback of a function was defined simply as a precomposition, which we needed to define the pushforward. With these two simple ideas we can generalize pullback to higher tensors. The idea is to define the action of a pullback tensor by the action of the original tensor on pushed-forward vectors. Thus, for a (0,p)-tensor T at $\phi(p)$ on N we define [2]

$$(\phi^* T)(v_1, ..., v_p) \equiv T(\phi_* v_1, ..., \phi_* v_p)$$

where $\phi^* T$ is a tensor at $p \in M$. From this we can find the components of the pullback tensor by

$$(\phi^* T)_{\mu_1 ... \mu_p} = (\phi^* T)(\partial_{\rho_1}, ..., \partial_{\rho_p}) = T(\phi_* \partial_{\rho_1}, ..., \phi_* \partial_{\rho_p}).$$

Writing the tensor on the form $T = T_{i_1 ... i_p} d\phi^{i_1} \otimes ... \otimes d\phi^{i_p}$ and using the result from our pushforward discussion we have that

$$(\phi^* T)_{\mu_1 ... \mu_p} = \left(\frac{\partial \phi^{j_1}}{\partial x^{\rho_1}} \right) ... \left(\frac{\partial \phi^{j_p}}{\partial x^{\rho_p}} \right) T_{j_1 ... j_p}.$$

The definition of the pullback does not change for differential forms, but we need to incorporate the antisymmetry somehow. Let ω be a element in $\Lambda^n T^* M$ locally expanded as usual

$$\omega = \omega_{i_1 ... i_n} d\phi^{i_1} \wedge ... \wedge d\phi^{i_n}$$

and let us pull this back to a n-form $\phi^* \omega$ in $\Lambda^n T^* M$. The components are found by

$$(\phi^* \omega)_{\mu_1 ... \mu_n} = \omega(\phi_* \partial_{\mu_1} ... \phi_* \partial_{\mu_n}) = \left(\frac{\partial \phi^{j_1}}{\partial x^{\mu_1}} \right) ... \left(\frac{\partial \phi^{j_n}}{\partial x^{\mu_n}} \right) \omega_{i_1 ... i_n} d\phi^{i_1} \wedge ... \wedge d\phi^{i_n} (\partial_{j_1} ... \partial_{j_n}).$$

Hence we need to calculate $d\phi^{i_1} \wedge ... \wedge d\phi^{i_n} (\partial_{j_1} ... \partial_{j_n})$. To do this, we recall the definition of the wedge product:

$$d\phi^{i_1} \wedge ... \wedge d\phi^{i_n} = \sum_P \mathrm{sgn}(P) d\phi^{i_{P(1)}} \otimes ... \otimes d\phi^{i_{P(n)}}$$

$$\therefore d\phi^{i_1} \wedge ... \wedge d\phi^{i_n} (\partial_{j_1} ... \partial_{j_n}) = \sum_P \mathrm{sgn}(P) \delta^{i_{P(1)}}_{j_1} ... \delta^{i_{P(n)}}_{j_n}$$

Inserting this into our expression for the components of the pullback form, and using the definition of the determinant we find

$$(\phi^* \omega)_{\mu_1 ... \mu_n} = \omega_{i_1 ... i_n} \det \left(\frac{\partial \phi^{i_1}}{\partial x^{\mu_1}} ... \frac{\partial \phi^{i_n}}{\partial x^{\mu_n}} \right)$$

For later reference we note the following . From [16] we have the identity

$$n!d^n x = \epsilon_{\mu_1 \ldots \mu_n} dx^{\mu_1} \wedge \ldots \wedge dx^{\mu_n}.$$

Hence

$$n!\epsilon^{\rho_1 \ldots \rho_n} d^n x = \epsilon^{\rho_1 \ldots \rho_n} \epsilon_{\mu_1 \ldots \mu_n} dx^{\mu_1} \wedge \ldots \wedge dx^{\mu_n} = (n!)^2 dx^{\rho_1} \wedge \ldots \wedge dx^{\rho_n}$$

where we used the Levi-Civita property that

$$\epsilon_{i_1 \ldots i_n} \epsilon_{j_1 \ldots j_n} = \det \begin{vmatrix} \delta_{i_1,j_1} & \ldots & \delta_{i_1,j_n} \\ \vdots & \ddots & \vdots \\ \delta_{i_n,j_1} & \ldots & \delta_{i_n,j_n} \end{vmatrix}.$$

We can then write

$$\phi^* \omega = (\phi^* \omega)_{\mu_1 \ldots \mu_n} dx^{\mu_1} \wedge \ldots \wedge dx^{\mu_n} = \frac{1}{n!} \epsilon^{\mu_1 \ldots \mu_n} (\phi^* \omega)_{\mu_1 \ldots \mu_n} d^n x.$$

Using the antisymmetry of the n-form and determinant properties we have

$$\omega_{i_1 \ldots i_n} \det \left(\frac{\partial \phi^{i_1}}{\partial x^{\mu_1}} \ldots \frac{\partial \phi^{i_n}}{\partial x^{\mu_n}} \right) = n! \omega_{i_1 \ldots i_n} \frac{\partial \phi^{i_1}}{\partial x^{\mu_1}} \ldots \frac{\partial \phi^{i_n}}{\partial x^{\mu_n}}.$$

So, we can finally write the pullback n-form as desired:

$$\phi^* \omega = \epsilon^{\mu_1 \ldots \mu_n} \frac{\partial \phi^{i_1}}{\partial x^{\mu_1}} \ldots \frac{\partial \phi^{i_n}}{\partial x^{\mu_n}} \omega_{i_1 \ldots i_n} d^n x.$$

As a final comment regarding pullbacks, we should mention the nice property that it commutes with exterior derivatives. Consider first the simple case of a function $f : \mathrm{N} \to \mathbb{R}$ that can be pulled back to a function on M by the map $\phi : \mathrm{M} \to \mathrm{N}$. Since $df = (\partial_i f) dy^i$ we have

$$df(v) = v(f) \tag{1.19}$$

for a vector $v = v^i \partial_i$ on N. Using the defining properties of pullbacks and push-forwards we can see that

$$\begin{aligned} \phi^* df(v) &= df(\phi_* v) \\ &\overset{(1.19)}{=} (\phi_* v)(f) \\ &= v(\phi^* f) \\ &\overset{(1.19)}{=} d(\phi^* f)(v). \end{aligned}$$

Hence $\phi^* df = d\phi^* f$. This can now be used to show that this results also holds for more general tensors. Consider $w = w(x) dy^{i_1} \wedge ...$. The pullback acting on the exterior derivative of this form yields

$$\phi^* dw = \phi^* dw \wedge \phi^*(dy^{\mu_1} \wedge ...) = d\phi^* w \wedge \phi^*(dy^{i_1} \wedge ...).$$

Since an exterior derivative applied to $(dy^{i_1} \wedge ...)$ vanishes, we can write this as

$$\phi^* dw = d[\phi^* w \wedge \phi^*(dy^{i_1} \wedge ...)] = d\phi^* w.$$

This will allow us to do some nice manipulations in later discussions. For example, in the context of Stokes theorem, we can now write

$$\int_{\mathrm{M}} \phi^* d\omega = \int_{\mathrm{M}} d\phi^* \omega = \int_{\partial \mathrm{M}} \phi^* \omega.$$

1.6 Geometric structures

So far the manifolds we have discussed have had only two interesting structures, namely smooth structure and the possible structure of fiber bundles. In this section we will discuss several geometrical structures that make the manifolds more interesting and more suited for our applications.

The smooth structure has so far been very generous, in that it has provided us with a myriad of geometric objects to study. From the smooth structure we realized that we could construct the tangent bundle, and from it the tensor bundles trough defining the dual bundle T*M. These bundles, and in particular their local features, has been the main focus so far. We here discuss how to add extra structure to a manifold.

When we discussed fiber bundles we noted that the structure group has to respect the structure of the fibers. In other words, the transition functions that glue the local trivializations together have to glue them together in a manner that is consistent with the wanted structure on the fibers. If the fibers are vector spaces for example, the transition functions must be linear maps $\mathrm{GL}_n(\mathbb{R})$ that preserve the vector structure.

This goes the other way as well, which is the origin of G-structures. We consider the tangent bundle $\mathrm{TM} \to \mathrm{M}$. By demanding that the structure group is a subgroup $\mathrm{G} \subset \mathrm{GL}_n(\mathbb{R})$, i.e.

$$\psi_{\beta\alpha} : \mathrm{U}_\alpha \cap \mathrm{U}_\beta \to \mathrm{G}$$

the fibers necessarily gains the structure for which G is the symmetry group. This is called a reduction of the structure group [40]. Of course, we need to pick a realization of the group G, and may thus think of a G-structure as a picking a group, acting as a sort of symmetry group, and a representation of it. The fibers of our fiber bundle will then be corresponding representation spaces.

For example, we know that if we on a finite dimensional vector space V add a inner product, we get a inner product space $\{V, (\cdot,\cdot)\}$. The maps preserving this structure is now not only the linear maps in GL(V) but rather the orthogonal transformations $O(\dim(V)) \subset GL(V)$. On the level of vector spaces we could then call the addition of a inner product a O(n)-structure on V. Generalizing this idea to the fibers over a manifold we get the notion of a O(n)-structure on M.

While this presents a nice and unified way of looking at geometric structures, it may not be practical, as we don't want to pay attention to local trivializations and projection maps all the time. The most convenient, and most used, way to give a manifold additional structure is by demanding the existence of certain tensor fields over M, which locally reproduces the wanted structure on the fibers. We take this approach, and comment on the relation to G-structure as we go along. It may also be worth mentioning that this discussion on G-structures will be fruitful when we discuss spin structures later, which are somewhat related to the idea of reducing the structure group. In fact, it is a sort of expansion of the structure group.

1.6.1 Riemannian structure

A Riemannian structure is a O(n)-structure on the tangent bundle that locally introduces the notion of orthogonal frames. First, consider a tensor field $g \in \Gamma(T^*M \otimes_{sym} T^*M)$ locally written $g = g_\mu dx^\mu \otimes dx^\nu$ with $g_{\mu\nu} = g_{\nu\mu}$. If this tensor is positive definite, in other words $g(v,v) \geq 0$ and only 0 if $v = 0$, for evert point on the manifold the tensor g is called a Riemannian metric. The pair (M, g) is then called a Riemannian manifold [60]. If in stead $g(u,v) = 0\, \forall u \in T_pM$ implies that $v = 0$, g is called a pseudo-Riemannian metric and the manifold is said to be pseudo-Riemannian.

Just as we did with the smooth manifolds we can consider a category where objects are n-dimensional Riemannian manifolds $\Sigma_{1,2}$ and the morphisms $(n+1)$-dimensional Riemannian cobordisms M [11]. Strictly speaking, once a cobordism M is chosen we should think of the corresponding isometry class as defining the morphism in the category. If $f : M \to N$ is a diffeomorphism and $g_M = f^* g_N$ we call f a isometry. The Riemannian manifold M and N are then isometric and lie is the same isometry class. This class consists of manifolds with the same notion of distance in some sense.

We recall that the inner product was introduced as a map $\mathrm{T}_p^*\mathrm{M} \times \mathrm{T}_p\mathrm{M} \rightarrow \mathbb{R}$. Given a Riemannian metric, which is a map $\mathrm{T}_p\mathrm{M} \otimes \mathrm{T}_p\mathrm{M} \rightarrow \mathbb{R}$ we may think of $g(u,\cdot)$, with one open slot, as a map from the tangent space at p to the reals. In other words, we should think of $g_p(u,\cdot)$ as a one-form. Thus, given a vector $v \in \mathrm{T}_p\mathrm{M}$ and this one-form we can clearly define a map $\mathrm{T}_p^*\mathrm{M} \times \mathrm{T}_p\mathrm{M} \rightarrow \mathbb{R}$ by $(g_p(u,\cdot), v) \rightarrow g_p(u,v) \in \mathbb{R}$. We also note that the vector u determines the one-form $g_p(u,\cdot)$ and inversely a one-form ω determines a vector by the identification $\omega \sim g_p(u_\omega,\cdot)$. Thus there is a one-to-one correspondence between the vectors and one-forms [60]. One says that the Riemannian metric has given us a isomorphism between $\mathrm{T}_p\mathrm{M}$ and $\mathrm{T}_p^*\mathrm{M}$. This correspondence is sometimes called a musical isomorphism, as the map between the dual spaces is often denoted $\sharp$.

Explicitly this isomorphism does the following. Let $v = v^\mu \partial_\mu$ be a vector in $\mathrm{T}_p\mathrm{M}$. The metric tensor $g_{\mu\nu} dx^\mu \otimes dx^\nu$ acts on a single vector as

$$g(v,\cdot) = g_{\mu\nu}(dx^\mu \otimes dx^\nu)(v^\rho \partial_\rho, \cdot) \tag{1.20}$$

$$= g_{\mu\nu} v^\rho \delta^\mu_\rho dx^\nu \tag{1.21}$$

$$= (g_{\mu\nu} v^\mu) dx^\nu. \tag{1.22}$$

The resulting 1-form $\omega_\nu = g_{\mu\nu} v^\mu$ is the unique 1-form corresponding to the vector v. Because of the one-to-one correspondence we often use the notation $(g_{\mu\nu} v^\mu) \equiv v_\nu$. We say that the metric lowered the vector index.

Now consider the case where the point p lies in a intersection of two open subsets. Then, the inner product $\langle a, b\rangle = g(a,b) = a_\mu b^\mu$ transforms as

$$\langle a, b\rangle \rightarrow a_\rho b^\sigma (\psi^\mu_\rho)^{-1} (\psi^\mu_\sigma). \tag{1.23}$$

With transition function $\psi^\nu_\mu = \partial y^\nu / \partial x^\mu$. Clearly $(\psi^\mu_\rho)^{-1}(\psi^\mu_\sigma) = (\psi^\mu_\rho)^{\mathrm{T}}(\psi^\mu_\sigma) = \delta^\rho_\sigma$, so the transition functions can be considered elements of $\mathrm{O}(n)$ with $n = \dim(\mathrm{M})$.

We also promised that we would discuss integrals again when we had introduced Riemannian structure. Recall that integration was defined as a map from top dimensional forms to the real numbers. Given a Riemannian structure, there is a canonical choice of such a n-form [60]

$$dvol \equiv \sqrt{\det(g_{\mu\nu})} dx^1 \wedge ... \wedge dx^n$$

where $g \in \Gamma(\mathrm{T}^*\mathrm{M} \otimes_{sym} \mathrm{T}^*\mathrm{M})$ is a metric tensor on M and x^μ are local coordinates. Note that under a coordinate transformation $x \rightarrow y$ we have [60]

$$dx^1 \wedge ... \wedge dx^n \rightarrow dy^1 \wedge ... \wedge dy^n = \det(\frac{\partial y^\mu}{\partial x^\nu}) dx^1 \wedge ... \wedge dx^n$$

while the prefactor transforms as

$$\sqrt{\det(g_{\mu\nu})} \to \sqrt{\det(g_{\mu\nu}\frac{\partial x^{\mu}}{\partial y^{\rho}}\frac{\partial x^{\nu}}{\partial y^{\sigma}})} = \sqrt{\det(g_{\mu\nu})} \cdot \left|\det(\frac{\partial x^{\mu}}{\partial y^{\nu}})\right|.$$

Hence the volume form $dvol$ in total will transform as

$$dvol \to \pm dvol$$

where the negative sign appears if we have a negative Jacobi determinant. In order to have a well defined volume form, we must add the additional requirement that all Jacobi determinants are positive [40]. Such a manifold is called orientable. Given a O(n)-structure (in the form of a Riemannian metric) the restriction to only positive determinant transitions means that the structure group is further reduced to SO(n). For short, we will from now on write $\sqrt{g} = \sqrt{|\det(g_{\mu\nu})|}$.

This may be best illustrated by an example. Consider the Mobius strip as a surface created by a process of twisting and gluing a long rectangular strip. In the twisting process we change coordinates (with the $\mathbb{Z}_2$-valued transition maps discussed in the section on fiber bundles), i.e. we let $y \to -y$. In this case the Jacobian will simply be $\mathrm{J} = -1$ which is *not* strictly positive. Thus the Mobius strip is a non-orientable surface. Whenever we do integrations we implicitly will assume that we are working on an orientable manifold.

Finally, for practical calculations we need the following. Assume that we are given such a orientable manifold M. Integration of a function $f : \mathrm{M} \to \mathbb{R}$ over a subset $\mathrm{U} \subset \mathrm{M}$ may then be defined [60] as

$$\int_{\mathcal{U}} f\ dvol = \int_{\phi(\mathcal{U})} f(x) dvol(x) \tag{1.24}$$

where the numerical value is calculated in a chart.

A Riemannian structure also allows a definition of the Hodge star operation. Given a r-form ω we can construct a so called Hodge dual form by [60]

$$*\omega = \frac{\sqrt{g}}{(n-r)!}\omega^{\mu_1 \ldots \mu_r}\epsilon_{\mu_1 \ldots \mu_r \mu_{r+1} \ldots \mu_n} dx^{\mu_1} \wedge \ldots \wedge dx^{\mu_n}.$$

I.e. the Hodge star is a map $* : \Omega^r(\mathrm{M}) \to \Omega^{m-r}(\mathrm{M})$. Given two r-forms we have the identity [60]

$$\omega \wedge *\eta = r!\omega_{\mu_1 \ldots \mu_r}\eta^{\mu_1 \ldots \mu_r} \mathrm{dvol}_{\mathrm{M}}. \tag{1.25}$$

An inner product on forms can then be defined by the integration of this top-dimensional form.

1.6.2 Complex structure

A complex structure presents a generalization of "multiplication by i" as familiar from complex analysis. A complex manifold is a manifold where the charts take values in $\mathbb{C}^n$ and where the coordinate transition functions are holomorphic. We will see that we can think of a complex structure as a $\mathrm{GL}_n(\mathbb{C})$-structure.

First we need some terminology. A holomorphic atlas is collection $\{(\mathrm{U}_i, \phi_i)\}$ of holomorphic charts such that the transition functions (coordinate transform functions) $\psi_{ij} = \phi_i \circ \phi_j^{-1} : \mathbb{C}^n \to \mathbb{C}^n$ are holomorphic. Let $\mathscr{A}$ and $\mathscr{A}'$ be holomorphic atlases with coordinate maps ϕ and ϕ'. The two atlases are said to be equivalent if all maps of the form $\phi \circ \phi'^{-1}$ are holomorphic [38]. A complex structure is an equivalence class of holomorphic atlases. A complex manifold M of complex dimension n is a real 2n-dimensional differentiable manifold with a complex structure on it [60][37][38]. The rest of this section is devoted to the notion of complex structure; how to endow a manifold with it, and the maps that preserve it.

We note that by definition all complex manifolds are real differentiable manifolds, while the opposite is clearly not always true, as complex manifolds have a somewhat stricter definition. A interesting question is thus when given an even dimensional real manifold, how does one endow it with a complex structure. Somewhat more precisely, we want to know what are the necessary and sufficient conditions for transition functions to be holomorphic.

At each point on a real 2n-dimensional manifold we define a (linear) map $\mathrm{J}_p : \mathrm{T}_p\mathrm{M} \to \mathrm{T}_p\mathrm{M}$ as a tangent space endomorphism such that $\mathrm{J}_p^2 = -\mathbb{I}_{\mathrm{T}_p\mathrm{M}}$. In terms on bundles, J is a tangent bundle endomorphism that fiberwise square to the negative identity. This is a tensor field of type (1,1) called an almost complex structure [60][37]. In a local trivialization of the bundle $\mathrm{TM} \otimes \mathrm{T}^*\mathrm{M}$ this tensor may be written

$$\mathrm{J}_p = \mathrm{J}_\mu{}^\nu dx^\mu \otimes \partial_\nu \tag{1.26}$$

at some point p. A real even dimensional smooth manifold M equipped with a almost complex structure is called an almost complex manifold, often denoted by the tuple (M, J). For some vector field $\mathrm{X} = \mathrm{X}^\mu \partial_\mu$ this tensor acts (locally) by

$$\mathrm{J}_p(\mathrm{X}) = \mathrm{J}_\mu{}^\nu \mathrm{X}^\mu \partial_\nu \tag{1.27}$$

$$\therefore \mathrm{J}_p^2(\mathrm{X}) = \mathrm{J}_p(\mathrm{J}_\mu{}^\nu \mathrm{X}^\mu \partial_\nu) = \mathrm{X}^\mu \mathrm{J}_\mu{}^\nu \mathrm{J}_\nu{}^\beta \partial_\beta. \tag{1.28}$$

Locally at a point p we must have that $\mathrm{J}_\mu{}^\nu \mathrm{J}_\nu{}^\beta = -\delta_\mu^\beta$ to fit the definition. To see the relation between almost complex manifolds and complex manifolds, we want to show explicitly that all complex manifolds are almost complex manifold.

Let M be a complex manifold, i.e. a real even dimensional manifold with holomorphic coordinate transition functions. It is convenient to complexity the tangent spaces, or equivalently the tangent bundle. For vector fields X, Y on M, viewed as a real manifold, we can define what we call the complexified fields in the obvious way

$$Z = \frac{1}{2}(X + iY), \tag{1.29}$$

$$\overline{Z} = \frac{1}{2}(X - iY). \tag{1.30}$$

We say that these fields are elements of the complexified tangent space $T_pM^{\mathbb{C}}$ where we now allow complex components. Since we have that $J_p^2 = -\mathbb{I}_{T_pM}$ its eigenvalues must clearly be $\lambda = \pm i$, and we may split the tangent space into two disjoint eigenspaces [60][37]

$$T_pM^{\mathbb{C}} = T_pM^+ \oplus T_pM^-$$

of vector fields with with eigenvalues $+i$ and $-i$ respectively. We call these fields holomorphic and anti-holomorphic vector fields respectively. We further claim that any vector field V can be split into its holomorphic and anti-holomorphic part as follows

$$Z = \frac{1}{2}(V - iJ(V)) \;\; ; \;\; \overline{Z} = \frac{1}{2}(V + iJ(V)).$$

This we can verify by letting the tensor J act on these. We then find

$$J(Z) = \frac{1}{2}(J(V) - iJ^2(V)) = \frac{1}{2}(J(V) + iV) \tag{1.31}$$

$$= i\frac{1}{2}(V - iJ(V)) = iZ \tag{1.32}$$

and similarly for $\overline{Z}$. With this splitting of the tangent spaces (and thereby tangent bundle) we may write the almost complex structure locally as

$$J = i\frac{\partial}{\partial z^\mu} \otimes dz^\mu - i\frac{\partial}{\partial \overline{z}^\mu} \otimes d\overline{z}^\mu \tag{1.33}$$

in what we will refer to as its canonical form. Clearly this tensor acts on holomorphic fields by multiplication by i and on anti-holomorphic fields by multiplication by $-i$ as wanted. Note that by writing out the complex coordinates in terms of the real ones

$$\frac{\partial}{\partial z^\mu} = \frac{1}{2}\left(\frac{\partial}{\partial x^\mu} - i\frac{\partial}{\partial y^\mu}\right) \;\; ; \;\; \frac{\partial}{\partial \overline{z}^\mu} = \frac{1}{2}\left(\frac{\partial}{\partial x^\mu} + i\frac{\partial}{\partial y^\mu}\right)$$

$$dz^\mu = dx^\mu + idy^\mu \;\; ; \;\; d\overline{z}^\mu = dx^\mu - dy^\mu$$

the the almost complex structure is of the form

$$\mathrm{J} = \frac{\partial}{\partial y^\mu} \otimes dx^\mu - \frac{\partial}{\partial x^\mu} \otimes dy^\mu \tag{1.34}$$

so that $\mathrm{J}(\partial/\partial x^\mu) = \partial/\partial y^\mu$ and $\mathrm{J}(\partial/\partial y^\mu) = -\partial/\partial x^\mu$, i.e. it acts as a $\pi/2$ rotation in the tangent spaces. We may interpret J as a guidance as to how we are to relate the $2n$ basis vectors on the underlying real manifold. This interpretation is particularly clear in the case of real dimension two.

Assume that we are working on a non-empty intersection of patches on M. Under a change of coordinates $z \rightarrow w(z,\bar{z})$ on these patches we have

$$\frac{\partial}{\partial z^\mu} \otimes dz^\mu = \left(\frac{\partial \omega^\nu}{\partial z^\mu} \frac{\partial}{\partial \omega^\nu} + \frac{\partial \overline{\omega}^\nu}{\partial z^\mu} \frac{\partial}{\partial \overline{\omega}^\nu} \right) \otimes \left(\frac{\partial z^\mu}{\partial \omega^\rho} d\omega^\rho + \frac{\partial z^\mu}{\partial \overline{\omega}^\rho} d\overline{\omega}^\rho \right) \tag{1.35}$$

$$= \frac{\partial z^\mu}{\partial w^\rho} \frac{\partial w^\nu}{\partial z^\mu} \frac{\partial}{\partial w^\nu} \otimes dw^\rho = \frac{\partial}{\partial w^\nu} \otimes dw^\nu \tag{1.36}$$

where we used the Cauchy Riemann conditions to eliminate the dependence on anti-holomorphic coordinates. Similar transformations apply to the anti-holomorphic part of J. In conclusion, the tensor J keeps its canonical form as we move from patch to patch on our manifold, given that we can do so holomorphically. Thus any complex manifold is a almost complex manifold, as one would expect from the names.

The converse is of course what we are interested about, and in general it is not true that almost complex manifolds are complex manifolds. The main result regarding this question is due to Newlander and Nirenberg, whose famous theorem gives us the sufficient condition for an almost complex manifold to be complex. It is a necessary condition that the manifold has an almost complex structure. It turns out we can find holomorphic coordinates if the Nijenhuis tensor vanishes. For X and Y vector fields, we have

$$\mathrm{N}(\mathrm{X},\mathrm{Y}) \equiv [\,\mathrm{X},\mathrm{Y}] + \mathrm{J}[\mathrm{JX},\mathrm{Y}] + \mathrm{J}[\mathrm{X},\mathrm{JY}] - [\mathrm{JX},\mathrm{JY}] \tag{1.37}$$

We then have the two important theorems.

THEOREM: *Let* M *be a real even dimensional manifold endowed with a almost complex structure* J. *Then this structure is said to be integrable if and only if the Nijenhuis tensor vanishes, i.e.* $\mathrm{N}(\mathrm{X},\mathrm{Y}) = 0$ *for any two vector fields* X *and* Y *on* M. For a proof see theorem 8.12 in [60].

NEWLANDER-NIRENBERG THEOREM: *Let* (M, J) *be a almost complex manifold. If the almost complex structure is integrable,* M *is a complex manifold. We then sometimes call* J *a complex structure.*

For a proof see original article [62].

In summary, the vanishing of the Nijenhuis tensor field given an almost complex structure J on an real 2n-manifold means that we may everywhere construct a coordinate system of holomorphic coordinates so that locally J takes its canonical form. In practice this means one have to be imaginative and construct some tensor that plays the role on an almost complex structure, and then simply check whether or not the Nijenhuis tensor vanishes.

In terms of the structure groups, the (almost) complex structure can be seen as the reduction to $\mathrm{GL}_n(\mathbb{C}) \subset \mathrm{GL}_{2n}(\mathbb{R})$ by the usual identification $\mathbb{R}^2 = \mathbb{C}$.

Some remarks

(i) Similarly to the vectors and one-forms more general tensors may be expressed in terms of holomorphic and antiholomorphic parts. For example [37] the space of n-forms $\Omega^n(\mathrm{M})$ may be decomposed

$$\Omega^n(\mathrm{M}) = \bigoplus_{p+q=n} \Omega^{p,q}(\mathrm{M})$$

where $\Omega^{p,q}(\mathrm{M})$ denotes the set of $(n = p + q)$-forms, expressed in basis

$$dz^{\mu_1}...dz^{\mu_p}d\overline{z}^{\nu_1}...d\overline{z}^{\nu_q}.$$

We will call these types of n-forms (p,q)-forms, not to be mistaken for (p,q)-tensors.

(ii) One can in the literature often find a somewhat confusing notation. Trough the identification $\mathrm{V} \to \frac{1}{2}(\mathrm{V} - i\mathrm{J}(\mathrm{V}))$ we can identify TM with the bundle of holomorphic fields, and trough $\omega \to \omega_{hol} = \frac{1}{2}(\omega - i\mathrm{J}(\omega))$ associate $\mathrm{T}^*\mathrm{M}$ with the holomorphic cotangent bundle. One therefore sometimes writes the complexified tangent bundle $\mathrm{T}^{\mathbb{C}}\mathrm{M} = \mathrm{TM} \oplus \overline{\mathrm{TM}}$, where the overline represents antiholomorphic fields.

(iii) Note that by giving a manifold any holomorphic atlas, a unique equivalence class is defined and so is a complex structure. This is how one in practice gives manifolds complex structures, similarly to the case of smooth structure where transition functions were smooth.

Assume next that we have two complex manifolds $\mathrm{M}_1, \mathrm{M}_2$. A continuous map

$$f : \mathrm{M}_1 \to \mathrm{M}_2$$

is said to be holomorphic if its coordinate representation $\phi' \circ f \circ \phi^{-1}$ is. Here we have used a (holomorphic) chart (U, ϕ) for M_1 and (U', ϕ') for M_2. A holomorphic map between complex manifolds with a holomorphic inverse is called a biholomorphism and the manifolds are said to be biholomorphic [60]. Just as diffeomorphisms gave us a notion of equivalence of differentiable manifolds, biholomorphic complex manifolds are considered the same. We say that they are biholomorphically equivalent. We will explore such equivalences of complex manifolds in more detail when we discuss the Riemann surfaces and the complex torus.

1.6.3 Conformal structure

Conformal structure can be seen as a relaxation of a Riemannian structure. Imagine that we for some reason only are interested in discussing angles locally on our manifold, not lengths. Just as in linear algebra, the angle between two vectors is given by

$$\cos\theta = \frac{g(a,b)}{\sqrt{g(a,a)}\sqrt{g(b,b)}}.$$

This local notion of angles remain unchanged if we rescale the metric

$$g_{\mu\nu} \to e^{\omega} g_{\mu\nu}$$

by a positive factor $e^{\omega(x)}$. The set of such metric transformations is called the Weyl transformations Weyl(M). The conformal class of metrics $[g_{\mu\nu}]$ constitute a conformal structure on the manifold. The interplay between Riemannian, complex and conformal structure will be of key importance when we discuss the theory of Riemann surfaces.

1.7 Connections and curvature

When we constructed the tangent spaces, we noted that there is no natural way to compare vectors belonging to different tangent spaces. This is in general true [39] for all vector bundles. This means trouble if we want to naively define a notion of differentiation of a vector, or more generally differentiation of sections of bundles. This section is inspired from [6],[60] and [39].

Consider a vector bundle $\pi : \mathcal{E} \to \mathrm{M}$ with fiber spaces V with a structure group G. We denote by $\mathfrak{g}$ the Lie algebra of this structure group. We wish to construct a derivative on the sections $\Gamma(\mathcal{E})$. We will mainly work in a local trivialization over $\mathrm{U} \subset \mathrm{M}$, where we chose a basis of local sections $\{e_i\}$. For future reference we consider the following product bundle

$$\mathfrak{g}\otimes \mathrm{T}^*\mathrm{M}\to .$$

We will work in a local trivialization over $\mathrm{U}\subset\mathrm{M}$. We consider the Lie algebra as a vector space, and expand a section locally as

$$\mathrm{A}=\mathrm{A}^a_\mu \mathrm{X}_a\otimes dx^\mu$$

where X_a are the generators of the structure group in a given representation. We write $\mathrm{A}=\mathrm{A}_\mu dx^\mu$ and think of A_μ as a collection of Lie algebra valued components. To see how this acts on vector fields v, as a section of $\mathrm{T}^*\mathrm{M}$, it is sufficient to see how it acts on the basis ∂_μ, namely

$$\mathrm{A}(\partial_\nu)=\mathrm{A}^a_\mu\delta^\mu_\nu\mathrm{X}_a=\mathrm{A}^a_\nu\mathrm{X}_a=\mathrm{A}_\nu\in\mathfrak{g}.$$

As this is simply an element of the Lie algebra, we can write it in terms of its matrix components $(\mathrm{A}_\mu)^i_j e_i\otimes e^j$, where we think of e^j as a local basis for the dual bundle. It can act on the local sections by $\mathrm{A}(\partial_\mu)s=\mathrm{A}(\partial_\mu)^i_j s^k e_i\otimes e^j(e_k)=(\mathrm{A}_\mu)^i_j s^j e_i$. This little analysis will help us realize what kind of geometric object the derivative is.

We now define the derivative relevant for sections of vector bundles. The covariant derivative is a map of sections

$$\mathrm{D}_\nu:\Gamma(\mathcal{E})\to\Gamma(\mathcal{E})$$

where $\nu\in\Gamma(\mathrm{TM})$ is a vector field on M giving the direction of differentiation [60][6]. Let c be some scalar, and f a function of M. Then, for sections s we want the derivative to have the following properties

$$\mathrm{D}_\nu(cs+s')=c\mathrm{D}(s)+\mathrm{D}(s')$$

$$\mathrm{D}_\nu(fs)=(\nu f)s+f\mathrm{D}_\nu(s)$$

$$\mathrm{D}_{\nu+w}(s)=\mathrm{D}_\nu(s)+\mathrm{D}_w(s)$$

$$\mathrm{D}_{f\nu}(s)=f\mathrm{D}_\nu(s).$$

Again we work locally on a subset U of M, and try to find coordinate representations of the derivative. Note that since $\mathrm{D}_{\partial_\mu}\equiv\mathrm{D}_\mu$ maps vectors to vectors, we can write

$$\mathrm{D}_\mu e_i=(\mathrm{A}_\mu)^j_i e_j.$$

Here the matrix $\mathrm{A}_\mu=(\mathrm{A}_\mu)^i_j e_i\otimes e^j$ is a element of $\mathfrak{g}$. With the above rules, we can write the covariant derivative of any local section [6]

$$\mathrm{D}_\nu s=\mathrm{D}_{\nu^\mu\partial_\mu}(s^i e_i)=\nu^\mu(\partial_\mu s^i+(\mathrm{A}_\mu)^i_j s^j)e_i.$$

In particular we have $D_\mu s = (D_\mu s)^i e_i = (\partial_\mu s^i + (A_\mu)^i_j s^j)e_i$. This last term is exactly the expression one gets from the sections of $\mathfrak{g} \otimes T^*M$ as we saw above. Hence we think of A, called the connection, as a Lie algebra valued 1-form.

One often uses the shorthand notation $D = d + A$ where d is the exterior derivative. To see that this makes sense, note the action of a $\Gamma(\mathcal{E})$ section

$$Ds = d(s^a)e_a + A^a_b s^b e_a = (\partial_\mu s^a + (A_\mu)^a_b s^b)dx^\mu \otimes e_a \equiv D_\mu s^a dx^\mu \otimes e_a.$$

We known that the exterior derivative satisfies d^2, but this is not true for D:

$$D^2 s = (d + A)(ds + As) = (dA)s - A(ds) + A(ds) + (A \wedge A)s$$

This expression is called the curvature 2-form of the connection $F = dA+A\wedge A$. By writing the 1-form in coordinates, and remembering that it is Lie algebra valued, one can easily see that

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu].$$

In matrix components when acting on a section this reads

$$(F_{\mu\nu})^a_b s^b = (\partial_\mu (A_\nu)^a_b - \partial_\nu (A_\mu)^a_b + [A_\mu, A_\nu]^a_b)s^b$$

As most of the manifolds we will study later will be Riemannian, we now consider the special case of the tangent bundle TM over a Riemannian manifold M. We here denote the connection by

$$\nabla : \Gamma(TM) \to \Omega^1(M) \otimes \Gamma(TM).$$

We define the numbers, called Christoffel symbols, $\Gamma^\sigma_{\mu\nu}$ by the equation $\Gamma^\sigma_{\mu\nu}\partial_\sigma = \nabla_\mu \partial_\nu$ just as for the general discussion above. Then, the connection has the action

$$\nabla_\mu v = (\partial_\mu v^\rho + v^\sigma \Gamma^\rho_{\mu\sigma})\partial_\rho$$

where we have the components $\nabla_\mu v^\nu = \partial_\mu v^\nu + \Gamma^\nu_{\mu\rho} v^\rho$. This connection may be generalized to one-forms and tensors as well [60], where on a one-form

$$\nabla_\mu \omega_\nu = \partial_\mu \omega_\nu - \Gamma^\rho_{\mu\nu}\omega_\rho$$

and on a (p,q)-tensor

$$\nabla_\mu T^{\rho_1 \dots \rho_p}_{\sigma_1 \dots \sigma_q} = \partial_\mu T^{\rho_1 \dots \rho_p}_{\sigma_1 \dots \sigma_q} + \Gamma^{\rho_1}_{\mu\nu} T^{\nu \dots \rho_p}_{\sigma_1 \dots \sigma_q} + \dots + \Gamma^{\rho_1}_{\mu\nu} T^{\rho_1 \dots \nu}_{\sigma_1 \dots \sigma_q}$$
$$- \Gamma^\nu_{\mu\sigma_1} T^{\rho_1 \dots \rho_p}_{\nu \dots \sigma_q} - \dots$$

For example, we can calculate the derivative of the metric tensor

$$\nabla_\mu g_{\rho\sigma} = \partial_\mu g_{\rho\sigma} - \Gamma^\nu_{\mu\rho} g_{\nu\sigma} - \Gamma^\nu_{\mu\sigma} g_{\rho\nu}.$$

In the case where the metric is covariantly constant $\nabla g = 0$ and the $\Gamma^\rho_{\mu\nu}$ are symmetric in the two lower indices we thus have

$$\partial_\mu g_{\rho\sigma} = \Gamma^\nu_{\mu\rho} g_{\nu\sigma} + \Gamma^\nu_{\mu\sigma} g_{\rho\nu}.$$

By adding a $(\mu \to \rho)$ term and subtracting a $(\mu \to \sigma)$ term we find an expression

$$\Gamma^\rho_{\mu\nu} = \frac{1}{2} g^{\rho\sigma}(\partial_\mu g_{\nu\sigma} + \partial_\nu g_{\mu\sigma} - \partial_\sigma g_{\mu\nu}).$$

The connection where these are symmetric in the lower two indices and the metric tensor is covariantly constant is called the Levi-Civita connection. From the fundamental theorem of Riemannian geometry [60] this connection is unique.

In the case of Riemannian manifolds the curvature reads

$$R^\rho_{\sigma\mu\nu} = \partial_\mu \Gamma^\rho_{\nu\sigma} - \partial_\nu \Gamma^\rho_{\mu\sigma} + \Gamma^\rho_{\mu\lambda}\Gamma^\lambda_{\nu\sigma} - \Gamma^\rho_{\nu\lambda}\Gamma^\lambda_{\mu\sigma}$$

and is called the Riemann curvature tensor. The contraction $R^\mu_{\sigma\mu\nu} = R_{\sigma\nu}$ is called the Ricci curvature tensor and $R = R^\mu_\mu$ the Ricci scalar curvature. For the case of two dimensions, the Riemann curvature tensor takes the form

$$R_{\mu\nu\rho\sigma} = K(x)(g_{\mu\rho} g_{\lambda\nu} - g_{\mu\sigma} g_{\nu\rho})$$

with Ricci tensor $R_{\mu\nu} = K g_{\mu\nu}$. By performing a final contraction we see that $K = R/2$ [60]. These results will be important when we discuss Riemann surfaces shortly.

For the sake of completeness we discuss geodesics. Consider a curve $\gamma : \mathbb{R} \to M$ with local coordinates $x^\mu(\lambda)$ the tangent vector

$$T_\gamma = \frac{\partial x^\mu(\lambda)}{\partial \lambda} \partial_\mu.$$

A vector v along this curve $v(\lambda) = v(\gamma(\lambda)) = v^i(\lambda) e_i$ is said to be covariantly constant along γ if $D_{T_\gamma} v = 0$. This can be seen as a set of differential equations determining the vector v, whose solution $v(\lambda)$ is called a parallel transport of $v(0)$ [39][37]. The curves that satisfy $\nabla_{T_\gamma} T_\gamma = 0$ are called geodesics, and can be shown [37] to satisfy the equation

$$\frac{d^2\gamma^\mu}{d\lambda^2} + \Gamma^\mu_{\nu\rho} \frac{d\gamma^\nu}{d\lambda} \frac{d\gamma^\rho}{d\lambda}.$$

These curves are the generalization of straight lines from Euclidian space to a Riemannian manifold.

1.8 Chern cohomology

Having discussed connections and curvature we can discuss Chern cohomology, which gives us a discrete topological invariant of vector bundles. We will use Chern cohomology and the associated invariants when we classify simple topological insulators in even dimensions in later chapters.

Let M be a closed manifold so that the integral of a form depends only on its cohomology class. Let also $\mathcal{E}$ be a rank k vector bundle over M with structure group G, which we usually will take to be $\mathrm{U}(n)$ for a complex vector bundle. The connection and curvature is as usual denoted A and F. A bundle is said to have Chern classes $c_i(\mathcal{E})$, which are rank $2i$ differential forms[2] [60]. One often defines the total Chern class by the sum

$$c(\mathcal{E}) = 1 + c_1(\mathcal{E}) + c_2(\mathcal{E}) + \ldots + c_k(\mathcal{E}).$$

In terms of the curvature of the connection, the Chern class can be written [6]

$$c_j = \frac{(i/2\pi)^j}{j!}\mathrm{tr}\mathrm{F}^j = \frac{(i/2\pi)^j}{j!}\mathrm{tr}(\mathrm{F} \wedge \ldots \wedge \mathrm{F})$$

We are being somewhat sloppy with the terminology here, as c_j is a (closed) form whose cohomology class should be called the Chern class [6]. However, since any representative of a class defines it, we allow ourselves to call c_j the j'th Chern class. These Chern classes are important in the topological classification of vector bundles. In particular, the integral of a Chern form is an integer [6] which is called the Chern number. Following [58] and [6] we discuss the U(1) case on 2-manifolds in some detail before discussing some general remarks. Over a 2-manifold only the first Chern class exists. The first Chern form is simply

$$c_1 = i\mathrm{F}/2\pi$$

where locally $\mathrm{F} = d\mathrm{A}$. We notice a few things. First of all, locally the curvature is closed since it can locally be written as an exact form. However, globally on M we need a connection for each open subset U_α, and the curvature is only closed on each subset. Recall that on overlaps $\mathrm{U}_\alpha \cap \mathrm{U}_\beta$ the connection transforms as

$$\mathrm{A}_\alpha = \mathrm{A}_\beta + i d\phi_{\beta\alpha}$$

under a U(1) transition function $\psi_{\beta\alpha} = \exp(i\phi_{\beta\alpha})$. We will absorb the factor of i into the connection. Recall also that the transition functions determine the topological non-triviality of the situation. In this sense, two connections should be

[2] Lie algebra valued forms, to be precise.

considered topologically of the same type if they transform identically on overlaps [58]

$$A_\alpha = A_\beta + d\phi_{\beta\alpha}$$
$$\tilde{A}_\alpha = \tilde{A}_\beta + d\phi_{\beta\alpha}.$$

The form $A_\alpha - \tilde{A}_\alpha = A_\beta - \tilde{A}_\beta$ is then clearly independent of the open subset and is globally well defined. The corresponding curvatures then satisfy

$$F_\alpha - \tilde{F}_\alpha = d(A - \tilde{A})$$

and hence is also globally well defined. In addition, the two curvatures are in the same cohomology class as a result of the two gauge fields being of the same topological type. In this sense, the Chern classes are topological.

We should also note how the curvature 2-form changes under a small deformation of the gauge field. We write $A' = A + \delta A$, which implies $\delta F = d\delta A$. In other words, the cohomology class of the curvature is independent of the choice of connection within a certain topological class, and in this sense only depends on the twisting of the bundle. These observations imply that over a closed 2-manifold M, the integral of the Chern form give a topological invariant classifying the bundle. Similar results can be found for the higher Chern classes, see for example [6]. As mentioned these numbers, called Chern numbers, can also be shown to take integer values. In proper normalization the general result is that

$$C_k(\mathcal{E}) = \frac{(i/2\pi)^k}{k!} \int_M \mathrm{tr}(F^k) \in \mathbb{Z}.$$

The Chern class has an important property known as the splitting principle. First, for direct sum vector bundles the total Chern class satisfies $c(\mathcal{E} \oplus \mathcal{F}) = c(\mathcal{E})c(\mathcal{F})$ [61]. The splitting principle states [61] that to prove any relation regarding Chern classes it is sufficient to assume that we are dealing with a direct sum of line bundles

$$\mathcal{E} = \mathcal{L}_1 \oplus \mathcal{L}_2 \oplus .. \oplus \mathcal{L}_k.$$

For line bundles the total Chern class is $c(\mathcal{L}_i) = 1 + c_1(\mathcal{L}_i)$, and for $\mathscr{E}$ we have

$$c(\mathcal{E}) = \prod_{i=1}^{k} [1 + c_1(\mathcal{L}_i)] = 1 + c_1(\mathcal{L}_1) + c_1(\mathcal{L}_2) + ... + c_1(\mathcal{L}_1)c_1(\mathcal{L}_2) + ...$$

Carrying on the product we can rad of the j'ht Chern class. For example

$$c_1(\mathcal{E}) = \sum_{i=1}^{k} c_1(\mathcal{L}_i),$$

$$c_2(\mathcal{E}) = \sum_{i<j} c_1(\mathcal{L}_i)c_1(\mathcal{L}_j).$$

A class that is somewhat better behaved under direct sums is the Chern character [61]. It is defined as

$$ch(\mathcal{E}) = \sum_{i=1}^{k} e^{c_1(\mathcal{L}_i)} = \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{i=1}^{k} c_1^n(\mathcal{L}_i).$$

For a direct sum bundle the Chern character simply adds $ch(\mathcal{E} \oplus \mathcal{F}) = ch(\mathcal{E}) + ch(\mathcal{F})$ [61]. As for the total Chern form the Chern character can be written as a sum $ch(\mathcal{E}) = ch_0(\mathcal{E}) + ch_1(\mathcal{E}) + ... + ch_k(\mathcal{E})$. By the above formula, we can read of the different j'th Chern characters

$$ch_0(\mathcal{E}) = k = \mathrm{rk}\mathcal{E},$$

$$ch_1(\mathcal{E}) = \sum_{i=1}^{k} c_1(\mathcal{L}_i) = c_1(\mathcal{E}),$$

$$ch_2(\mathcal{E}) = \frac{1}{2}[c_1^2(\mathcal{L}_1) + ... + c_1^2(\mathcal{L}_k)] = \frac{1}{2}[c_1^2(\mathcal{E}) - 2c_2(\mathcal{E})].$$

The integral of the j'th Chern character over a closed manifold M is another way to find the Chern numbers. A trivial bundle $\mathrm{M} \times \mathbb{C}^\ell$ admits a flat connection with $\mathrm{F} = 0$. In this case the Chern characters are trivial, and the Chern numbers zero. Hence we can modify a bundle $\mathcal{E}$ by a trivial bundle without changing its Chern numbers. This will be important when we discuss the so called A class topological insulators in even dimensions.

1.9 Spin structures

To define spinors on a manifold, we need the notion of a spin structure. This is similar to the other types of geometric structure we have discussed, in that it can be seen as a modification of the structure group of a bundle over M. Physically we often view M as spacetime, which already has some geometric structures. Formally, a spacetime is a manifold with a (Lorentzian) metric tensor field and a smooth structure. Causality can also be seen as a required structure. The manifold M is covered by light-cones, defined by the null trajectories calculated in each tangent space, and is covered by a vector field T that picks, in a continuous fashion, a notion of future directed cones.

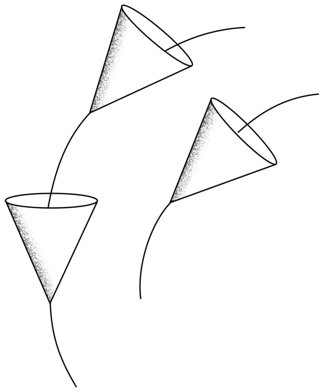


Figure 1.8: Causal structure as smoothly varying light cones.

For spinors we also need the notion of spin structures. Thus, a suitable spacetime is formally the quintuple

$$(\mathrm{M}, \mathcal{A}, g, \mathrm{T}, \#)$$

where $\mathcal{A}$ is a smooth atlas, g a Riemannian metric, and # denotes a spin structure.

In the non-relativistic case, when all motion takes place at small angles inside the light-cone, there is no mixing of space and time. This simplifies the discussion of spinors somewhat. This is the case in most of our discussions, where we simply view spacetime as a product $\Sigma \times \mathrm{I}$ of a space manifold Σ and a time interval with a Euclidian metric signature. It is now Σ that needs a spin structure. We will discuss spinors in flat space from a group theoretic perspective before generalizing to curved spaces.

Spinors are, from a group theoretical point of view, defined as representations of the double covering groups of the special orthogonal groups

$$\mathrm{SO}(n) = \mathrm{Spin}(n)/\mathbb{Z}_2.$$

This can be seen as a consequence of the projective nature of quantum states. In dimensions larger than two the double cover is in fact a universal cover. In low dimensions we have the accidental group isomorphisms

n	Spin(n)	π_1(SO(n))
1	O(1)	$\{e\}$
2	U(1)	$\mathbb{Z}$
3	SU(2)	$\mathbb{Z}_2$
4	SU(2)× SU(2)	$\mathbb{Z}_2$

We should note that the above table of isomorphisms gives a false impression of a pattern. It is not that case that higher dimensional spin groups continue to be a combination of SU(2). In fact, there are no known patters whatsoever.

Most familiar is maybe the case of three dimensions. Here the Lie algebras $\mathfrak{so}(3)$ and $\mathfrak{su}(2)$ coincide, and the irreducible representations are vector spaces V_ℓ of dimension $2\ell+1$, where the *spin* ℓ can be integer or half integer [35]. The fundamental representation is the two dimensional spin-half representation. However, the half integer representations to not correspond to proper representations of the group SO(3), but rather its double cover $\text{Spin}(3) = \text{SU}(2)$. Roughly peaking, the Lie algebra in exponentiated into the covering group. The two types of spin, integer and half-integer, corresponds to the two inequivalent ways to quantize a classical system of many particles, i.e. bosons and fermions. In general, if $V(n)$ is a representation space of $\text{Spin}(n)$, the tensor product space $L^2(\mathbb{R}^n)\otimes V(n)$ is the appropriate Hilbert space of a quantum mechanical particle moving in flat space.

In dimension 2 things are somewhat more strange. Here the universal cover is the real line $\mathbb{R}$, which is not a double cover but a infinite covering. One may then be tempted to relate the unitary representations of the real line with spinors, in a similar fashion to the higher dimensional cases. This would lead to spins characterized by any real number, which is related to anyonic statistics. However, we will here claim that spinors are still associated with double covering groups, which in the case of a circle $\text{SO}(2) = S^1 = U(1)$ is simply a circle "traversed with double speed". By the identification $\mathbb{R}^2 = \mathbb{C}$ the spin transformations are phases

$$\psi \to e^{i\theta/2}\psi$$

e.g. rotation operators that square to normal SO(2) rotations.

To have a well defined notion of spinors on a general manifold, we need yet another geometric structure. As opposed to the earlier G-structures where the we reduced the structure group, spin structure is a geometric structure where one considers not a reduction but a covering of the structure group. We will start with a (oriented) Riemannian manifold, i.e. a manifold where the structure group is the special orthogonal group. To generalize the concept of spinors to curved spaces, we must construct a vector bundle where we lift the SO(n) structure group to the Spin(n) group. This section is by no means meant as a exhaustive discussion on spin geometry, but rather a introduction. For more detailed discussions see for example [60].

This lifting is achieved by the 2:1 homomorphism $\rho : \text{Spin}(n) \to \text{SO}(n)$. Given transition functions

$$\psi_{\alpha\beta} : U_\alpha \cap U_\beta \to \text{SO}(n)$$

on a manifold with Riemannian structure, we wish to lift these to the transitions

$$\tilde{\psi}_{\alpha\beta} : \mathrm{U}_\alpha \cap \mathrm{U}_\beta \to \mathrm{Spin}(n)$$

so that on triple intersections the cocycle condition is satisfied

$$\tilde{\psi}_{\alpha\beta}\tilde{\psi}_{\beta\gamma}\tilde{\psi}_{\gamma\alpha} = 1.$$

A spin structure on M is defined [60] by the existence of transition functions satisfying this cocycle condition. In this case the vector bundle is called a spin bundle $\mathcal{S}$M, and M a spin manifold. Spinors are associated with sections of this bundle. In a local trivialization this gives back the familiar notion of spinors, e.g. where a local section maps $x \to (x, \psi(x))$ and $\psi : \mathrm{M} \to \mathrm{V}$ where V is a vector space with a spin representation.

Note that the inverse group homomorphism acting on the cocycle condition satisfies

$$\rho^{-1}(1) = \rho^{-1}(\psi_{\alpha\beta}\psi_{\beta\gamma}\psi_{\gamma\alpha}) = \pm 1$$

since the homomorphism is 2:1. These sign ambiguities are the possible obstruction to define a spin structure.

As discussed above, in a local trivialization of the spinor bundle, spinors are simply elements of a vector space with a representation $\rho : \mathrm{Spin}(n) \to \mathrm{Aut(V)}$. Let $\psi(x)$ be such a spinor. To define a Dirac operator, we need the notion of covariant derivative of spinors. Just as in the case of vectors where

$$\mathrm{D}_\mu \mathrm{V}^\nu = \partial_\mu \mathrm{V}^\nu + \Gamma^\nu_{\mu\rho}\mathrm{V}^\rho$$

with connection coefficients $\Gamma^\nu_{\mu\rho}$, the covariant derivative of spinors is

$$\mathrm{D}_\mu \psi = \partial_\mu \psi + \omega_\mu^{ab}\Sigma_{ab}\psi$$

where Σ_{ab} are the generators in the spinor representation and ω_μ^{ab} is the connection one-form. See [58] or [60] for more details on this spin connection. This covariant derivative can be used to construct Dirac operators [60]. The gamma matrices γ^μ satisfy the Clifford algebra $\{\gamma^\mu, \gamma^\nu\} = 2\delta^{\mu\nu}$. The Dirac operator is defined as $\mathcal{D} = i\gamma^\mu\partial_\mu = i\partial\!\!\!/$. If we use curved space gamma matrices $\{\gamma^\mu, \gamma^\nu\} = 2g^{\mu\nu}$ the Dirac operator is $i\gamma^\mu \mathrm{D}_\mu = i\mathrm{D}\!\!\!\!/$. We should note that the Dirac operator as it is defined does not act on any kind of spinor, but rather the Dirac spinors. This is not a irreducible representation of the spin group, but rather is a combination of spinors with different chirality. From the above gamma matrices, we define the matrix [60]

$$\gamma^{n+1} = i^{(n/2)}\gamma^1 \cdots \gamma^n,$$

$$\gamma^{n+1^2} = 1.$$

This definition only holds for even space-time dimensions as n must be a even number. The eigenvalues of γ^{n+1} must be ±1 and are called the chirality. The space of sections $\Gamma(\mathcal{S}M)$ can then be divided into disjoint eigenspaces [60]

$$\Gamma(\mathcal{S}M) = \Gamma^+(\mathcal{S}M) \oplus \Gamma^-(\mathcal{S}M).$$

These two classes of spinors are often called Weyl spinors.

1.10 Riemann surfaces

We will here introduce the concept of a Riemann surface based on earlier discussions of geometric structures. For further discussions on Riemann surfaces, see for example [40] or [84].

DEFINITION: *A Riemann surface Σ is a 1 dimensional complex manifold, e.g. a complex curve [84].*

This may be seen as our main definition. However, as we discussed in the section on complex structures it may be rewarding to ask when a real surface can be given a complex structure. The following theorem is thus invaluable.

THEOREM: *Every orientable Riemannian 2-manifold (Σ,g) is a Riemann surface.*

We will show the equivalence to Definition 1. We will follow the line of thought presented in [58] and [60]. We include these proofs in their entirety as they give insight into exactly what makes two dimensions so special.

This first proof is due to Green, Schwarz and Witten [58], sketched in a footnote in their discussion on complex manifolds. We elaborate on this in the hope to gain further insight. We also choose to work not in component form with the Nijenhuis tensor as is done in their discussion. We know from our discussions of tensors and forms that on a (orientable) Riemannian manifold there is an antisymmetric tensor ϵ_{ij} determined by a single parameter. We may chose it in a normalized way such that $\epsilon_{12} = -\epsilon_{21} = 1$. We note that this tensor squares to $-\mathbb{I}$, as is clear if we view it as a 2-by-2 matrix. Together with a Riemannian metric we can construct the tensor with components $J^l_{\ i} \equiv \sqrt{g}\epsilon_{mi}g^{lm}$ which is of type (1,1) that also squares to the negative identity. This is in other words an almost complex structure on $T\Sigma$ which fiberwise square to the negative identity. We now claim that the tangent spaces are given by

$$T_p\Sigma = \operatorname{span}_{\mathbb{R}}\{u, Ju\}$$

for some choice of vector field u. We show the $\mathbb{R}$-linear independence of these fields, which is really quite trivial. It is clear that the equality $Ju = \lambda u$ cannot be true for real numbers λ as may be seen simply by acting on both sides with J. Intuitively the linear independence is also clear from the fact that we can interpret J as a rotation by $\pi/2$ making them orthogonal.

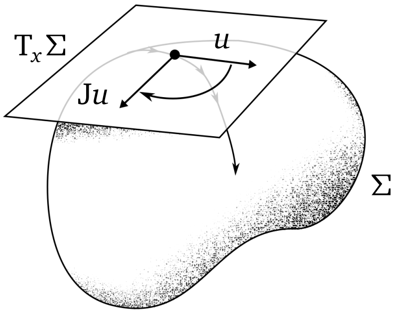


Figure 1.9: We can use the almost complex structure to construct a basis for the tangent spaces.

Then it should be sufficient to evaluate $N(u, Ju)$ as any vector field may be written in terms of this basis:

$$\begin{aligned} N(u,Ju) &= [u,Ju] + J[Ju,Ju] + J[u,J^2u] - [Ju,J^2u] \\ &= [u,Ju] + [Ju,u] \\ &= 0. \end{aligned}$$

Thus, from the Newlander-Nirenberg theorem, any two dimensional orientable Riemannian manifold is a complex manifold.

The second proof is more constructive, and can be found in [60] in the chapter on Bosonic string theories. We will here go trough the same proof in a more detailed way , as it makes use of a useful form of the metric and explicitly shows that transition functions are holomorphic without the need to invoke the machinery of almost complex structures. The proof uses the result that on a Riemann surface the metric is conformally flat [84], meaning

$$g(V,V) = ds^2 = f^2(x,y)(dx^2 + dy^2)$$

or equivalently in terms of complex coordinates $f^2(z,\bar{z})dzd\bar{z}$. We now want find the expression for this in another chart:

$$z \to w(z,\bar{z}) = u(x,y) + iv(x,y),$$

$$f^2(z,\overline{z})dzd\overline{z} \to g^2(w,\overline{w})dwd\overline{w}.$$

On some non zero overlap of these patches we may write the differentials

$$dw = \frac{\partial w}{\partial z}dz + \frac{\partial w}{\partial \overline{z}}d\overline{z},$$

$$d\overline{w} = \frac{\partial \overline{w}}{\partial z}dz + \frac{\partial \overline{w}}{\partial \overline{z}}d\overline{z}.$$

However, on the overlap the metric should match so we can set

$$f^2(z,\overline{z})dzd\overline{z} = g^2(w,\overline{w})dwd\overline{w} \tag{1.38}$$

$$= g^2(w,\overline{w})\left(\frac{\partial w}{\partial z}dz + \frac{\partial w}{\partial \overline{z}}d\overline{z}\right)\left(\frac{\partial \overline{w}}{\partial z}dz + \frac{\partial \overline{w}}{\partial \overline{z}}d\overline{z}\right) \tag{1.39}$$

and we may equate term by term on the left and right hand side. The $dzdz$ term then yields

$$\frac{\partial w}{\partial z}\frac{\partial \overline{w}}{\partial z} = 0. \tag{1.40}$$

If we want to show that the transition functions are holomorphic, i.e. that w is complex analytic in z, we must show that $\partial w/\partial z$ can not be zero. We assume that it is, and do a proof by contradiction. The coordinate change would lead to a Jacobian of the form

$$\text{Jac} = \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{bmatrix} = \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial u}{\partial y} & -\frac{\partial u}{\partial x} \end{bmatrix} \tag{1.41}$$

where we used the Cauchy - Riemann conditions for anti-holomorphic functions. Clearly the determinant det(Jac) is strictly negative as it is a negative sum of squares. In other words, our manifold is non-orientable which is agains our initial assumptions. Thus, $\partial w/\partial z$ can't be zero and we must have that

$$\frac{\partial \overline{w}}{\partial z} = 0$$

or equivalently $\partial w/\partial \overline{z} = 0$. This concludes the proof as the transition functions have to be holomorphic.

Let us take a step back and see what really has happened. In particular, we have seen that given a metric tensor, we can construct the complex structure

$$\text{J}^{\mu}_{\nu} = \sqrt{g}\epsilon_{\nu\rho}g^{\mu\rho}.$$

However, note that under a Weyl(M) transformation $g \to e^{\omega}g$ the complex structure transforms

$$\text{J} \to e^{(n/2-1)\omega}\sqrt{g}\epsilon_{\nu\rho}g^{\mu\rho} = \text{J}$$

as $n = \dim_{\mathbb{R}}(M) = 2$. Hence only the conformal equivalence class $[g_{\mu\nu}]$ contributes to the definition of the complex structure. This is related to the fact the J can be seen as a rotation in the tangent spaces, telling us how to orient vector fields in relation to each other but says nothing about their lengths. This presents an alternative view on the classification of complex structures. The space of conformal equivalence classes of Riemannian metrics corresponds to the space of complex structures of the surface.

Before we state and discuss the so called Uniformization of Riemann surfaces, we need to discuss quotients of Riemann surfaces. We have discussed quotient from the point of view of topology, where constructed quotient spaces from an equivalence relation (see appendix). However, we have not claimed that we are ever dealing with a quotient *manifold*. For the quotient space to be a manifold, there are certain requirements. We will here assume that the equivalence relation is obtained from a group action of G on a space X.

First a few definitions. Let the group G have a group action $(g,x) \to g(x)$ on a space X. The set of points $\mathcal{O}(x,G) = \{y \in X | y = g(x), g \in G\}$ is called the orbit of x under G. The group action is called properly discontinuous when the two following conditions are met [38]

1. Let $x \in U_x \subset X$, $g \in G$ and $x \notin \mathcal{O}(y,G)$. For a properly discontinuous action we must have that $g(U_x) \cap U_x = \emptyset$.

2. Let $x, y \in X$ with neighborhoods U_x and U_y, and $x \notin \mathcal{O}(y,G)$. For a properly discontinuous action we must have that $U_x \cap g(U_y) = \emptyset$.

See also [84] for alternate discussion on the subject. Consider now the case where X is a complex manifold. Let us assume that we are given a quotient map $\pi : X \to X/G$. The quotient space X/G is then a complex manifold if the group action is properly discontinuous. If X is covered by the subsets $\{U_i\}$, the charts on the quotient are naturally given by $\phi_i \circ \pi^{-1} : \pi(U_i) \to \phi_i(U_i) \subset \mathbb{C}$ [84][38].

Another tool useful for studying quotients of Riemann surfaces is the notion of a fundamental domain. These contain, in particular, topological information regarding the quotient.

DEFINITION: A fundamental domain D of a group G acting on a space X is a subset $D \subset X$ such that

1. For point $x \in X$ there is a $g \in G$ such that $g(x) \in D$.

2. Two (or more) points in D can not lie in each others orbits.

3. The union $\cup_g g(\mathrm{D}) = \mathrm{X}$ tesselates X. It is clear from this that be have to be careful when discussing the boundaries of D.

For a trivial example we can consider the quotient $\mathbb{R}/\mathbb{Z}$. The interval $[0,1)$ then is a fundamental domain. Clearly no two points in this interval lies in each others orbits, and any point can be shifted to the right or the left to lie in $[0,1)$. Since we do not include the endpoint 1 the translates $[1,2)$ and $[-1,0)$ do not overlap and in total tesselates the line trivially.

The second main result, here stated without proof, is the so called Uniformization theorem of Riemann surfaces. This theorem does not only give us insight into how we may construct Riemann surfaces, but also gives a first classification of them.

UNIFORMIZATION THEOREM: *Let Σ be a connected Riemann surface and $\tilde{\Sigma}$ its universal covering space. Then this covering space is isomorphic to either the Riemann sphere, the complex plane or the upper half plane.*

For a larger discussion of this theorem see [84]. From our discussion on homotopy groups and universal covering spaces, it should then be clear that every Riemann surface Σ can be obtained by a quotient of one of these simply connected spaces by a group of automorphisms. This group is the first homotopy group $\pi_1(\Sigma)$ of the resulting Riemann surface. Thus, the study on Riemann surfaces can largely be reduced to studying (subgroups of) the automorphism groups of these simply connected spaces, and their group action. This is not a complete classification of Riemann surfaces as complex manifolds however, since there may exist inequivalent complex structures. A particularly important class of surfaces for us is the case where the universal covering space is $\mathbb{C}$ and group of automorphisms is a translational group isomorphic to $\mathbb{Z} \times \mathbb{Z}$. Note that for the complex plane the automorphisms are either such translations or rotations, but rotations necessarily have fixed points and do not act properly discontinuous on $\mathbb{C}$. Hence, the only Riemann surface with covering the complex plane is the genus one torus. Similar observations can be made for the spherical case, but in this case all automorphisms have fixed points. For the hyperbolic plane, the automorphisms are the Fuchsian groups. In conclusion, all Riemann surfaces are either a sphere, a torus or quotients of hyperbolic space by Fuchsian groups.

As Riemannian manifolds the Riemann surfaces inherit a metric form their covers. The curvature of these can be obtained by the Gauss-Bonnet theorem [58]

$$\frac{1}{4\pi}\int_{\Sigma} d^2x\sqrt{g}\mathrm{R} = \chi(\Sigma) = 2-2g$$

where R is the Ricci curvature scalar of the surface. Hence we can relate the genus to the curvature. Since the Riemann surfaces gain a metric from their covers, and therefore also curvature, we can relate genus to the three model geometries of the uniformization theorem. Genus 0 correspond to the positively curved sphere, genus zero to the flat torus, and higher genus to the negatively curved hyperbolic surfaces.

1.10.1 Line bundles over Riemann surfaces

Before moving on to studying the complex torus we discuss the geometry of Riemann surfaces in somewhat more detail. In particular we want to understand line bundles. For a larger discussion on the subject see [84].

From our discussion of complex structures we know that the tangent bundle over a complex manifold may be decomposed into holomorphic and antiholomorphic line bundles

$$\mathrm{T}^{\mathbb{C}}\Sigma = \mathrm{T}\Sigma_{\mathrm{Hol}} \oplus \mathrm{T}\Sigma_{\mathrm{aHol}}$$

and similarly for the cotangent bundle. We will mainly be focusing on holomorphic vector fields and holomorphic one-forms, e.g. holomorphic vector fields and the canonical bundle. From the section on fiber bundles, we remember that a local choice of coordinates on $\mathrm{U} \subset \Sigma$ also trivialized the tensor bundles. Consider for example the canonical line bundle of holomorphic one-forms. On the intersections $\mathrm{U}_a \cap \mathrm{U}_b$ the one-forms will transform as

$$f(z)dz \to f(w)dw = f(w(z))\frac{\partial w}{\partial z}dz$$

with the transition functions $\frac{\partial w}{\partial z} = \psi_{ba} : \mathrm{U}_a \cap \mathrm{U}_b \to \mathrm{SO}(2)$. These clearly satisfy the conditions discussed in the section on fiber bundles, in particular

$$\psi_{ab}\psi_{bc}\psi_{ca} = \frac{\partial z_a}{\partial z_b}\frac{\partial z_b}{\partial z_c}\frac{\partial z_c}{\partial z_a} = 1$$

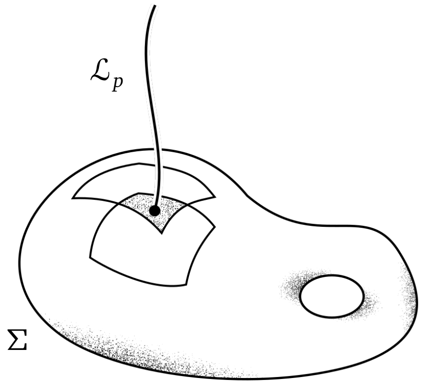


Figure 1.10: A line bundle over a point p in the intersection of charts U_a and U_b.

The uniformization theorem gives us a simple way to construct Riemann surfaces using quotients. Let us see how this extends to the line bundles. Let X denote the universal covering of a Riemann surface Σ and G the (discrete) group action by which we want to quotient. The family of points $x_1, x_2, \ldots$ in X are the orbits of G which all are projected to $\mathrm{P}(x_i) \equiv p$ in Σ. Let also $\mathcal{L}$ be the line bundle of either holomorphic vector fields or holomorphic one-forms. We purpose the following identification: *A line bundle $\mathcal{L}$ over $\Sigma = \mathrm{X}/\mathrm{G}$ is identified with a G-equivariant line bundle $\tilde{\mathcal{L}}$ over* X.

We could view $\mathcal{L}$ as a pullback bundle by the projection P, but as the inverse projection is one-to-many the bundles over points related by G needs to be identified. For a small discussion of a similar[3] case see [13].

[3] In this reference a similar statement is made regarding line bundles over quotient of Lie groups.

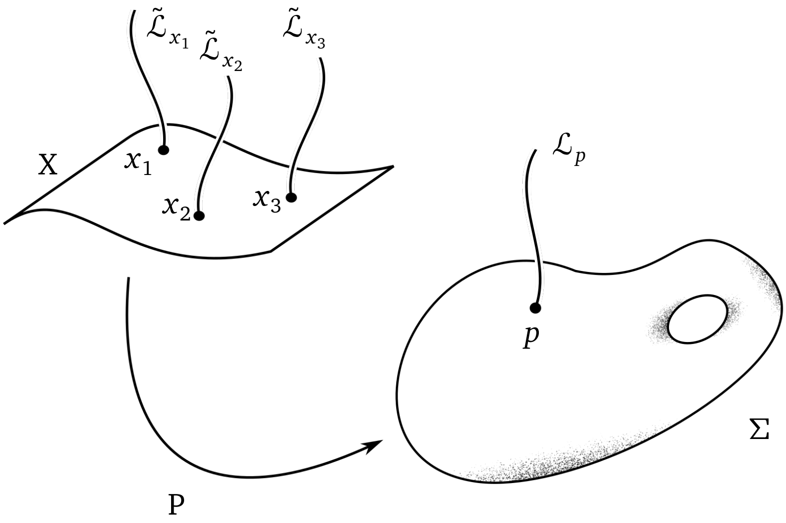

$\tilde{\mathcal{L}}_{x_1}$
$\tilde{\mathcal{L}}_{x_2}$
$\tilde{\mathcal{L}}_{x_3}$
X
x_1
x_2
x_3
$\mathcal{L}_p$
p
Σ
P

Figure 1.11: As the points $x_1, x_2, ...$ are G-equivalent, the fibers above should be identified when taking the quotient. In this way the bundles are "projectable" to the Riemann surface.

Let us see what this identification means for the line bundle of holomorphic one-forms. Let z be a point that is projected to p, and $g(z)$ for $g \in G$ a group translate of it. Under the G-action we identify

$$f(z)dz \overset{!}{=} f(g(z))dg(z) = f(g(z))\frac{\partial g(z)}{\partial z}dz$$

implying the transformation rule

$$f(g(z)) = f(z)\left(\frac{\partial g(z)}{\partial z}\right)^{-1}.$$

Hence the study of one-forms on a Riemann surface obtained by a quotient X/G can be seen as a study of functions on X invariant up to an extra factor.

1.10.2 Counting spin structures

In the case of two dimensions, spin structures are particularly kind. As we will see there is a more geometric interpretation of spinors that will be useful when counting spin structures. We consider the holomorphic line bundle $T^*\Sigma_{\mathrm{Hol}} = \mathcal{L}$ of one-forms over a Riemann surface Σ, where in a local trivialization the one-forms take the form

$$\omega = \omega(z)dz.$$

Since Riemann surfaces are Riemannian manifolds, and hence have SO(2)-structure, the transition functions ψ_{ab} are SO(2)-valued functions on the chart overlaps $U_a \cap U_b$. We write $\psi_{ab} = R_{ab}(\theta)$. From this line bundle we can construct what is known as a square root of the line bundle. A square root of this line bundle is another line bundle $\mathcal{S}$ such that $\mathcal{S} \otimes \mathcal{S} = \mathcal{L}$ [10]. The transition functions on $\mathcal{S}$ satisfy $\tilde{R}^2_{ab}(\theta) = R_{ab}(\theta)$, hence

$$\tilde{R}_{ab} = \pm\sqrt{R_{ab}}.$$

The sign ambiguity in taking the square root of the transition functions imply that the cocycle condition is only satisfied mod $\mathbb{Z}_2$, as we discussed earlier in more generality. The bundle $\mathcal{S}$ contains half-order forms, written $\omega(z)\sqrt{dz}$, where the square root is nothing more than a formal symbol reflecting the fact that the tensor product of two such half-forms yields a 1-form. The sections of this line bundle will act as spinors under the rotation group [3][10]. However, constructing such a square root is highly non-canonical.

We here sketch the construction of the square root line bundle. We consider a Riemann surface obtained by a quotient $\pi : X \to \Sigma = X/G$, where $G = \pi_1(\Sigma)$. We recall that the first homotopy group acts on the cover X by permuting preimages $\pi^{-1}(p)$. Let $\tilde{\mathcal{S}}$ be the square root of $\tilde{\mathcal{L}}_{can}$ on the cover X with half-forms $\omega = f(z)\sqrt{dz}$. The half-forms on Σ is then identified with G-equivariant half-forms

$$f(g(z))\sqrt{\frac{\partial g(z)}{\partial z}dz} = f(z)dz$$

$$\therefore f(g(z)) = f(z)\left(\frac{\partial g(z)}{\partial z}\right)^{-1/2}.$$

Taking this square root results in a sign ambiguity. The transformations $z \to g(z)$ are generated by transporting points along cycles in Σ.

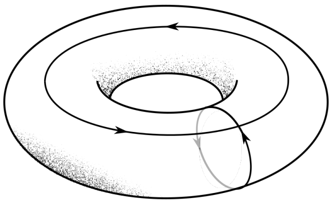


Figure 1.12: On the torus the first homotopy group is $\mathbb{Z} \times \mathbb{Z}$, corresponding to winding numbers around the two non-homotopic closed loops. As there are two possible signs to chose for each path, there are 4 square roots of $\mathcal{L}_{can}$

For each genus g there are 2 such loops, resulting in

$$\underbrace{2 \cdot 2 \cdot \ldots \cdot 2}_{\text{2g times}} = 2^{2g}$$

sign choices. These are the 2^{2g} inequivalent ways to construct the bundle $\mathcal{S}$, and hence the inequivalent spin structures on Σ [10][3]. Thus on the torus, as a genus 1 surface, there are four spin structures, corresponding to the four different combinations of periodic or anti-periodic boundary conditions.

2

Geometry of the ring of modular forms

The theory of modular forms is an old mathematical theory that still is an active area of research. We wish to present this theory in an almost purely geometric form. Not only is this an attractive approach to modular forms, but it offers a nice introduction to the subject for people more familiar with geometry than complex analysis and number theory. The hope is that this chapter can serve as a clear introduction to this wonderful mathematical theory for physicist familiar with the geometry of general relativity and field theory.

2.1 Overview

The theory of modular forms is a large field with many components. Before we start we want to present a brief overview of the subjects we will discuss.

The story starts with complex tori. These are the complex counterpart of the familiar genus 1 surface in two real dimensions. As a complex manifold these tori are classified by invertible holomorphic mappings, as we discussed in the previous chapter. The moduli space of these objects is a quotient space. By placing further structure on the complex torus, this moduli space will change so that the holomorphic mappings respect this additional structure. The holomorphic 1-forms dz and their k-fold tensor products over these moduli spaces are modular forms. In this way, we can see modular forms as sections of a certain line bundle, on which we can find a connection 1-form. This will occupy much of our time, and lead to a geometric understanding of the so-called differential ring of quasi-modular forms.

The modular forms also appear when considering functions on the torus. These can be seen as doubly periodic functions, and are often referred to as elliptic functions in the holomorphic case. By performing a series expansion of a certain class of these functions, we will see that certain modular forms appear as coefficients. Furthermore, these elliptic functions and their derivative satisfy certain algebraic relations known as elliptic curves. These algebraic objects can be shown to be isomorphic to the original complex torus.

2.2 Complex tori

From the point of view of the uniformization theorem of Riemann surfaces, the complex torus is the genus one, flat Riemann surface with universal covering the entire complex plane. For alternative discussions see for example [60], [84] or more or less any book on Riemann surfaces. The automorphisms of the complex plane can be taken to be rotations and translations. Since the rotations have fixed points, we have to quotient by the translations. The discrete translations form a lattice $\Lambda \subset \mathbb{C}$, viewed as a subgroup of $(\mathbb{C},+)$. We write

$$\Lambda_{\omega_1,\omega_2} = \omega_1\mathbb{Z} \times \omega_2\mathbb{Z}$$

where ω_1 and ω_2 are $\mathbb{R}$-linearly independent as vectors in the plane. The quotient obtained by the equivalence $z \sim z+\omega$ for $\omega \in \Lambda_{\omega_1,\omega_2}$ is the complex torus

$$\mathrm{E}_{\omega_1,\omega_2} = \mathbb{C}/\Lambda_{\omega_1,\omega_2}$$

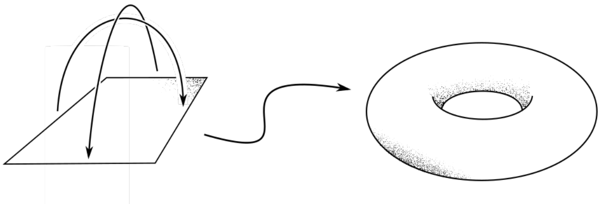


Figure 2.1: The 2-torus obtained from a identification of the sides of a parallelgram in the complex plane.

The complex tori are also called complex elliptic curves, hence the above notation $\mathrm{E}_{\omega_1,\omega_2}$. We will later justify this terminology by discussing the relation between the theory of complex tori and the theory of elliptic curves over $\mathbb{C}$. We will use the two names interchangeably.

Recall that in general, the complex structure is obtained by having holomorphic transition functions $\phi_i \circ \phi_j^{-1}$. The charts on a quotient can be chosen to be $\phi_i \circ \pi^{-1}$, with ϕ_i the charts of the covering. The transition functions on the quotient manifold is then the composition

$$(\phi_i \circ \pi^{-1}) \circ (\phi_j \circ \pi^{-1})^{-1} = (\phi_i \circ \pi^{-1}) \circ (\pi \circ \phi_j^{-1}) = \phi_i \circ \phi_j^{-1}.$$

In this sense the complex structure of the quotient is inherited from the covering. In the case of the torus the complex structure is obtained by a choice (ω_1, ω_2). We will therefore refer to this choice of lattice as a complex structure.

2.3 Moduli space of complex tori

We want to classify the complex tori by biholomorphic equivalence. Recall that classification problems often comes in two steps [1]. First one identifies the discrete, often topological, classification of the objects one wants to study. The moduli space first appears when one for fixed discrete structures searches for parameters that further divides the objects into continuous families. In our present case, the discrete invariants are the genus of the surface. This section discusses the moduli space associated with genus one surfaces following [60][21].

According to our discussion on complex manifolds, the moduli space of complex tori should schematically be

$$\mathscr{M}(\mathrm{E}_{\omega_1,\omega_2}) = \{\text{Complex Tori } \mathrm{E}_{\omega_1,\omega_2}\} \,/\, \text{Biholomorphisms}.$$

Before we construct the biholomorphisms, we need the notion of equivalent lattices.

The idea behind the classical approach to complex torus classification is to realize that the operation that gives equivalent lattices induces a biholomorphic map between the tori. This proof follows [60] quite closely, and is included for the sake of completeness. We consider two lattices $\Lambda_{\omega_1,\omega_2}$ and $\Lambda_{\omega_1',\omega_2'}$. If the orbits of these lattices are to coincide we must at least have that the ω_i lie in the orbit of $\Lambda_{\omega_1',\omega_2'}$, i.e. $(\omega_1,\omega_2) = (a\omega_1' + b\omega_2', c\omega_1' + d\omega_2)'$. In matrix notation this reads

$$\begin{bmatrix}\omega_1\\ \omega_1\end{bmatrix} = \begin{bmatrix}a & b\\ c & d\end{bmatrix}\begin{bmatrix}\omega_1'\\ \omega_2'\end{bmatrix},$$

which we write $\omega = \mathrm{M}\omega'$ for short. The opposite must also be true, i.e. $\omega_i{}'$ must lie in the orbit of $\Lambda_{\omega_1,\omega_2}$. By inverting the above matrix relation we get

$$\det(\mathrm{M})\begin{bmatrix}\omega_1'\\ \omega_2'\end{bmatrix} = \begin{bmatrix}d & -b\\ -c & a\end{bmatrix}\begin{bmatrix}\omega_1\\ \omega_2\end{bmatrix}.$$

Hence if $\det(\mathrm{M}) = \pm 1$ the ω_i' also lie in the orbit of Λ, and the two lattices coincide. The linear independence of the two generating vectors can be expressed as $\Im\tau > 0$ for $\tau = \omega_2/\omega_1$. For the transformed lattice $\omega' = \mathrm{M}\omega$ with $\omega_1' = a\omega_1 + b\omega_2$, $\omega_2' = c\omega_1 + d\omega_2$ we get

$$\Im\tau' = \Im\left[\frac{c + d\tau}{a + b\tau}\right] = \frac{\det(\mathrm{M})}{|a + b\tau|^2}\Im\tau.$$

So, to preserve the linear independence the determinant must be positive. Hence we are dealing with a $\mathrm{SL}_2(\mathbb{Z})$ transformation. Note that the above arguments still

[1] See the appendix for a discussion on structure and classification.

hold if we in stead used the matrix $-M$. Hence we could mod out this $\mathbb{Z}_2$ action to get the projective special linear group $\mathrm{PSL}_2(\mathbb{Z})$. This is also called the modular group, which we will denote $\Gamma(1)$ for reasons that will become clear later. The modular group is an infinite discrete group generated by two elements T,S that satisfies $S^2 = (ST)^3 = 1$ [21]. On the upper half plane, these generators act by $T : z \to z+1$, $S : z \to -1/z$. More generally, a modular transformation acts by

$$\gamma(z) = \begin{bmatrix} a & b \\ c & d \end{bmatrix}(z) = \frac{az+b}{cz+d}.$$

This group plays a crucial role in the classification of tori. Being a complex manifold, the equivalent complex tori are related by a invertible holomorphic map, i.e. a biholomorphism

$$\varphi : \mathbb{C}/\Lambda_{\omega_1,\omega_2} \to \mathbb{C}/\Lambda_{\omega_1',\omega_2'}.$$

The main idea is to use the projection $\pi : \mathbb{C} \to \mathbb{C}/\Lambda_{\omega_1,\omega_2}$ to lift φ to a holomorphic map f on the complex plane. Intuitively, the job of this map is to deform one lattice into the other so that the following diagram commutes

$$\begin{tikzcd}
\mathbb{C} \arrow[r, "f"] \arrow[d, "\pi"'] & \mathbb{C} \arrow[d, "\tilde{\pi}"] \\
\mathbb{C}/\Lambda_{\omega_1,\omega_2} \arrow[r, "\varphi"] & \mathbb{C}/\Lambda_{\omega_1',\omega_2'}
\end{tikzcd}$$

The following theorem, taken from [21], summarizes the classification of complex tori.

THEOREM: *Let $\varphi : \mathbb{C}/\Lambda_{\omega_1,\omega_2} \to \mathbb{C}/\Lambda_{\omega_1',\omega_2'}$ be a invertible holomorphic map between complex tori. Then there are complex numbers λ, b such that $\varphi(z+\Lambda_{\omega_1,\omega_2}) = \lambda z + b + \Lambda_{\omega_1',\omega_2'}$ and $\lambda\Lambda_{\omega_1,\omega_2} = \Lambda_{\omega_1',\omega_2'}$. Here $z+\Lambda$ denotes the equivalence class of $z \in \mathbb{C}$ under the lattice. The two tori are in this case said to be biholomorphically equivalent.*

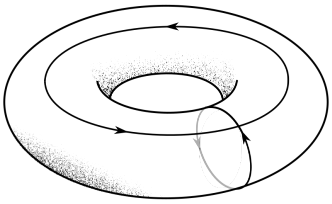
In particular, consider the lattice $\Lambda_{\omega_1,\omega_2}$ written as $\omega_1(\mathbb{Z}\times\tau\mathbb{Z})$, where as before $\tau = \omega_2/\omega_1$. This can be seen as a map $\mathbb{C}/\Lambda_{\omega_1,\omega_2} \to \mathbb{C}/\Lambda_{1,\tau}$ with lattices equivalent up to a overall scaling. By the above theorem, these tori are isomorphic as complex manifolds. However, this choice of τ is not unique as there are many equivalent lattices under the modular group, i.e. $\tau' = \gamma(\tau)$ for $\gamma \in \Gamma(1)$. Thus, a complex structure τ determines a complex torus up to modular transformations. The moduli space of complex tori is [21] the orbit space

$$\mathcal{M}(E_{1,\tau}) = \mathfrak{H}/\mathrm{PSL}_2(\mathbb{Z}) = \mathfrak{H}/\Gamma(1).$$

The geometry of this moduli space will play a central role in most upcoming discussions. Such quotients of the upper half plane by the modular group or its subgroups are called modular curves.

2.4 Moduli space of complex tori with spin structure

When the holomorphic map between equivalent tori have to respect more geometric structure, the moduli space changes. By adding spin structures on the torus, we will see that the moduli space can be constructed as a quotient as before, but now with a subgroup of the modular group. Recall that when we counted spin structures on Riemann surfaces, we identified spinors on Σ with the square root of the canonical line bundle of holomorphic 1-forms. The number of spin structures corresponded to the number of ways we could take this square root, which we found to be 2^{2g} for a genus g surface. On a torus there are four spin structures. A choice of spin structure is equivalent to choosing either periodic (P) or anti-periodic (A) boundary conditions along the two fundamental loops on the torus.



These four spin structures we denote (P,P), (P,A), (A,P) and (A,A). We will denote the torus with a particular spin structure $\# = (\mathrm{P},\mathrm{P}),(\mathrm{P},\mathrm{A}),(\mathrm{A},\mathrm{P}),(\mathrm{A},\mathrm{A})$ simply $\mathrm{E}^{\#}_{1,\tau}$.

To connect the above four spin structures to particular level 2 subgroups Γ of the full modular group $\Gamma(1)$, we should first study how spinors transform under the two generators of the modular transformations. Once the subgroups that fixes particular spin structures are identified, we can study the holomorphic geometry on the moduli space $\mathfrak{H}/\Gamma$ of complex tori with spin structure $\mathrm{E}^{\#}_{1,\tau}$ to learn about the modular forms.

The modular transformations can be seen as operations that cut, twist and glue the complex tori, and hence will mix the four different boundary conditions. If we insist of having a fixed spin structure, we need to identify the subgroups that preserve the given boundary conditions.

First we should see how the modular group generators T and S changes the spin structures. Consider coordinates $x+\tau y$ on $\mathrm{E}_{1,\tau}$. Under $\mathrm{S}:\tau\to -1/\tau$ we have the transformation

$$x+\tau y \to \frac{\tau x - y}{\tau}$$

so this generator essentially maps $(x,y)\to(-y,x)$, up to a scaling. Similarly, under the translations $\mathrm{T}:\tau\to\tau+1$ the coordinates get mapped as $(x,y)\to(x+y,y)$. Knowing this, we can see how spinors on $\mathrm{E}_{1,\tau}$ transform. Consider first the (A,P) spin structure. Then

$$\psi(x,y)\overset{(\mathrm{A},\cdot)}{=}-\psi(x+1,y)$$

$$\downarrow \mathrm{S}$$

$$\psi(-y,x)=-\psi(-y,x+1)$$

By this notation we mean that the above equality is true because of the anti periodicity in the first argument, and that we transform both the left and right hand side of this equation by S to get a new boundary condition on the transformed torus. Similarly

$$\psi(x,y)\overset{(\cdot,\mathrm{P})}{=}\psi(x,y+1)$$

$$\downarrow \mathrm{S}$$

$$\psi(-y,x)=\psi(-y-1,x)$$

In total this means that $\mathrm{S}:(\mathrm{A},\mathrm{P})\to(\mathrm{P},\mathrm{A})$. We treat the action of the second generator likewise.

$$\psi(x,y)\overset{(\mathrm{A},\cdot)}{=}-\psi(x+1,y)$$

$$\downarrow \mathrm{T}$$

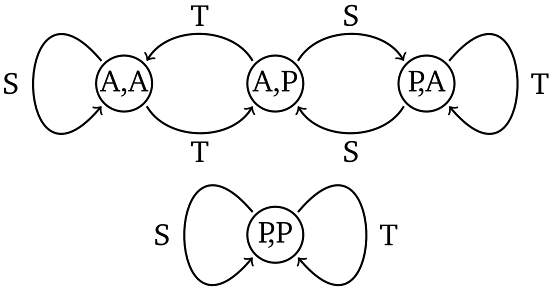
$$\psi(x+y,y)=-\psi(x+y+1,y) \tag{2.1}$$

$$\psi(x,y)\overset{(\cdot,\mathrm{P})}{=}\psi(x,y+1)$$

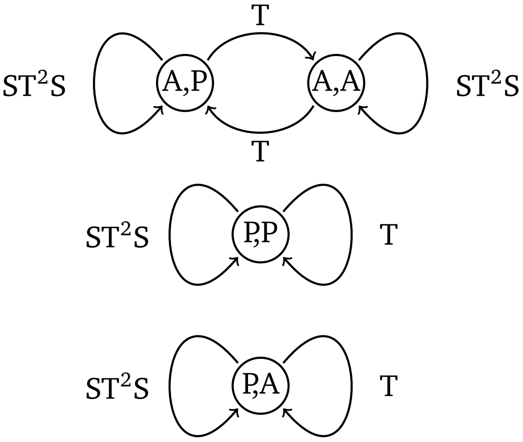
$$\downarrow \mathrm{T}$$

$$\psi(x+y,y)=\psi(x+y+1,y+1)$$

Since anti-periodicity in the first argument is conserved under T by equation (2.1) we can write $\psi(x+y,y)=-\psi(x+y,y+1)$. Hence $\mathrm{T}:(\mathrm{A},\mathrm{P})\to(\mathrm{A},\mathrm{A})$. Similarly we can proceed to find how the three other spin structures transform. The results are summarized in the below graph.



These transformation rules coincides with the ones discussed in for example [75]. The only spin structure preserved by the full modular group is the trivial (P,P) spin structure. Knowing how the spin structures mix under the modular generators we can draw a similar diagram for the subgroups $\Gamma \subset \Gamma(1)$. Most relevant in our discussions is the subgroup generated by T and ST^{2}S, denoted $\Gamma_0(2)$. Using the above diagram we can easily draw a diagram for this subgroup.



We see that now also the (P,A) spin structure is fixed. Similarly, there are other subgroups that preserve the other spin structures. In terms of combinations of the generators for the full modular group, these are given by [68]

$$\Gamma(2) = \left\langle T^2, ST^2S \right\rangle \tag{2.2}$$

$$\Gamma_0(2) = \Gamma_T = \left\langle T, ST^2S \right\rangle = \left\langle T, R^2 \right\rangle \tag{2.3}$$

$$\Gamma^0(2) = \Gamma_R = \left\langle T^2, STS \right\rangle = \left\langle R, T^2 \right\rangle \tag{2.4}$$

$$\Gamma_\theta(2) = \Gamma_S = \left\langle S, T^2 \right\rangle \tag{2.5}$$

$$\Gamma_2 = \Gamma_P = \langle ST, TS \rangle = \langle P, SPS \rangle \tag{2.6}$$

where $R = TST = ST^{-1}S$. The group $\Gamma(2)$ is called the principal congruence subgroup. The other groups are the four groups that lie between this principal subgroup and the full modular group $\Gamma(1)$ [68]. There are several interesting conjugacies between these groups. Recall that two subgroups H_1,H_2 of a group G are said to be conjugate if there is a $g \in G$ such that $H_1 = gH_2g^{-1}$. In this way subgroups fall into conjugacy classes. The groups $\Gamma_{T,R,S}$ are conjugate by $\Gamma_R = S\Gamma_T S^{-1}$ and $\Gamma_S = T\Gamma_R T^{-1}$ [64]. We can also consider conjugation by the $GL_2(\mathbb{Q})$ operation $G(z) = 2z$. We then also have the relation $\Gamma_R = G\Gamma_T G^{-1}$. This conjugacy is also used to prove the isomorphism of the groups $\Gamma_0(4) = \{\gamma \in \Gamma(1) | c = 0 \text{ mod} 4\}$ and $\Gamma(2)$ [12]. We will discuss these conjugacy classes in more detail in later chapters, as well as in the appended paper where we discuss these modular groups in the context of the quantum Hall effect. For more information on the subgroups of the modular group see for example [68].

Using the above diagrams one can verify that the (A,A) spin structure is fixed under Γ_S, and (A,P) under Γ_R. In other words, the three subgroups correspond exactly to the subset of modular transformations that preserve one of the spin structures. When we constructed the moduli space of complex elliptic curves we considered the equivalence classes of lattice generators $(1, \tau)$ under the full modular group. To have a moduli space of a complex elliptic curve with additional spin structure we have to quotient by one of the above subgroups. For general $\Gamma_X \subset \Gamma(1)$ we get the complex curve $\mathfrak{H}/\Gamma_X$. For example, the moduli space of an elliptic curve with (P,A) spin structure is the modular curve $\mathfrak{H}/\Gamma_T$. We now study holomorphic tensors on these modular curves.

2.5 Modular forms from holomorphic geometry

We start by analyzing holomorphic forms on the moduli spaces $\mathfrak{H}/\Gamma_X$. We will for ease of notation write $\mathcal{L}_* = T^*\mathfrak{H}_{holo}$ and $\mathcal{L} = T\mathfrak{H}_{holo}$. Recall from our earlier discussions on line bundles on Riemann surfaces that to construct holomorphic one-forms on a quotient we in stead work on the covering space but demand

$$f(g(z)) = f(z)\left(\frac{\partial g(z)}{\partial z}\right)^{-1}.$$

For holomorphic vectors, e.g. sections of $\mathcal{L}$, the inverse transformation rule holds

$$f(g(z)) = f(z)\left(\frac{\partial g(z)}{\partial z}\right)^{+1}$$

First, let $\gamma \in \Gamma(1)$ and consider $z \to \gamma(z)$ for $z \in \mathfrak{H}$. Then

$$\partial_z \gamma(z) = \frac{\partial}{\partial z}\frac{az+d}{bz+c} = \frac{a}{cz+d} - c\frac{az+b}{(cz+d)^2} = (cz+d)^{-2}.$$

Hence, to construct holomorphic 1-forms on the moduli space $\mathscr{M}(E_{\omega_1,\omega_2})$ we must demand

$$f(\gamma(z)) = f(z)(cz+d)^2$$

where $f : \mathfrak{H} \to \mathbb{C}$. If we construct the k-fold tensor product bundle $\mathcal{L}_*^{\otimes k} = \mathcal{L}_* \otimes \ldots \otimes \mathcal{L}_*$ with fiber basis $dz \otimes \ldots \otimes dz$, the transformation rule will similarly be

$$f(\gamma(z)) = f(z)(cz+d)^{2k}.$$

These functions are called modular forms of weight 2k [41]. If we in stead consider the bundle $\mathcal{L} \otimes \ldots \otimes \mathcal{L}$ we would get

$$f(\gamma(z)) = f(z)(cz+d)^{-2k}$$

which transform as modular forms of negative weight.

We often work only with the components of these tensors. Under the generators S, T of the modular group, the requirement for a modular form reads

$$f(S(z)) = z^k f(z),$$

$$f(T(z)) = f(z).$$

Hence modular forms are periodic under T and can be expanded in a Fourier series. Defining $q = \exp(2\pi i z)$ we have a q-expansion

$$f = \sum_{n\in\mathbb{Z}} a_n q^n.$$

The holomorphicity at infinity is now equivalent with the fact the $a_n = 0$ for negative n. A modular form that vanishes at $i\infty$ is called a cusp form [21]. Note that since q vanishes in the limit $\tau \to i\infty$, we have $f(i\infty) = a_0$. Often one normalizes the modular form so that $a_0 = 1$. We will explicitly do this later for the Eisenstein series.

We denote the space of modular forms $M_k(\Gamma)$ for some subgroup Γ. From the above, it should be clear that this space is closed under both addition and (complex) scalar multiplication, so $M_k(\Gamma)$ is a vector space over $\mathbb{C}$. Note that if $f(z)$ is a weight k modular form and $g(z)$ a form of weight l, the product $f(z)g(z)$ is a modular form of weight $(k+l)$. If we consider the space of all modular forms

$$M(\Gamma) = \bigoplus_k M_k(\Gamma)$$

this space has the structure of a ring. We will call it the modular ring.

We now turn to modular forms for the subgroups Γ. Just as for the full modular group we consider k-fold tensor products of the line bundles of holomorphic 1-forms $\mathcal{L}_*$ on the upper half plane. Under the quotient by one of the modular subgroups these tensors again have to satisfy an automorphy relation. This time however, the group element belongs to the subgroup. Recall that Γ_T is generated by T and ST^2S, while Γ_R is generated by T^2 and STS. Since there is no factor of automorphy associated with a T^n transformation, we need only consider the other generator for the two subgroups to identify forms.

	Transformation	Automorphic factor for $\mathcal{L}_*$	Automorphic factor for $\mathcal{L}_*^{\otimes N}$
STS	$\tau \to (1/\tau - 1)^{-1}$	$(\tau-1)^2$	$(\tau-1)^{2N}$
ST^2S	$\tau \to (1/\tau - 2)^{-1}$	$(2\tau-1)^2$	$(2\tau-1)^{2N}$

Again we clearly see the G-conjugacy between Γ_T and Γ_R. Note that for the last subgroup $\Gamma_S = \left\langle S, T^2 \right\rangle$ the factor of automorphy is identical to that of the full modular group. As usual we will replace this even exponent 2N with k and consider only even numbers. We could also consider the principal level 2 congruence subgroup $\Gamma(2) = \langle T^2, ST^2S\rangle$ even though this does not correspond to a spin structure. Modular forms for the principal subgroup then have to satisfy

$$w(z+2) = w(z),$$

$$w(ST^2S(z)) = w(z)(1-2z)^k.$$

We see that the spaces $M_k(\Gamma_T)$ and $M_k(\Gamma(2))$ differ essentially by what q-expansion the forms have. The dimensions of spaces of low-weight forms, which can be found in [64] or [68], is listed in the below table.

	$\Gamma(1)$	Γ_T	Γ_R	Γ_S	$\Gamma(2)$
$\dim(M_0(\Gamma))$	1	1	1	1	1
$\dim(M_2(\Gamma))$	0	1	1	1	2

The fact that dimensions for the three subgroups $\Gamma_{T,R,S}$ coincide is not surprising, as the groups are conjugate of each other. Since conjugacy is an isomorphism, every statement made regarding one of the groups should be translatable to the others. We want to construct the spaces $\mathrm{M}_2(\Gamma_\mathrm{X})$ for the three level 2 subgroups. The main presentation of these that we will use is in terms of the Eisenstein series, which we will come to in a moment. First we briefly discuss a nice way to obtain these level 2 forms by using the Jacobi theta functions, following [64]. The theta functions read

$$\theta_1(\tau) = 2\sum_{n=0}^{\infty}(-1)^n q^{\frac{1}{2}(n+1/2)^2}, \tag{2.7}$$

$$\theta_2(\tau) = \sum_{n=-\infty}^{\infty} q^{\frac{1}{2}(n+1/2)^2}, \tag{2.8}$$

$$\theta_3(\tau) = \sum_{n=-\infty}^{\infty} q^{n^2/2}, \tag{2.9}$$

$$\theta_4(\tau) = \sum_{n=-\infty}^{\infty} (-1)^n q^{n^2/2}. \tag{2.10}$$

Under the two generators of the full modular group T,S these functions transform as

$$\theta_2(\mathrm{T}(\tau)) = e^{i\pi/4}\theta_2(\tau), \theta_2(\mathrm{S}(\tau)) = \sqrt{-i\tau}\theta_4(\tau), \tag{2.11}$$

$$\theta_3(\mathrm{T}(\tau)) = \theta_4(\tau), \theta_3(\mathrm{S}(\tau)) = \sqrt{-i\tau}\theta_3(\tau), \tag{2.12}$$

$$\theta_4(\mathrm{T}(\tau)) = \theta_3(\tau), \theta_4(\mathrm{S}(\tau)) = \sqrt{-i\tau}\theta_2(\tau). \tag{2.13}$$

The theta functions are also related by the Jacobi identity $\theta_3^4 = \theta_4^4 + \theta_2^4$. We start with the space $\mathrm{M}_2(\Gamma(2))$, which is to be 2 dimensional. On can easily verify that

$$\mathrm{ST}^2\mathrm{S} : \theta_2^4 \to (1-2\tau)^2\theta_2^4$$

and similarly for θ_3^4 and θ_4^4. However, due to the Jacobi identity these three forms are linearly related, reducing the dimension to 2. We take as basis θ_2^4 and θ_3^4. Approaching the other subgroups similarly, one can show [64] that

$$\theta_3^4 - \frac{1}{2}\theta_2^4 \in \mathrm{M}_2(\Gamma_\mathrm{T}),$$

$$\theta_3^4 + \theta_2^4 \in \mathrm{M}_2(\Gamma_\mathrm{R}),$$
$$\theta_3^4 - 2\theta_2^4 \in \mathrm{M}_2(\Gamma_\mathrm{S}).$$

To prove this one have to use the Jacobi identity. It may not be clear at the moment that the weight 2 forms for Γ_T and Γ_R are related by conjugation by $\mathrm{G}(\tau) = 2\tau$, but when we later express these modular forms in terms of Dedekind eta functions this becomes manifest. Since these three modular forms are of very similar form, we can consider the linear combination

$$\theta_3^4 - a\theta_2^4 = \theta_3^4(1 - a\lambda).$$

This is modular on $\Gamma(2)$ for all values of a, but for particular choices the symmetry is enhanced to one of the congruence subgroups [64]. Here $\lambda = \theta_2^4/\theta_3^4$ is the $\Gamma(2)$ invariant modular lambda function.

2.6 Eisenstein series and their q-expansion

We have just seen that the weight 2 modular forms on the subgroups Γ_{X} can be understood from Jacobi theta functions. We can equivalently use the Eisenstein series and the Dedekind eta function to understand these spaces. In particular, we wish to understand the ring of modular forms for Γ_{T} in a geometric manner. This section discusses these Eisenstein series, their modular properties and their q-expansions.

2.6.1 Eisenstein series on $\Gamma(1)$

Let k be an even integer bigger than or equal to 4. The Eisenstein series are defined as [41] a sum over lattices of the type

$$\mathrm{G}_k(z) = \sum_{m,n\in\mathbb{Z}\times\mathbb{Z}-\{(0,0)\}} \frac{1}{(mz+n)^k},$$

where the sum runs over all nonzero lattice cites. This function is clearly invariant under $\mathrm{T} : z \to z + 1$. Under the other generator of the modular group we have

$$\mathrm{G}_k(-1/z) = \sum_{n,m} \frac{1}{(1-1/z)^k} = z^k \mathrm{G}_k(z).$$

Hence the Eisenstein series transforms as a modular form of weight k. To check for behavior at the cusp we take the limit

$$\lim_{z\to\infty} \mathrm{G}_k(z) = \lim_{z\to\infty} \sum_{m,n\in\mathbb{Z}\times\mathbb{Z}} \frac{1}{(mz+n)^k}.$$

All $m \neq 0$ terms will vanish in the limit. That leaves the $m = 0$ case, which gives

$$\lim_{z\to\infty} \mathrm{G}_k(z) = \sum_{n\in\mathbb{Z}-\{0\}} 1/n^k = 2\zeta(k)$$

where $\zeta(n)$ is the Riemann zeta function. It is customary to define the normalized series

$$\mathrm{E}_k(z) = \frac{1}{2\zeta(k)}\mathrm{G}_k(z).$$

With this normalization the Eisenstein series may alternatively be written as the q-expansion

$$\mathrm{E}_k(z) = 1 + \frac{(-2\pi i)^k}{(k-1)!\zeta(k)}\sum_{\ell=0}^{\infty}\frac{\ell^{k-1}q^\ell}{1-q^\ell}.$$

We will prove this expansion in later sections. As mentioned above, the Eisenstein series are defined only for forms of weight larger than or equal to 4. The second Eisenstein series does not transform as a modular form due to convergence problems [71], but in stead transforms as

$$\mathrm{E}_2(\gamma(z)) = (cz+d)^2\mathrm{E}_2(z) - \frac{6ic}{\pi}(cz+d).$$

Being almost a modular form, this series is often called a quasi-modular form.

A invaluable property of the spaces of modular forms is that they are finite dimensional. In fact, their dimensions are often small. We list some properties of low weights, taken from [41].:

$$\begin{aligned}
\mathrm{M}_0(\Gamma(1)) &= \mathbb{C},\\
\mathrm{M}_2(\Gamma(1)) &= 0,\\
\mathrm{M}_{\mathrm{k\ odd}}(\Gamma(1)) &= 0,\\
\mathrm{M}_{-k}(\Gamma(1)) &= 0.
\end{aligned}$$

Further, the non-zero spaces are the complex span of combinations of the Eisenstein series E_4 and E_6 in the following way.

$$\begin{aligned}
\mathrm{M}_4(\Gamma(1)) &= \mathbb{C}\mathrm{E}_4,\\
\mathrm{M}_6(\Gamma(1)) &= \mathbb{C}\mathrm{E}_6,\\
\mathrm{M}_8(\Gamma(1)) &= \mathbb{C}\mathrm{E}_4^2,\\
\mathrm{M}_{10}(\Gamma(1)) &= \mathbb{C}\mathrm{E}_4\mathrm{E}_6,\\
\mathrm{M}_{14}(\Gamma(1)) &= \mathbb{C}\mathrm{E}_4^2\mathrm{E}_6.
\end{aligned}$$

In this way, the above ring of modular forms is generated by these two modular forms of weight 4 and 6.

The Dedekind eta function $\eta(z)$ will be important as we proceed. Although it is not a modular form, we will see that it can be used to construct both modular forms and covariant derivatives on modular forms. The eta function has the following definition and transformation rules under the modular transformations [75]

$$\eta(z) = q^{-1/24}\prod_{n=1}^{\infty} 1-q^n,$$
$$\eta(z+1) = e^{\pi i/12}\eta(z),$$
$$\eta(-1/z) = \sqrt{-iz}\eta(z).$$

where $q = e^{2\pi i z}$. Under the modular group, we see that this function transforms almost as a weight $1/2$ modular form. There is a relation between the eta function and the quasi-modular E_2. This can be seen by explicit calculation

$$\begin{aligned}\partial \log\eta(\tau) &= \partial\left\{\frac{\pi i\tau}{12} + \sum_{n=1}^{\infty}\log(1-q^n)\right\}\\ &= \frac{\pi i}{12}\left\{1-24\sum_{n=1}^{\infty}\frac{nq^n}{1-q^n}\right\}\\ &= \frac{\pi i}{12}\mathrm{E}_2.\end{aligned}$$

From the perspective offered by this relation, the anomalous transformation properties of E_2 is simply a result of the product rule applied to $\eta(\gamma(z))$.

2.6.2 Eisenstein series on Γ_{T}

We want to discuss the Eisenstein series at level 2 and compute the q-expansion of the series corresponding to the Γ_{T} subgroup. Similar to the Eisenstein seres for the full modular group, the level 2 series are defined as a sum over a lattice. They are now defined [41] by summing over a sublattice as opposed to the whole $\mathbb{Z}\times\mathbb{Z}$ as we did for the full modular group. Let (a,b) be some combination of $(0,1)$. Then the Eisenstein series for the level 2 subgroups are defined [41] by

$$\mathrm{G}_k^{(a,b)} = \sum_{\substack{m,n\\ m=a\mathrm{mod}2\\ n=b\mathrm{mod}2}} \frac{1}{(m+n\tau)^k}.$$

Note that $\mathrm{G}_k^{(a,b)}(\tau) = (-1)^k\mathrm{G}_k^{(a,b)}(\tau)$, so k must be an even integer. Note also that the series (0,0) reproduces the full Eisenstein series up to an overall factor

$$\mathrm{G}_k^{(0,0)} = \sum_{m,n\in\mathbb{Z}}\frac{1}{(2m+2n\tau)^k} = 2^{-k}\mathrm{G}_k.$$

We will compute the q-expansion of this series also, as it is very similar to the $(1,0)$ series. First we will see that these series in fact corresponds to modular forms. Let us consider the transformations of $G_k^{(1,0)}$ under ST^2S:

$$\begin{aligned}G_k^{(1,0)} = \sum_{m,n} \frac{1}{[2m+1+2n\tau]^k} \to \sum_{m,n} \frac{1}{\left[2m+1+2n\frac{1}{2-1/\tau}\right]^k} \\ = (2\tau-1)^k \sum_{m,n} \frac{1}{[-(2m+1)+2(2m+n+1)\tau]^k} \\ = (2\tau-1)^k G_k^{(1,0)}.\end{aligned}$$

According to our previous discussions, this is the factor of automorphy of a weight k modular form on Γ_T. Under T the series transforms

$$G_k^{(1,0)} \to \sum_{m,n} \frac{1}{[2(m+n)+1+2n\tau]^k} = G_k^{(1,0)}.$$

Hence, this series corresponds to weight k modular forms for Γ_T when it converges. Likewise, we can verify that the $(0,1)$ series corresponds to the Γ_R subgroup:

$$STS : G_k^{(0,1)} = \sum_{m,n} \frac{1}{[2m+(2n+1)\tau]^k} \to (\tau-1)^k G_k^{(0,1)},$$

$$T^2 : G_k^{(0,1)} \to \sum_{m,n} \frac{1}{[2\{m+2n+1\}+(2n+1)\tau]^k} = G_k^{(0,1)}.$$

Similar manipulations show that the $(1,1)$ series corresponds to modular forms for the Γ_S group.

We now focus on the Γ_T subgroup. To understand the ring of quasi-modular forms on this group, we will need the q-expansion of the corresponding Eisenstein series. First though, we derive some identities we will need. We will follow a procedure similar to what it done in Serre's Arithmetic [72] for the Eisenstein series on the full modular group. We begin with the the identity

$$\sum_{m\in\mathbb{Z}} \frac{1}{z+m} = \pi\cot(\pi z).$$

We will see a sketch of a proof of this later. Now in addition we have the identity for the cotangent [72]

$$\pi\cot(\pi z) = i\pi - 2\pi i \sum_{n=0}^{\infty} q^n.$$

If we do a sum over odd or even integers, we in stead end up with the relations

$$\sum_{m\in\mathbb{Z}}\frac{1}{z+(2m+1)}=\frac{1}{2}\pi\cot\left(\pi\frac{z+1}{2}\right)=\frac{\pi i}{2}-\pi i\sum_{n=0}^{\infty}(-1)^n q^{n/2},$$

$$\sum_{m\in\mathbb{Z}}\frac{1}{z+(2m)}=\frac{1}{2}\pi\cot\left(\pi\frac{z}{2}\right)=\frac{\pi i}{2}-\pi i\sum_{n=0}^{\infty}q^{n/2}.$$

To end up with something that looks more like a part of the Eisenstein series, we should differentiate these relations (k-1) times. Recalling that $\partial_z=(2\pi i q)\partial_q$ we get the identities

$$\sum_{m\in\mathbb{Z}}\frac{1}{(z+2m+1)^k}=\frac{(-2\pi i)^k}{(k-1)!2^k}\sum_{n=0}^{\infty}(-1)^n n^{k-1}q^{n/2}, \tag{2.14}$$

$$\sum_{m\in\mathbb{Z}}\frac{1}{(z+2m)^k}=\frac{(-2\pi i)^k}{(k-1)!2^k}\sum_{n=0}^{\infty}n^{k-1}q^{n/2}. \tag{2.15}$$

With these identities we can calculate the q-expansions. As promised we will compute the q-expansion of both the $(1,0)$ and the $(0,0)$ series. These Eisenstein series can be written

$$\mathrm{G}_k^{(a,0)}(z)=\sum_{\substack{m=a \bmod 2\\ m\neq 0}}m^{-k}+2\sum_{\substack{n=1\\ n=0 \bmod 2}}^{\infty}\sum_{m=a \bmod 2}(m+nz)^{-k}.$$

We see that our derived identities appear as the last sum over m in this expression. It is then simply a matter of inserting the right identity into the right Eisenstein series to get a q-expansion. First, however, we consider the $\tau\to i\infty$ limit. We first consider the case of even numbers $a=0$. Splitting the sum over positive and negative integers we get the normalization

$$\sum_{\substack{m\in\mathbb{Z}\\ m=0 \bmod 2}}\frac{1}{m^k}=2^{-k}\left(\sum_{m=1}^{\infty}\frac{1}{m^k}+(-1)^k\sum_{m=1}^{\infty}\frac{1}{m^k}\right)=2^{1-k}\zeta(k)$$

for even values of k. The difference from the Eisenstein series on the full modular group in the additional factor of 2^{-k}. Similarly we can find the normalization when the sum is taken over odd integers. This is ever so slightly more involved. Again we can split the sum into two parts

$$\sum_{\substack{m\in\mathbb{Z}\\ m=1 \bmod 2}}\frac{1}{m^k}=\left\{\sum_{n=0}^{\infty}\frac{1}{(n+1/2)^k}+(-1)^k\sum_{n=1}^{\infty}\frac{1}{(n-1/2)^k}\right\}.$$

The first of these sums can be recognized as the Hurwitz zeta function. The second term is too, if we rewrite it so that the sum starts at $n=0$. This is easily achieved by shifting the value of n by one. For even values of k, the normalization can be written

$$\sum_{\substack{m\in\mathbb{Z}\\ m=1 \bmod 2}} \frac{1}{m^k} = 2^{1-k}\zeta_{\rm H}(k,1/2) = (2-2^{1-k})\zeta(k)$$

where we used a relation between the Hurwitz zeta at $1/2$ and the Riemann zeta. We see that the Eisenstein series over odd integers is normalized with the Hurwitz zeta in stead of the Riemann zeta.

With these values at infinity, we can finally write down the q-expansion for the Eisenstein series. By the identities (2.14) and (2.15) we have

$$\mathrm{E}_k^{(1,0)} = 1 + \frac{(-2\pi i)^k}{(k-1)!\zeta(k)(2^k-1)} \sum_{\ell=0}^{\infty} \frac{(-1)^\ell \ell^{k-1} q^\ell}{1-q^\ell}, \tag{2.16}$$

$$\mathrm{E}_k^{(0,0)} = 2^{-k}\mathrm{E}_k = 2^{-k}\left\{1 + \frac{(-2\pi i)^k}{(k-1)!\zeta(k)} \sum_{\ell=0}^{\infty} \frac{\ell^{k-1} q^\ell}{1-q^\ell}\right\}. \tag{2.17}$$

This also proves the previously states q-expansion of the Eisenstein series for the full modular group. Notice that in the beginning of our derivations we interchanged the order of the double sum, so these expressions hold only for weight $k \geq 4$ when the sums are uniformly convergent. The second Eisenstein series is again quasi-modular [33].

2.7 Weight two modular forms on $\Gamma_{\rm X}$

Of the subgroups, it is the $\Gamma_{\rm T}$ subgroup that has had most of our attention. While this will still be true for the rest of this chapter, we will need some information regarding the remaining two subgroups as we proceed. Specifically, in order to construct modular beta functions, we will need the weight 2 modular forms on the subgroups. These can be obtained from modular invariant functions by differentiation.

Invariant functions for all three subgroups can be found in for example [68]. For the three subgroups $\Gamma_{\rm X}$, $\rm X = T, S, R$ these functions are

$$\begin{aligned} f_{\rm T} &= \frac{\lambda-1}{\lambda^2}, \\ f_{\rm S} &= \lambda(1-\lambda), \\ f_{\rm R} &= \frac{-\lambda}{(1-\lambda)^2}. \end{aligned}$$

where λ is the modular lambda function we met above. The functions $\partial \log f_{\mathrm{X}}$ will be modular of weight 2 as the derivative adds a covariant index. By using well known q-expansions of the modular lambda and the eta function, one can verify the equalities

$$\begin{aligned}
\partial \log f_{\mathrm{T}} &= -24\partial \log \frac{\eta(2\tau)}{\eta(\tau)}, \\
\partial \log f_{\mathrm{S}} &= 24\partial \log \frac{\eta(2\tau)\eta(\tau/2)}{\eta^2(\tau)}, \\
\partial \log f_{\mathrm{R}} &= -24\partial \log \frac{\eta(\tau/2)}{\eta(\tau)}.
\end{aligned}$$

We denote the respective eta quotients simply by η_{X}. As usual we would like to normalize the modular forms so that the q-expansions start at unity. The properly normalized weight 2 forms we denote $\mathrm{E}_2^{\mathrm{X}} = \mathrm{N}_{\mathrm{X}}\partial \varphi_{\mathrm{X}}$ where $\varphi_{\mathrm{X}} = \log \eta_{\mathrm{X}}$. Here the G-conjugacy of Γ_{T} and Γ_{R} is evident. By expanding in q, we can find the proper normalization N_{X}, yielding

$$\begin{aligned}
\mathrm{E}_2^{\mathrm{T}} &= \frac{12}{\pi i}\partial \varphi_{\mathrm{T}}, \\
\mathrm{E}_2^{\mathrm{S}} &= \frac{24}{\pi i}\partial \varphi_{\mathrm{S}}, \\
\mathrm{E}_2^{\mathrm{R}} &= -\frac{24}{\pi i}\partial \varphi_{\mathrm{R}}.
\end{aligned}$$

We would like the q-expansions of these forms, which we can find by using the definition of the Dedekind eta function. For example, to express $\mathrm{E}_2^{\mathrm{T}}$ as a q-expansion, we first note that

$$\begin{aligned}
\frac{\eta(2\tau)}{\eta(\tau)} &= \frac{q^{1/12}\prod_{n=1}^{\infty}(1-q^{2n})}{q^{1/24}\prod_{n=1}^{\infty}(1-q^{n})} \\
&= q^{1/24}\prod_{n=1}^{\infty}\frac{1-q^{2n}}{1-q^{n}} \\
&= q^{1/24}\prod_{n=1}^{\infty}(1+q^{n}).
\end{aligned}$$

When taking a logarithm of this expression, the infinite product is turned into a infinite sum, which is exactly the q-expansion we are looking for, after differentiation

$$\mathrm{E}_2^{\mathrm{T}} = 1 + 24\sum_{n=1}^{\infty}\frac{nq^n}{1+q^n}. \tag{2.18}$$

Note that since we have derived this from a Γ_T invariant function, this second "Eisenstein series" E_2^T is a weight 2 modular form. As in the case of the full modular group, the "naive" Eisenstein series $\mathrm{E}_2^{(1,0)}$ defined by the lattice sum is quasi-modular [33].

The q-expansions of the other weight 2 forms can be found in a similar manner.

$$\mathrm{E}_2^\mathrm{S} = 1 - 24\sum_{n=0}^{\infty}\frac{(2n+1)q^{\frac{2n+1}{2}}}{1+q^{\frac{2n+1}{2}}}, \tag{2.19}$$

$$\mathrm{E}_2^\mathrm{R} = 1 + 12\sum_{n=0}^{\infty}\frac{nq^{n/2}}{1+q^{n/2}}. \tag{2.20}$$

At this point the notation has become quite messy. For the sake of simplicity, we will from now on use a new notation for the Eisenstein series both for the full modular group and the subgroup. We will for the full modular group denote the modular Eisenstein series by E_k. For the Γ_T subgroup we will denote the Eisenstein series with $k \geq 4$, i.e. (2.16), by E_k^T, and the modular weight 2 form on Γ_X discussed in this chapter we denote $\mathcal{E}_\mathrm{X}$. The quasi-modular forms for the full modular group we denote $\mathcal{H}_2$ and for the Γ_X subgroup $\mathcal{H}_2^\mathrm{X}$.

2.8 Holomorphic connections generated by η-quotients

As we have mentioned, we are studying the theory of modular forms from a geometric perspective where everything is tailored to fit the modular group. This is reminiscent of the Erlangen program by Klein, where the study of geometry was connected with invariant objects under certain group transformations. Here we are considering objects that behave like tensors under the modular group or its subgroups. In this section we consider connections and covariant derivatives in the same spirit.

To define a notion of differentiation on a fiber bundle, we needed a connection. Recall that a connection A is a Lie algebra valued 1-form, where the Lie algebra generates the endomorphisms of the fibers. If the fiber basis is e_i, the connection coefficients will satisfy $\mathrm{D}_\mu e_i = (\mathrm{A}_\mu)_i^j e_j$. Under a transformation $e_i \to \mathrm{M}_i^j e_j$ one can easily show that the connection must transform as

$$(\tilde{\mathrm{A}}_\mu)_i^j = \mathrm{M}_r^j(\partial_\mu \mathrm{M}_i^r) + \mathrm{M}_r^j \mathrm{M}_i^\ell (\mathrm{A}_\mu)_\ell^r.$$

This transformation law can be derived by identifying $\mathrm{D}_\mu e_i$ with the component of a dual vector that must transform to $\mathrm{M}_i^r \tilde{\mathrm{D}}_\mu e_r$, where the new covariant derivative contains the new connection. In the case on the tensor bundles, we recall that the transition functions M^i_j are inherited from the coordinate transformations on the base manifold. In particular, when we are dealing with holomorphic geometry on a surface, this transformation law simplifies substantially. We have $\mathrm{M} = \partial z/\partial z'$, and the connection transforms as

$$\tilde{\mathrm{A}} = \frac{\partial z}{\partial z'}\mathrm{A} + \frac{\partial z'}{\partial z}\frac{\partial^2 z}{\partial z'^2}.$$

The method we will use to construct connections is inspired by remarks made in [32]. It turns out there is a very natural way to construct connections. Let the function $f(z,\overline{z})$ be the component of the bidegree $(1,a)$-form on $\mathfrak{H}$ where a is 0 or 1, i.e.

$$f = f(z,\overline{z})dz \wedge d\overline{z}^{\wedge a}.$$

As this tensor is invariant under coordinate transformations $z \to z'$, the component must satisfy the standard tensorial transformation law

$$f(z',\overline{z}') = f(z,\overline{z})\frac{\partial z}{\partial z'}\left(\frac{\partial \overline{z}}{\partial \overline{z}'}\right)^a.$$

Given such a form, we can construct a connection by taking the logarithmic derivative

$$\mathrm{A} = \partial \log(f).$$

By the above transformation law for f one can easily verify that A in fact transforms as a connection

$$\begin{aligned}
\mathrm{A}' &= \partial' \log\left\{f(z,\overline{z})\frac{\partial z}{\partial z'}\left(\frac{\partial \overline{z}}{\partial \overline{z}'}\right)^a\right\} \\
&= \frac{\partial z}{\partial z'}\partial\log(f) + \frac{\partial z'}{\partial z}\frac{\partial^2 z}{\partial z'^2} + a\partial' \log\frac{\partial \overline{z}}{\partial \overline{z}'} \\
&= \frac{\partial z}{\partial z'}\mathrm{A} + \frac{\partial z'}{\partial z}\frac{\partial^2 z}{\partial z'^2}.
\end{aligned}$$

In this way, we can construct connections by finding candidate functions $f(z)$ transforming as (1,a)-forms.We will from now set $a = 0$ and deal only with holomorphic f. The construction of such a connection is valuable if we want to understand interrelations between modular forms of different weights from a tensorial perspective.

Recall from our discussions on fiber bundles that a covariant derivative $\mathscr{D}$ can be seen as a map

$$\mathscr{D} : \Gamma(\mathcal{E}) \to \Gamma(\mathcal{E}) \otimes \Gamma(\mathrm{T}^*\mathrm{M})$$

where $\mathcal{E}$ is a vector bundle over a manifold M. Let us now consider a space $\mathcal{A}_k$ of functions that transform as (k,0)-tensors under a certain group G but not necessarily all other transformations. If we can find a connection A under G we can construct a covariant derivative $\mathcal{D}$ that from the above arguments take $\mathcal{A}_k \to \mathcal{A}_{k+1}$. In the case of modular forms we have $\mathcal{A}_k = \Gamma(\mathcal{L}_*^{\otimes k}) = \mathrm{M}_{2k}(\Gamma)$, and the connection A can be constructed by finding a weight 2 modular form f for Γ. We first consider the full modular group. Immediately we run into problems, since there are no weight 2 modular forms on this group. However, we can solve this problem by slightly easing the restrictions on f.

Consider not a modular form, but rather a "projective" version that transforms as

$$f \to \alpha \frac{\partial \gamma(z)}{\partial z} f$$

where γ as usual is a Γ transformation, and α is some overall complex factor. One can easily convince oneself that this constant factor will have no effect on the transformation of A since it is killed by the combination ∂ log. This allows us to find candidate functions f also for the full modular group, even though it has no proper weight 2 modular forms. Another possible way around this issue that we will not discuss is to allow for anti-holomorphic in the tensor f.

As an example, consider $f = \eta^a(z)$ for some a to be determined. Using the above modular transformation rules for the Dedekind eta function, we have

$$\mathrm{T} : \partial \log \eta^a(z) \to \partial \log[e^{ia\pi/12}\eta(z)] = \partial \log \eta^a(z),$$

$$\mathrm{S} : \partial \log \eta^a(z) \to z^2 \partial \log[(-iz)^{a/2}\eta(z)] = z^2 \partial \log[z^{a/2}\eta(z)].$$

Hence, if we chose $a = 4$ this will effectively transform as a 1-form because of the logarithm and derivative. Hence, we can consider the connection

$$\mathrm{A} = \partial \log(\eta^4(z)).$$

We can also find a connection for the subgroups Γ_{X}. Here the automorphic factor is different, as we discussed in the above chapters. Consider for the moment the eta function with scaled coordinate $\eta(nz)$. Under the two generating transformations we have

$$\mathrm{T} : \eta(nz) \to e^{n\pi i/12}\eta(nz),$$

$$\mathrm{S} : \eta(nz) \to \eta(-n/z) = \eta(-1/(z/n)) = (-iz/n)^{1/2}\eta(z/n).$$

Note that since the argument nz is turned into a fraction, these scaled eta functions can not be used to make connections for the full modular group. However, for the subgroups where one of the generator is built from two S transformations, there still is a chance. For the Γ_T group generated by translations and ST^2S, we have

$$ST^2S : \eta(nz) \to e^{2\pi i/12n}(1-2z)^{1/2}\eta(nz).$$

Surprisingly, the only n dependence is in the exponential. Hence $\partial \log \eta^4(nz)$ could be a candidate connection. Slightly more generally, we can consider

$$A^{(n)} = \partial \log \frac{\eta^8(nz)}{\eta^4(z)}.$$

We will see that the $n = 2$ connection is related to the Eisenstein series on Γ_T, which will enable us to find more interesting structure on the corresponding ring of modular forms.

2.9 Differential structure on the ring of quasi-modular $\Gamma(1)$ forms

With the quasi-modular weight 2 Eisenstein series we can extend the modular ring $\oplus_k M_k(\Gamma(1))$ to the ring of quasi-modular forms on the full modular group, generated by $(\mathcal{H}_2, E_4, E_6)$. We will here derive the famous Ramanujan identities which can be seen as relations defining a differential structure on this quasi-modular ring. We will formulate this geometrically.

Recall that we view modular forms of weight $2k$ as sections of the line bundle $\mathcal{L}_*^{\otimes k}$. We would like to find a way to go between forms of different weight, similarly to how the exterior derivative maps differential p-forms to (p+1)-forms et cetera. The analogue sequence in this case would be a map $\mathcal{D}$ that in someway represents going from one bundle $\mathcal{L}_*^{\otimes k}$ to another $\mathcal{L}_*^{\otimes k+1}$. First recall some facts regarding differential forms. We can write the action of the exterior derivative as

$$\omega = \omega_\mu(x)dx^\mu \to [d\omega_\mu(x)] \wedge dx^\mu.$$

Under a coordinate change $x \to \tilde{x}$, say by a group action, this changes to

$$d[\omega_\mu(x)\frac{\partial x^\mu}{\partial \tilde{x}^\rho}] \wedge d\tilde{x}^\rho.$$

However, after using the product rule, the second derivative terms will vanish since they will be contracted with the antisymmetric wedge product. This is why the exterior derivative maps forms to forms, while the derivative of a generic tensor will not be tensorial. This automatic removal of the double derivative term is the luxury we do not have in the case of $\mathcal{L}_*^{\otimes k}$. This motivates the introduction of a covariant derivative.

Based on the above discussions of connection, we can construct covariant derivatives $\mathscr{D} = \partial - k\mathrm{A}$, where $2k$ is the weight of the modular form. This can be seen as a map generating the chain

$$\dots \xrightarrow{\mathscr{D}} \mathrm{M}_{2k}(\Gamma) \xrightarrow{\mathscr{D}} \mathrm{M}_{2k+2}(\Gamma) \xrightarrow{\mathscr{D}} \mathrm{M}_{2k+4}(\Gamma) \xrightarrow{\mathscr{D}} \dots$$

For the full modular group, we saw that one possible connection was

$$\mathrm{A} = \partial \log(\eta^4(z)).$$

As we discussed in earlier sections, the second quasi-modular Eisenstein series $\mathcal{H}_2$ can in a similar way be written $\mathcal{H}_2 = (12/\pi i)\log\eta(z)$. By working out the normalization factors, we have the relation

$$\mathrm{A} = \frac{2\pi i}{6}\mathcal{H}_2.$$

When the resulting covariant derivative acts on $(0,k)$ tensors, it takes the form $\mathscr{D} = \partial - k\mathrm{A}$. In conclusion, we have found a covariant derivative

$$\mathscr{D} = \partial - \frac{\pi i}{6}(2k)\mathcal{H}_2.$$

Here $2k$ is the modular weight of the holomorphic $(0,k)$-tensor. This operator is the same as the one mentioned in the exercises in chapter 9 of [44], up to an overall scaling by $2\pi i$. It is also the operator referred to as the Serre derivative in [12].

This covariant derivative can be used to construct the Ramanujan identities which relates modular forms and their derivatives. In a geometric language, they are simply tensor identities [50]. This is a consequence of the relation between the holomorphic connections and the quasi-modular Eisenstein series. Consider the tensor $\omega = \partial\mathrm{A} - \frac{1}{2}\mathrm{A}^2$, where A is the above connection [50]. This can be shown to transform as a weight 4 modular form [25] but is not in general a (0,2) tensor. In a similar way, we only know that η^4 effectively transforms as a modular form of weight 2 under modular transformations, not general coordinate transformations. In this sense, this is geometry tailored to fit the modular group. Since $\mathrm{M}_4(\Gamma(1))$ is one dimensional, ω and E_4 must be proportional. By considering the q-expansions one can see that

$$\omega = \frac{\pi^2}{18}\mathrm{E}_4.$$

Similarly $\mathscr{D}\omega \in \mathrm{M}_6(\Gamma(1))$ and $\mathscr{D}^2\omega \in \mathrm{M}_8(\Gamma(1))$. By again considering q-expansions one can verify that

$$\mathscr{D}\omega = -\frac{27}{\pi^3 i}\mathrm{E}_6,$$

$$\mathscr{D}^2\omega = -\pi i \mathrm{E}_4^2.$$

These tensor identities can equivalently be written in terms of the logarithmic derivative $\mathrm{D} = q\partial_q$ as the differential equations

$$\mathrm{D}\mathcal{H}_2 = \frac{\mathcal{H}_2^2 - \mathrm{E}_4}{12}, \tag{2.21}$$

$$\mathrm{D}\mathrm{E}_4 = \frac{\mathcal{H}_2\mathrm{E}_4 - \mathrm{E}_6}{3}, \tag{2.22}$$

$$\mathrm{D}\mathrm{E}_6 = \frac{\mathcal{H}_2\mathrm{E}_6 - \mathrm{E}_4^2}{2}. \tag{2.23}$$

These three differential equations are the well-known Ramanujan identities. They can be seen as a proof of the fact that the quasi-modular ring is closed under the differential operator $\mathrm{D} = q\partial_q$. One says that the ring has a differential structure. For an alternate but similar proof see [12]. From these three identities, higher order differential equations can be generated by applying D. For example, by differentiating the identity for $\mathrm{D}\mathcal{H}_2$ and using the other Ramanujan identities, we get

$$12\mathrm{D}^2\mathcal{H}_2 = \frac{1}{6}(\mathcal{H}_2^3 - 3\mathcal{H}_2\mathrm{E}_4 + 2\mathrm{E}_6).$$

In principle one can keep going to construct infinitely many differential equations. Note that if we look at these results backwards we have quite a non-trivial result: we know of infinitely many differential equations which the Eisenstein series solve.

2.10 Differential structure on the ring of quasi-modular Γ_T forms

In the case of the full modular group, we saw that the ring of modular forms was generated by E_4 and E_6. By extending this to the quasi-modular case, we saw that the ring of quasi-modular forms was generated by $(\mathcal{H}_2, \mathrm{E}_4, \mathrm{E}_6)$. This ring was also equipped with a further structure, making it into a differential ring. The Ramanujan identities defined this differentiable structure. We want to understand these ring structures for the subgroup Γ_T.

In contrast to the level 1 case, we now have a modular form of weight 2. The ring in interest is the direct sum $\mathrm{M}(\Gamma_\mathrm{T}) = \oplus_k \mathrm{M}_k(\Gamma_\mathrm{T})$ which is generated by $(\mathcal{E}_\mathrm{T}, \mathrm{E}_4^\mathrm{T})$ [84]. By also including the quasi-modular weight 2 form, we have the quasi-modular ring generated by $(\mathcal{H}_2^\mathrm{T}, \mathcal{E}_\mathrm{T}, \mathrm{E}_4^\mathrm{T})$. We will show that this is a differential ring in a similar manner to what we did for the full modular group.

We start with constructing a connection. In the level 1 case, the connection was proportional to the quasi-modular Eisenstein series $\mathcal{H}_2$. We may therefore expect that a connection proportional to the quasi-modular $\mathcal{H}_2^{\mathrm{T}}$ can be constructed as well. Recall that the eta function is a weight half modular form. From above discussions we know that on thus subgroup we have a connection

$$\mathrm{A}^{(2)} \equiv \mathrm{A} = \partial \log \frac{\eta^8(2\tau)}{\eta^4(\tau)}.$$

This can be expressed in terms of the quasi-modular $\mathcal{H}_2^{\mathrm{T}}$. By considering the q-expansions of the Eisenstein series we can verify that $3\mathcal{H}_2^{\mathrm{T}} = 4\mathcal{H}_2(2\tau) - \mathcal{H}_2(\tau)$ [33]. By using the previously derived relation between $\mathcal{H}_2$ and the eta function, we can write

$$\mathcal{H}_2^{\mathrm{T}} = \frac{1}{\pi i} \partial \log \frac{\eta^8(2\tau)}{\eta^4(\tau)}.$$

Hence the connection satisfies $\mathrm{A} = \pi i \mathcal{H}_2^{\mathrm{T}}$, and the covariant derivative reads

$$\mathcal{D} = \partial - k\mathrm{A} = \partial - k\pi i \mathcal{H}_2^{\mathrm{T}}.$$

We can rescale this derivative by $2\pi i$ to get

$$q\partial_q - \frac{2k}{4}\mathcal{H}_2^{\mathrm{T}}$$

where again $2k$ is the modular weight. This is the operator considered in [33], achieved by deforming the connection we used in the level 1 case by the modular $\mathcal{E}_{\mathrm{T}}$. This is the natural connection on the modular curve $\mathfrak{H}/\Gamma_{\mathrm{T}}$ since we want to act on tensors invariant under only the subgroup Γ_{T}.

Recall that in the level 1 case, we derived the Ramanujan identities from tensor identities involving $\omega = \partial \mathrm{A} - \frac{1}{2}\mathrm{A}^2$. This was a modular form of weight 4. Consider again this tensor, now with the new connection $\mathrm{A} = \pi i \mathcal{H}_2^{\mathrm{T}}$. By comparing q-expansions to find normalization factors, we have

$$\omega = \frac{\pi^2}{2}\mathrm{E}_4^{\mathrm{T}}.$$

We let the covariant derivative act on this form. This should give a weight 6 modular form, which can be verified to be

$$\mathcal{D}\omega = -\pi^3 i \mathrm{E}_6^{\mathrm{T}} = -\pi^3 i \mathcal{E}_{\mathrm{T}} \mathrm{E}_4^{\mathrm{T}}.$$

However, in contrast to the level 1 case we now have a proper modular form of weight 2, namely $\mathcal{E}_{\mathrm{T}}$. The covariant derivative takes this to a weight 4 form:

$$\mathcal{D}\mathcal{E}_{\mathrm{T}} = -\pi i \mathrm{E}_4^{\mathrm{T}}.$$

These three tensor identities shows that the quasi-modular ring generated by $(\mathcal{H}_2^{\mathrm{T}}, \mathcal{E}_{\mathrm{T}}, \mathrm{E}_4^{\mathrm{T}})$ at level 2 is closed under differentiation. As in the level 1 case we can write these identities in terms of $\mathrm{D} = q\partial_q$ in the form

$$\begin{aligned} \mathrm{D}\mathcal{E}_{\mathrm{T}} &= \frac{\mathcal{H}_2^{\mathrm{T}}\mathcal{E}_{\mathrm{T}} - \mathrm{E}_4^{\mathrm{T}}}{2}, \\ \mathrm{D}\mathcal{H}_2^{\mathrm{T}} &= \frac{\mathcal{H}_2^{\mathrm{T}2} - \mathrm{E}_4^{\mathrm{T}}}{4}, \\ \mathrm{D}\mathrm{E}_4^{\mathrm{T}} &= \mathcal{H}_2^{\mathrm{T}}\mathrm{E}_4^{\mathrm{T}} - \mathcal{E}_{\mathrm{T}}\mathrm{E}_4^{\mathrm{T}}. \end{aligned}$$

Note that if we had chosen another connection, we would just get tensor identities, but by choosing the connections that are proportional to the Eisenstein series we get equations involving only Eisenstein series.

2.11 Elliptic functions and elliptic curves

This section considers the algebraic representation of the complex torus. We will see that there is a close connection between the theory of elliptic functions, Eisenstein series and elliptic curves. An elliptic curve over a field $\mathbb{K}$ is a cubic equation [86] of the type

$$y^2 = x^3 + \mathrm{A}x^2 + \mathrm{B}x + \mathrm{C}$$

where the constants $\mathrm{A}, \mathrm{B}, \mathrm{C}$ take values in $\mathbb{K}$. This type of equation is sometimes also called a Weierstrass equation. For technical reasons one also adds a single point at infinity, working in a projective version of the space. For example, in the case of the real line or the plane, the projective versions are the circle and the sphere respectively. We will see that these elliptic curves naturally appear from algebraic relations satisfied by canonical functions on the torus. We start with some general remarks on invariant functions for any group G.

2.11.1 Constructing G-invariant functions

From our time of birth we learn that sines and cosines are the fundamental trigonometric functions. However, there is a sense in which these are not the canonical periodic functions. The familiar trigonometric functions can be seen as circle functions, e.g. maps

$$f : \mathrm{S}^1 \to \mathbb{R}.$$

Functions on a quotient M/G can be seen as G-invariant functions on the cover M. In the case of a circle we have $\mathrm{S}^1 = \mathbb{R}/\mathbb{Z}$, which yields the notion of trigonometric functions we are more familiar with, namely periodic functions of a real variable. One could generalize this to n-dimensional tori $\mathbb{R}^n/\mathbb{Z} \times \ldots \times \mathbb{Z}$ where we could consider $\mathbb{Z}^{\times n}$-invariant functions on flat n-space. For a more general situation, consider the diagram

$$\begin{tikzcd}
\mathrm{M} \arrow[r, "f"] \arrow[d, "\pi"'] & \mathbb{C} \\
\mathrm{M}/\mathrm{G} \arrow[ur, "\tilde{f}"'] &
\end{tikzcd}$$

We define the function on the quotient simply by $\tilde{f} = f \circ \pi^{-1}$. Equivalently, given a function $\tilde{f}$ that takes in equivalence classes $[x]$ we have $f = \tilde{f} \circ \pi$. On some point $x \in \mathrm{M}$ this acts as $f(x) = \tilde{f}([x])$. However, for any other point $g(x)$ in the same G-orbit as x we must also have $f(g(x)) = \tilde{f}([x])$. Hence f in the above diagram must be a G-invariant function for $\tilde{f}$ to be well-defined. Given a projection, such a function defines the function on the quotient by $\tilde{f} = f \circ \pi^{-1}$.

We want a way to construct G-invariant functions. The brute-force "canonical" way is the following. Pick F as a function on M that is rapidly decreasing with $x \in \mathrm{M}$. For example $\mathrm{F}(x) = 1/x^n$ for some sufficiently large n. Given a G-action on M we consider functions of the form

$$f(x) = \sum_{g \in \mathrm{G}} \mathrm{F}(g(x)).$$

This is clearly invariant under G as any group transformation on M would simply lead to a relabeling of the group elements in the sum. In the case of functions on a circle, this reduces to a infinite sum over $\mathbb{Z}$ for some appropriate F.

2.11.2 Lattice sums and the Weierstrass cubic

Consider the complex torus $\mathbb{C}/\Lambda_{1,\tau}$ and the complex valued functions on this space. By the above line of thought we should consider lattice sums of converging functions [51] of the type

$$\wp_n(z;\tau) = \sum_{w \in \Lambda} (z+w)^{-n} = \sum_{m_1, m_2} (z + m_1 + m_2\tau)^{-n}. \tag{2.24}$$

For $n = 2$ this is very similar to a famous mathematical function, namely the Weierstrass $\wp$-functions. It has the definition [86]

$$\wp = \sum_{m,n} \left\{ \frac{1}{(z+m+n\tau)^2} - \frac{1}{(m+n\tau)^2} \right\},$$

and has a series expansion

$$\wp = \frac{1}{z^2} + 3\mathrm{G}_4 z^2 + 5\mathrm{G}_6 z^4 + \ldots$$

From our definition (2.24), it seems obvious that $\wp_2(z;\tau) = \wp(z;\tau) + \mathrm{G}_2(\tau)$, but because the second Eisenstein series is not uniformly convergent while the Weierstrass function as a whole is, we can not necessarily split the sum in $\wp$. However, we can compare power series order by order.

We will need the series expansion of the function $f(z) = (1+z)^{-a}$. The n'th derivative can be easily verified to be

$$f^{(n)}(z) = (-1)^n \frac{(a+n-1)!}{(a-1)!}(1+z)^{-1}.$$

Using well known relations between the binomial coefficient and the factorial, the series expansion takes the form

$$\frac{1}{(1+z)^a} = \sum_{k=0}^{\infty} \binom{-a}{k} z^k.$$

We write the $\wp_\ell$ function as

$$\wp_\ell = \sum_{m,n} \frac{1}{(z+m+n\tau)^\ell} = \frac{1}{z^\ell} + \sum_{m,n}{}' \frac{1}{(m+n\tau)^\ell} \frac{1}{\left(1+\frac{z}{m+n\tau}\right)^\ell},$$

where the prime indicates a sum over non-zero elements. Using the above series expansion we can express this torus function as the series

$$\wp_\ell = \frac{1}{z^\ell} + \sum_{k=0}^{\infty} \sum_{m,n}{}' \frac{1}{(m+n\tau)^{-\ell-k}} \binom{-\ell}{k} \left(\frac{z}{m+n\tau}\right)^k = \frac{1}{z^\ell} + \sum_{k=0}^{\infty} \mathrm{G}_{\ell+k}(\tau) \binom{-\ell}{k} z^k.$$

Here we see that the Eisenstein series have appeared. We recall from the discussion of the Eisenstein series that the G_n's are only non-zero for even values on n. In our case this means that $k+\ell = 2a$ for some integer a. The final form of the series expansion then reads

$$\wp_\ell = \frac{1}{z^\ell} + \sum_{\substack{a\in\mathbb{N} \\ a\geq \ell/2}} \mathrm{G}_{2a}(\tau) \binom{-\ell}{2a-\ell} z^{2a-\ell}.$$

We can now compare this to the similar series expansion of the Weierstrass function. The first few terms of these expansions are

$$\wp_2 = \frac{1}{z^2} + \mathrm{G}_2 + 3\mathrm{G}_4 z^2 + \ldots$$

Hence $\wp = \wp_2 - \mathrm{G}_2$ as one would naively believe. This subtraction is made so that the coefficients of the series expansion of the Weierstrass function consists only of proper modular forms, not the quasi-modular G_2. We set $x_w = \wp$ and $y_w = \wp'$. Using this series expansion, we can easily verify the algebraic equation

$$y_w{}^2 = 4x_w{}^3 - 60\mathrm{G}_4(\tau)x_w - 140\mathrm{G}_6(\tau). \tag{2.25}$$

We denote this elliptic curve $\mathrm{E}_{1,\tau}$. There is in other words a map taking $z \to \mathrm{E}_{1,\tau}$, which is lattice invariant since both the Weierstrass function and its derivative is [86]. We can therefore view the map as a map from the torus to the complex elliptic curve. This can be shown to be a isomorphism [86], which is why we so far have used the names complex torus and complex elliptic curve interchangeably.

Part II

Quantum field theory, universality and duality

3

A geometric approach to quantum field theories

The goal of this chapter is to present a clear and geometric formulation of field theories. We will focus on formal aspects of these theories in this chapter, while the next discusses the application of quantum field theory as a tool for studying universal features of many-particle systems.

Field theory is a subject not lacking in good literature. For a good and concise introduction see M. Maggiores book [57] or the book of Peskin and Schoeder [67]. The books of Zinn-Justin [91], Di Francesco et all [28] and Altland and Simons [2] are also good introductions if one wants field theory presented in contexts outside scattering problems and particle physics. These are particularly relevant for applications in critical phenomena, renormalization and condensed matter. There are also many good mathematically oriented books like K. Hori et als Mirror Symmetry [37], de Faria and de Melos book [19] or the two-volume wonder [20] of Witten, Freed and company. This chapter is inspired by the three latter references.

3.1 Geometric structures, fields and actions

Quantum field theory seeks to integrate over spaces of certain geometric objects on a manifold M [37]. The classical counterpart is interested in a fixed subset of such objects. The objects are called fields and can often be seen as sections of some bundle over M. There are three pieces of data needed to define a classical field theory:

$$(\mathrm{M}, \mathcal{F}_{\mathrm{M}}, \mathrm{S}),$$

where $\mathcal{F}_{\mathrm{M}}$ is a space of fields over a spacetime manifold M and S is a real valued function on $\mathcal{F}_{\mathrm{M}}$ called the action. The manifold M can be a topological manifold with no additional structure, or a highly equipped smooth manifold with additional geometric structures. The type of field theory one ends up with depends on the category to which M belongs (topological, Riemannian, conformal etc). Generically the situation is the following [11]. A spacetime manifold M is a cobordism[1] between two spatial manifolds of dimension $d-1$. Formally the spacetime cobordisms constitute the morphisms in the corresponding spacetime category[11]. If we need to refer to the spacetime category in a general setting we will denote it $dCob_*$, where $*$ indicates some geometric structure.

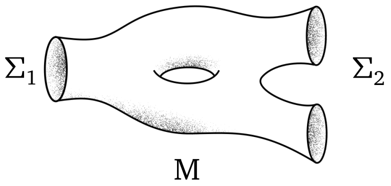


Figure 3.1: Spacetime as a cobordism between two spatial slices. M can be seen as a geometric analogue of time passing in a general theory.

The different spacetime categories divides the classical field theories into large classes. We will meet field theories in different classes as we go along.

- When no geometric structure is put on M other than an orientation and a smooth structure, the corresponding field theory is called topological. Here the morphisms are (homeomorphism classes of) smooth d-manifolds whose boundaries are oriented $(d-1)$-manifolds. A theory is also called topological if the manifold *is* equipped with geometric structures but the theory is independent of these.

- When spacetime belongs to the category of Riemannian cobordisms, the field theory is called Euclidian. Here space is a $(d-1)$-dimensional Riemannian manifold, and the spacetime morphism (an isometry class of) a d-dimensional manifold M. When the signature of the metric on M is Minkowskian the theory is called relativistic.

- When the theory only depends on the conformal equivalence class of metrics on M the theory is said to be a conformal field theory. As we have seen, this corresponds to a complex structure in the case of surfaces.

[1] Recall that a cobordism is simply a manifold that starts at a boundary Σ_1 and ends at another boundary Σ_2. These boundaries are said to be cobordant (= "jointly bound").

Within each of these large classes the field theories are further refined according to their field content and action. As mentioned, the fields in the theory are in a generic situation sections of some fiber bundle $\pi : \mathcal{E} \to M$. For example, trivial real or complex line bundles correspond to real of complex scalar fields, vector bundles to vector fields, tensor bundles to tensor fields and so on. The connection 1-forms on a vector bundle can also be considered as a field, which is the case in gauge theories. The space of field configurations $\mathcal{F}_M$ is in most cases associated with the space of sections $\Gamma(\mathcal{E})$, but may in some cases be constructed from these sections by suitable identifications. For example, in Yang-Mills theory the relevant field space is a quotient by the gauge group.

The last data needed for a classical field theory is the action. Let $\mathcal{L} : \mathcal{F}_M \to \Omega^n(M)$, the Lagrangian, be a map from the fields to a top dimensional form on M. Integrating this yields the action

$$S = \int_M \mathcal{L},$$

as a real valued function on the field space. This action is the fundamental quantity is a classical field theory. The classical fields satisfy the variational principle $\delta S = 0$, while in a quantum theory fluctuations must be taken into account.

Let of briefly discuss the Lagrangian in more detail, and in particular so-called boundary terms. Typically the Lagrangian of a theory is written $\mathcal{L}(\varphi) d\mathrm{vol}_g$ if M is equipped with a Riemannian/Lorentzian structure. In any case, one assumes that the Lagrangian is local in the sense that it is an expression in the fields expressed at a single point in M. However, the expression for the action should not be dependent on the arbitrary choice of local trivialization if the fields are obtained from some bundle. In some generic situation of a fiber bundle over M with structure group G this translates into writing only G-invariant local expressions in the Lagrangian.

Note that since all top dimensional forms are closed we have $d\mathcal{L} = 0$ and the Lagrangian defines a cohomology class on M. The addition of an exact form $\mathcal{L} \to \mathcal{L} + d\omega$ induces

$$S \to S + \int_{\partial M} \omega$$

by Stokes theorem. Since the Lagrangian, and hence also ω, is a local expression in the fields, this boundary term will vanish if the fields vanish at ∂M or if the manifold is boundaryless. This is often the case in particle physics applications, where one can imagine spacetime to be a sphere with infinite radius, and any local process should not be affected by events with large spatial separation.

However, in a general setting one should take some care regarding these boundary terms.

In the remainder of this section we will be dealing with spacetime manifolds M in the Riemannian category. The Lagrangian can then be written $\mathcal{L}(\varphi)dvol_g$. When the relevant field bundle is a vector bundle with an inner product in each fiber space, one often constructs an Lagrangian of the form $\langle\varphi,\mathrm{Q}\varphi\rangle$ where φ are the fields and Q is a operator constructed (or rather guessed or postulated) based on symmetry or geometry. The action then takes the form

$$\mathrm{S}(\varphi)=\int_{\mathrm{M}} d\mathrm{vol}_g\,\langle\varphi,\mathrm{Q}\varphi\rangle\,.$$

We write the vector space fibers V, which we take to be of finite dimension. In a local trivialization the sections (fields) are indexed as φ^a where a runs over the vector space dimensions (the rank rk($\mathcal{E}$) of the bundle) . The covariant derivative is as usual of the form

$$\mathrm{D}_\mu\varphi^a=\partial_\mu\varphi^a+(\mathrm{A}_\mu)^a_b\varphi^b$$

for a Lie algebra valued connection 1-form A. The classical requirement that the variation of the action vanishes can in a local trivialization be turned into a differential equation for the field components. Consider the variation of the action

$$\delta\mathrm{S}=\int d\mathrm{vol}_g\delta\mathcal{L}=\int d\mathrm{vol}_g\left\{\frac{\partial\mathcal{L}}{\partial\varphi^a}\delta\varphi^a+\frac{\partial\mathcal{L}}{\partial(\mathrm{D}_\mu\varphi^a)}\mathrm{D}_\mu\delta\varphi^a\right\}.$$

We will assume that if M has a boundary $\partial\mathrm{M}$, the field variations vanish there. By using the product rule and Stokes theorem we can write the variation of the action as

$$\begin{aligned}\delta\mathrm{S}&=\int_{\mathrm{M}} d\mathrm{vol}_g\left\{\frac{\partial\mathcal{L}}{\partial\varphi^a}\delta\varphi^a-\mathrm{D}_\mu\frac{\partial\mathcal{L}}{\partial(\mathrm{D}_\mu\varphi^a)}\delta\varphi^a\right\}+\int_{\mathrm{M}} d\mathrm{vol}_g\mathrm{D}_\mu\left[\frac{\partial\mathcal{L}}{\partial(\mathrm{D}_\mu\varphi^a)}\delta\varphi^a\right]\\&=\int d\mathrm{vol}_g\left\{\frac{\partial\mathcal{L}}{\partial\varphi^a}\delta\varphi^a-\mathrm{D}_\mu\frac{\partial\mathcal{L}}{\partial(\mathrm{D}_\mu\varphi^a)}\delta\varphi^a\right\}+\int_{\partial\mathrm{M}} d\mu_{\partial\mathrm{M}}\left[\frac{\partial\mathcal{L}}{\partial(\mathrm{D}_\mu\varphi^a)}\delta\varphi^a\right].\end{aligned}$$

Since the field variations vanish at the boundary the variational principle $\delta\mathrm{S}=0$ reduces to the classical equation of motion

$$\frac{\partial\mathcal{L}}{\partial\varphi^a}-\mathrm{D}_\mu\frac{\partial\mathcal{L}}{\partial(\mathrm{D}_\mu\varphi^a)}=0.$$

These are the Euler-Lagrange equations, which defines a subset of physical field configurations in $\mathcal{F}_{\mathrm{M}}$. Note that we only assumed here that the field variations vanish at the boundary, while the fields themselves need not. In this sense, a total derivative term $d\omega$ will not affect the local equations of motion, since the variation of the action would be

$$\delta S + \int_{\partial M} \delta\omega$$

and all variations on the boundary vanishes. The boundary term may however hold global information not contained in the local equations of motion. This global information can be accessible in a quantum theory where one integrates over all field configurations.

When the field variations stem from a group action we can get another useful result. For a Lie group G and elements X of its Lie algebra we write

$$(g(t)\varphi)^a = (g(t))^a_b\varphi^b = (e^{tX})^a_b\varphi^b = \varphi^a + tX^a_b\varphi^b + \ldots$$

where we can read of $\delta\varphi^a = X^a_b\varphi^b$. If these classical fields satisfy the Euler-Lagrange equations, we know from above that the variations satisfy

$$\delta S = \int_M d\mathrm{vol}_g D_\mu\left[\frac{\partial\mathcal{L}}{\partial(D_\mu\varphi^a)}\delta\varphi^a\right] = \int_M d\mathrm{vol}_g D_\mu\left[\frac{\partial\mathcal{L}}{\partial(D_\mu\varphi^a)}X^a_b\varphi^b\right].$$

We denote the expression in the square brackets as J^μ. If the group G is a symmetry of the theory, the variation of the Lagrangian must at most be a total derivative $\delta\mathcal{L} = D_\mu F^\mu$ [67], and we have $D_\mu j^\mu = 0$ for $j^\mu = J^\mu - F^\mu$. Most important maybe is the stress-energy tensor current, obtained by translational invariance $x^\mu \to x^\mu + \epsilon^\mu$ for constant ϵ. The fields and Lagrangian transform as $\varphi^a \to \varphi^a + \epsilon^\mu D_\mu\varphi^a$, $\mathcal{L} \to \mathcal{L} + \epsilon^\mu\partial_\mu\mathcal{L}$ and gives a current

$$T^\mu_\nu = \frac{\partial\mathcal{L}}{\partial(D_\mu\varphi^a)}D_\nu\varphi^a - \delta^\mu_\nu\mathcal{L}.$$

In fact, by letting ϵ be coordinate dependent we can repeat the above to get a more general current $j^\mu = T^\mu_\nu\epsilon^\nu(x)$.

There is also another expression for the stress-energy tensor. This expression can be obtained by promoting the metric to a dynamical field of the theory. If one also promotes the parameters specifying the transformation to be spacetime dependent, we can view the transformation simply as a diffeomorphism. Since the theory should be independent of the chosen coordinates, i.e. has Diff(M) invariance, the change in the action originating from the change in fields must be canceled by the change stemming from the coordinate transformation of the metric. Hence the two are equal and opposite. In our case, note that by promoting ϵ to $\epsilon(x)$ the change in action must be on the form

$$\delta S = -\int d\mathrm{vol}_M J^\mu_\nu \partial_\mu \epsilon^\nu$$

since in the spacetime independent case this variation vanishes. By doing a integration by parts one can easily see that J is a conserved current. Using $\delta g_{\mu\nu} = \partial_\mu \epsilon_\nu + \partial_\nu \epsilon_\mu$, the change in the action due to the metric transformation is

$$\delta S = -\int d\mathrm{vol}_g \frac{\delta S}{\delta g_{\mu\nu}} \delta g_{\mu\nu} = -2\int d\mathrm{vol}_g \frac{\delta S}{\delta g_{\mu\nu}} \partial_\mu \epsilon_\nu.$$

The conserved current can be seen to be $T_{\mu\nu} = -2\delta S/\delta g^{\mu\nu}$, up to normalization factors. This expression will be useful when we discuss conformal field theory. For a more detailed explanation see [28].

When we study a theory with interactions, one often deforms the original theory by adding terms

$$S = \int_M d\mathrm{vol}_g \langle \varphi, Q\varphi \rangle + \sum_i c_i \int_M d\mathrm{vol}_g \mathcal{O}_i(\varphi),$$

where the c_i's are called coupling constants and $\mathcal{O}_i(x)$ are called local operators, consisting of combinations of the fields. Note that a field theory can consist of several fields by considering tensor products of different bundles $\times_i \mathcal{E}_i$. We will from now on refer to the first term as the free action $S_0(\varphi)$. We imagine the c^i's to be the local coordinates on a space called parameter space or moduli space of the theory. We will return to this space in more detail later, once we have discussed the renormalization group. We will see that not all such deformation are relevant at all length scales.

3.2 Quantum aspects

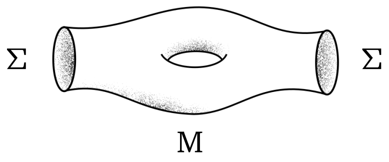
Quantum field theory is a wonderful subject both from a physical and mathematical perspective. Having discussed the basic ingredients in a classical theory of fields, we are ready for quantization. As in any quantum theory, we need a Hilbert space of states. We will assume that at a given time the system lives on a spatial manifold Σ_1 of dimension $d-1$. To this spatial slice of spacetime we associate a Hilbert space $\mathcal{H}_{\Sigma_1}$ of quantum states. The most general type of quantum field theory assumes nothing of the geometric structures on spacetime, and at a later time the system could find itself on a spatial manifold Σ_2, with spacetime a cobordism M as above. This spatial slice comes with its own Hilbert space, and the spacetime M can be interpreted as a geometric analogue of time passing.

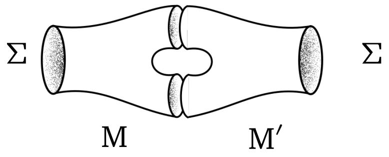
If we recall back to the chapter on mathematical structures in the beginning of this thesis we see that the above rules of associating to $(d-1)$-manifolds a Hilbert space and to d-manifolds a time translation operator is exactly a functor of the spacetime category

$$\mathcal{Z} : \mathrm{dCob}_* \to \mathrm{Hilb}$$

into the category of Hilbert spaces. This approach to quantum field theories was pioneered by Segal [77] and Atiyah [5] in the setting of topological and conformal field theories. As in the classical case, the quantum field theory is given different names depending on the geometric structures on spacetime. The study of different types of quantum field theories is then tantamount to studying functors on different cobordism categories. The most familiar case in for example particle physics is when spacetime has a Riemannian structure are we consider the cobordisms $\Sigma \times [0, \mathrm{T}]$, i.e. generalized cylinders of different length. The length T is the time passed. In any case there are certain axioms the functor should satisfy. We will not go into too much detail regarding this, but the interested reader can see [22] for a more detailed discussion.

We mention some of the key features of the field theory functor. First of all, when a boundary is a disjoint union of Σ_i's the Hilbert space is a tensor product space $\otimes_i \mathcal{H}_{\Sigma_i}$. This is simply the standard rule in quantum theory for combining systems into larger systems. When the boundary is the empty set the associated Hilbert space is $\mathbb{C}$. From our brief discussion of categories we recall that there is also a composition rule for the morphisms in a category. The composition rule in dCob is by gluing one cobordism on top of another, along an identical boundary.





The functor has to be compatible with this composition, in the sense that $\mathcal{Z}(\mathrm{MM}') = \mathcal{Z}(\mathrm{M})\mathcal{Z}(\mathrm{M}')$. When the cobordism M has two identical boundaries Σ, we may glue the manifold together along Σ to get a closed manifold M_c. This cobordism starts and ends at the empty set, so $\mathcal{Z}(M_c) : \mathbb{C} \to \mathbb{C}$, while $\mathcal{Z}(M) : \mathcal{H}_\Sigma \to \mathcal{H}_\Sigma$. These are to be compatible in the sense that $\mathcal{Z}(M_c) = \mathrm{Tr}_{\mathcal{H}_\Sigma} \mathcal{Z}(M)$. This is the so called partition function of the theory in the canonical formulation [22].

In this case of Riemannian structure with simple cylindrical cobordisms the functor maps the cobordism to

$$\mathcal{Z}(\Sigma \times [t_1, t_2]) \equiv U(t_2, t_1) : \mathcal{H} \to \mathcal{H}$$

where $\mathcal{H}$ is the Hilbert space associated with Σ. We write this operator $U(t_2, t_1) = \exp(-(t_2 - t_1)H)$ where the operator H is the Hamiltonian generating time translations [37]. By the composition of cobordisms, this satisfies the normal composition rules of time translations.

We will be working mainly in the functional approach to quantum field theory. Here the partition function is calculated by a integral over fields rather than as a trace over Hilbert space. The fundamental but somewhat schematic equation in a quantum field theory is

$$\mathcal{Z} = Tr_{\mathcal{H}} U = \int \mathscr{D}\varphi e^{-S(\varphi)}.$$

In other words, the number the QFT functor associates to the closed manifold can be calculated by an integral over the field space.

If we interpret $\mathscr{D}\varphi e^{-S}$ as a weighted measure of the field space, the partition function is a sort of effective volume of $\mathcal{F}_M$. Roughly speaking, the partition function is a measure of weighted degrees of freedom. We should note that in general the partition function should be defined with an additional sum over topological sectors of the field configuration space. This is clear, since the field integral can only take into account the configurations that can be continuously deformed into each other. For example, consider a theory of maps $\varphi : \mathbb{S}^d \to T$ where T is some manifold with possibly non-trivial topology. The field configuration space $\mathcal{F}$ can in this case be considered as combination of subsets $\mathcal{F}_i$ consisting of field configurations in a particular class of $\pi_d(T)$.

In a quantum theory the quantum fluctuations around the classical solution must be taken into account. An observable $\mathcal{O}$ has an expectation value calculated by

$$\langle \mathcal{O} \rangle = \frac{\int \mathscr{D}\varphi \mathcal{O}(\varphi) e^{-S(\varphi)}}{\int \mathscr{D}\varphi e^{-S(\varphi)}} = \mathcal{Z}^{-1} \int \mathscr{D}\varphi \mathcal{O}(\varphi) e^{-S(\varphi)}$$

which we can read as a standard expectation value in a probability distribution. Recall that we often deform the action to include $\delta S = \sum_i c^i \int \mathcal{O}_i$ as perturbations.

In general, one expands amplitudes $\mathcal{A}$ like the above expectation value as a formal power series

$$\mathcal{A} = \sum_n \mathcal{A}_n c^n$$

around a point $c = 0$ is the moduli space of the theory. Each of the terms $\mathcal{A}_n$ can be calculated by diagrammatic techniques, i.e. Feynman rules [67]. Very schematically, each of the terms in a perturbative expansion take the form

$$\mathcal{A}_n = \sum_\Gamma \frac{a_\Gamma}{|\mathrm{Aut}\Gamma|}$$

where one sums over graphs Γ, calculates the so called value of the graph a_Γ by Feynman rules and divides by the size of the symmetry group of the graph [23]. A lot of interesting things can be said about these graphs. Typically the Hilbert space of a QFT is constructed as a Fock space where the single particle Hilbert space is a representation space of the symmetry group of the problem. The diagrams in a perturbative expansion can be seen as representation homomorphisms, mapping a n-particle state to a m-particle state [36]. These maps are also called intertwiners as they "intertwine" two representations. The corresponding complex number a_n is obtained by forming an inner product between the relevant states. We will not discuss these perturbation aspects any further, as only non-perturbative aspects of quantum field theory will be relevant in our later discussions. The interested reader can see for example Wittens chapter in [20].

Notice that we can expand an observable in a power series so that

$$\langle \mathcal{O} \rangle = \sum_n \frac{1}{n!} \int dx_x ... dx_n \frac{\delta^n \mathcal{O}}{\delta\varphi(x_1)...\delta\varphi(x_n)} \langle \varphi(x_1)...\varphi(x_n) \rangle \, .$$

We here used scalar fields in a one dimensional theory for simplicity. In this sense, any calculation relies on the so-called n-point functions $\langle \varphi(x_1)...\varphi(x_n) \rangle$. These are the basic objects we calculate in a quantum field theory. We would like to digress slightly. Consider the number sequence $\{a_n\} = \{0, 1, 1, 2, 3, 5, 8, 13, ...\}$ i.e. the Fibonacci numbers. Define a function f by the formal series

$$f(x) = x + x^2 + 2x^3 + ... = x + \sum_{n=2}^{\infty} a_n x^n .$$

Since the Fibonacci numbers satisfy $a_n = a_{n-1} + a_{n-2}$ we can by shifting the indexes in the sum write

$$f(x) = x + f(x)x + f(x)x^2 \rightarrow f(x) = \frac{x}{1 - x - x^2} .$$

This function now has the property that its nth derivative is exactly the nth Fibonacci number. This is a part of a general idea when dealing with number sequences, called generating functions. These functions hold a lot of information regarding the number sequence. A natural question in our quantum field setting is then the following. What is the generator of the number sequence $\{\langle\varphi(x_1)...\varphi(x_n)\rangle\}$? The answer turns out to be the "deformed" partition function

$$\mathcal{Z}[\mathrm{J}] = \int \mathcal{D}\varphi e^{-\mathrm{S}(\varphi)-\int dx \mathrm{J}(x)\varphi(x)}.$$

This can be seen by explicitly doing a expansion in the field $\mathrm{J}(x)$

$$\mathcal{Z}[\mathrm{J}] = \sum_n \frac{(-1)^n}{n!}\mathcal{Z}[0]\int dx_1...dx_n \langle\varphi(x_1)...\varphi(x_n)\rangle \mathrm{J}(x_1)...\mathrm{J}(x_n).$$

Hence we have a relation between the nth derivative of the partition function and and n-point functions. Such relations can also be found for other field types. The important point is that the partition function encodes all the geometric information in our theory, from which everything else may be calculated by the above arguments.

3.3 Gauge theories

A large class of field theories have as fields the connections on some G-bundle over M. These are the gauge theories, where the group G is called the gauge group. We will consider two types of gauge theories here, first the topological Chern-Simons theory before we consider Yang-Mills theories.

3.3.1 Chern-Simons theory

Chern-Simons theory is a topological quantum field theory, i.e. the spacetime belongs to the topological category. Recall from our discussion of Chern cohomology the definition

$$c_k = \frac{(i/2\pi)^k}{k!}\mathrm{T}r(\mathrm{F}^{\wedge k})$$

where the complex prefactor is chosen to that the integral of this form over a closed $2k$-manifold takes integer values. The Chern-Simons form CS_{2k-1} is the (2k-1)-form defined (up to normalization) by the relation

$$d\mathrm{CS}_{2k-1} = c_k.$$

For example, in the case of a U(1)-valued connection the relation between Chern form and Chern-Simons form read

$$d(\mathrm{A}\wedge d\mathrm{A}) = d\mathrm{A}\wedge d\mathrm{A} + \mathrm{A}\wedge d^2\mathrm{A} = \mathrm{F}\wedge\mathrm{F} = -8\pi^2 c_2.$$

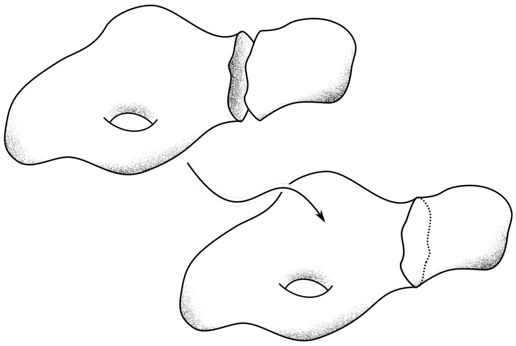
In a more general setting, Chern-Simons theory is defined by the classical action

$$\mathrm{S}_{\mathrm{CS}} = k \int_{\mathrm{M}} \mathrm{CS}_{2k-1}$$

where M is a $2k-1$ dimensional manifold. We will assume this manifold to be closed. The constant k is called the level of the theory. To quantize the theory, let us first note that we can use Stokes theorem to rewrite the action as

$$k \int_{\mathrm{Y}_1} c_k$$

where Y_1 is a $2k$-manifold such that $\partial \mathrm{Y}_1 = \mathrm{M}$. However, such a geometric extension is not unique. We can also consider a manifold Y_2 such that $\partial \mathrm{Y}_2 = \overline{\mathrm{M}}$, where we mean the manifold M with reversed orientation.



By combining the two manifolds Y_i one can show that the number k must be quantized in order for the partition function of the theory to be well behaved. Consider the difference in the extended actions

$$\Delta \mathrm{S} = k \int_{\mathrm{Y}_1} c_k - k \int_{\mathrm{Y}_2} c_k = k \int_{\mathrm{Y}_1 \cup \mathrm{Y}_2} c_k = k\mathrm{C}_k.$$

Here $\mathrm{Y}_1 \cup \mathrm{Y}_2$ is a closed manifold, so the integral over the Chern form yields the integral Chern number. In this way, one of the extended Chern-Simons actions can be written in terms of the other and an additional term $k\mathrm{C}_k$. In the partition function for the theory this takes the form

$$\int \mathscr{D}\mathrm{A} e^{ik \int_{\mathrm{Y}_1} c_k} = \int \mathscr{D}\mathrm{A} e^{ik\mathrm{C}_k} e^{ik \int_{\mathrm{Y}_2} c_k}.$$

For the partition function to be well-defined, this additional phase must be unity. Hence $k\mathrm{C}_k \in 2\pi\mathbb{Z}$ so, since the Chern number is an integer, we must have $k = 2\pi n$. This will be important when we discuss the so-called class A topological insulators later. Here each topological phase corresponds to a Chern-Simons theory with a particular level. Since the level must be integral, this means that the physical phases must also be labeled by integers.

3.3.2 Yang-Mills theory

We here briefly discuss Yang-Mills theory on a Riemannian manifold M. This topic is covered is more detail in for example [60][19]. We consider on M the overlapping subsets U_a and U_b and local sections of a vector bundle $\mathcal{V} \to M$

$$s_a : U_a \to \mathcal{V},$$

$$s_b : U_b \to \mathcal{V}.$$

Note that the latin indices here denote to which subset the sections belong, not a vector index. The transition functions

$$\phi_{ba} : U_a \cap U_b \to G$$

necessarily maps $\phi_{ba} s_a = s_b$. Hence we can also use this to see how the differentiated section Ds changes from one local trivialization to another:

$$\phi_{ba}(d + A_a)s_a = \phi_{ba} d\phi_{ba}{}^{-1} s_b + ds_b + \phi_{ba} A_a \phi_{ba}{}^{-1} s_b \overset{!}{=} (d + A_b)s_b$$

$$\therefore A_a = \phi_{ba}{}^{-1} d\phi_{ba} + \phi_{ba}{}^{-1} A_b \phi_{ba}.$$

By introducing a coupling constant the covariant derivative is often written $D_\mu = \partial_\mu + icA_\mu$. Writing $\phi_{ba}(x) = g(x)$, the above transformation reads

$$A_b \to A_a = -\frac{i}{c} g^{-1}(x)\partial_\mu g(x) + g^{-1}(x) A_b(x) g(x).$$

The first term implies that the connection does not transform as a tensor. In fact, under such a local gauge transformation the connection transforms in the adjoint representation of the gauge group [57]. From the connection 1-form we can construct the curvature 2-form

$$F = dA + A \wedge A = \frac{1}{2}(\partial_\mu A^a_\nu - \partial_\nu A^a_\mu + f^a_{bc} A^b_\mu A^c_\nu) t_a \otimes dx^\mu \wedge dx^\nu$$

where the f_{abc}'s are the structure constants of the Lie algebra $\mathfrak{g}$. The transformation behavior of this curvature is determined from the transformation of the connection, and one can show that it indeed transforms tensorially.

The (local, classical) Yang-Mills action on a Riemannian manifold M with metric g is defined by

$$\mathrm{S}_{\mathrm{YM}}[\mathrm{A}] = \frac{1}{4}\int_{\mathrm{M}} \mathrm{dvol}_g \mathrm{F}^a_{\mu\nu}\mathrm{F}^{\mu\nu}_a.$$

The partition function of the quantized theory reads

$$\mathcal{Z}[\mathrm{A}] = \int \mathcal{D}\mathrm{A}e^{-\mathrm{S}_{\mathrm{YM}}}.$$

Note however that in this case the relevant space of field configurations is not the full space of connections $\mathcal{A}$, as we have a local equivalence given by the gauge transformations. If we think of the partition function as a measure of the volume of the space of field configurations, the naive volume would be all too large. Hence the integral should be performed over $\mathcal{A}/\mathrm{G}$, namely over individual gauge equivalence classes. Rewriting the integral in terms of these physically different field configurations leads to the Faddeev-Popov determinant and ghost fields [67] [19].

The relevant field configuration space for the theory considering both the gauge fields and the fields defined by our original vector bundle $\mathcal{V}$ where the gauge group acts is both $\mathcal{A}/\mathrm{G}$ and $\Gamma(\mathcal{V})$, and the partition function is formally a map from the product space of these to the real numbers. Often one considers scalar or spinor fields, as defined by a line bundle or spinor bundle respectively, and makes these $\mathcal{V}$-valued to extend the internal symmetry G to these fields. More precisely, if we have fields defined as sections $\Gamma(\mathcal{E})$ and a Yang-Mills theory as above we can consider the theory with field configuration space

$$\Gamma(\mathcal{E}\otimes\mathcal{V}).$$

Now the matter fields are sections of the tensor product bundle $\mathcal{E}\otimes\mathcal{V}$ where the first factor represents the degrees of freedom (e.g. scalar, vector, spinor and so on) and the second factor represents the gauge symmetry. Note that on this bundle there is a natural connection defined by the connections of the respective factor bundles. If s and v are sections of $\mathcal{E}$ and $\mathcal{V}$ respectively we have

$$\mathrm{D}(s\otimes v) = (\mathrm{D}^{\mathcal{E}}s)\otimes v + s\otimes(\mathrm{D}^{\mathcal{V}}v).$$

In the case of spinors on M the action would be

$$\mathrm{S}(\mathrm{A},\psi) = \mathrm{S}_{\mathrm{YM}}(\mathrm{A}) + \mathrm{S}_{Dirac}(\psi,\mathrm{A})$$

with Dirac action $\int\overline{\psi}i\gamma^\mu\mathrm{D}_\mu\psi$, where the covariant derivative acts on the product bundle.

3.4 The non-linear sigma models

The nonlinear sigma models are a large class of field theories with a highly geometric origin that belongs to the Riemannian category. We here review the basics of these field theories and discuss some simple examples. Nice discussions on these field theories can be found for example in [1].

3.4.1 Geometric setup

The nonlinear sigma models are field theories of maps between Riemannian manifolds $\varphi : \Sigma \to \mathrm{M}$, or equivalently fields which are sections of a trivial M-bundle over Σ. We will refer to Σ as a base manifold and M as the target manifold. We denote by $g_{\mu\nu}$ the base manifold metric tensor, and by h_{ij} the metric on the target space. Greek indices will be used for the base, while latin indices for the target. In our discussion of quantum field theories we mentioned that the base manifold could be equipped with all sorts of geometric structures. In the nonlinear sigma model case we have two manifolds on which to chose structures. In addition to Riemannian structure the two manifolds can be equipped with for example complex structure or spin structure. A particular interesting question is how the choice of target manifold structures are reflected in the theory. Later we will see an example where the complex structure of the target has an interesting physical interpretation.

Given a chart on the target manifold $\psi : \mathrm{M} \to \mathbb{R}^m$ we may construct the map

$$(\psi \circ \varphi)(p) = \psi(\varphi(p)) \in \mathbb{R}^m$$
$$\varphi^i \equiv (\psi \circ \varphi)^i$$

which are the local coordinates along the field. As in any field theory we want to construct some (scalar) Lagrangian from the data given. In this case we only have the geometric data $(\varphi, g_{\mu\nu}, h_{ij})$. The simplest real valued map we can construct from this is

$$\mathcal{L} : \varphi \to \mathcal{L}(\varphi) = g^{-1}(\varphi^* h)$$

where in local coordinates we have

$$g^{-1} = g^{\mu\nu}\partial_\mu \otimes \partial_\nu$$
$$h = h_{ij} d\varphi^i \otimes d\varphi^j$$
$$\varphi^* h = h_{ij}(\varphi)\partial_\mu\varphi^i\partial_\nu\varphi^j dx^\mu \otimes dx^\nu.$$

Equivalently we can use the explicit (local) form of the differential of φ we discussed in the chapter on pullbacks and pushforwards

$$d\varphi = \partial_\mu \varphi^a dx^\mu \otimes \partial_a$$

here viewed as a local section of the bundle $\mathscr{B} = \mathrm{T}^*\Sigma \otimes \varphi^*\mathrm{TN}$. On the cotangent bundle over the base we locally have an inner product given by contraction with g^{-1}, while on the pullback bundle over the target space by contraction with $h_{ij}(\varphi)$. These combine to give a inner product on $\mathscr{B}$. The inner product, or "size", of $d\varphi$ thus reads

$$\langle d\varphi, d\varphi\rangle_{\mathscr{B}} = g^{\mu\nu}\partial_\mu\varphi^i\partial_\nu\varphi^j h_{ij}(\varphi).$$

Hence we can write the action

$$\mathrm{S}[\varphi] = \int_{\mathrm{M}} d^n x\sqrt{g}g^{-1}(\varphi^* h) = \int_{\mathrm{M}} d^n x\sqrt{g}\,\langle d\varphi, d\varphi\rangle_{\mathscr{B}} = \int_{\mathrm{M}} d^n x\sqrt{g}g^{\mu\nu}\partial_\mu\varphi^i\partial_\nu\varphi^j h_{ij}(\varphi).$$

The role of this action is as usual to weigh the different field configurations properly in field integrals. The space of field configurations in the sigma models are often referred to as $\mathrm{Maps}(\Sigma, \mathrm{M})$. As the Lagrangian of this theory is constructed in a purely tensorial way, the action is necessarily invariant under target space diffeomorphisms, i.e. field redefinitions $\varphi \to \varphi'$.

Another interesting thing to note is the following. If we expand the target space metric in the fields, we formally have

$$h_{ij}(\varphi) = h_{ij}(0) + \partial_k h_{ij}(0))\varphi^k + \partial_k\partial_l h_{ij}(0))\varphi^k\varphi^l + ...$$

and one can interpret the nonlinear sigma model as a theory of scalar fields φ^i with infinitely many coupling constants

$$\{h_{ij}(0), \partial_k h_{ij}(0), ...\}.$$

Hence we are dealing with a free theory, in the sense that we have no potential or interaction term, but as a result of the non-trivial target space geometry we have infinitely many interactions.

We want to briefly sketch how spinors fit into this geometric picture. We here assume that the sigma model $\varphi : \Sigma \to \mathrm{M}$ consists of a spin manifold Σ where $\mathcal{S}\Sigma$ denotes its bundle of spinors. As spinors are associated to sections of this bundle there is a priori no connection between the spinor fields and the target space geometry. As the target is often chosen as to incorporate some symmetry into the physical model, we need to include its geometry somehow. This is usually done by twisting the spinor bundle over the base by the pullback tangent bundle of the target. The (twisted) spinors normally used for sigma models are then associated to sections

$$\psi \in \Gamma(\mathcal{S}\Sigma \otimes \varphi^*\mathrm{TM}).$$

In some way, we need the scalar field to define the spinor fields. As discussed in the chapter on spinors, we can locally write them

$$\psi = (\psi^\alpha)^i s_\alpha \otimes \partial_i = \psi^i \otimes \partial_i$$

where $\partial_i = \partial / \partial \varphi^i$ and $\{s_\alpha\}$ is a set of basis sections for the spinors. We want each of these spinors to anti commute when we treat them in the functional formalism. Just as we have n scalar fields φ^i, acting as coordinates on the target, we have n spinors ψ^i in the tangent space of the target. We will in the remainder of this section focus on the scalar theory.

It will be useful to consider the local classical equations of motion, i.e. the covariant Euler-Lagrange equations

$$\nabla_\mu \frac{\partial \mathcal{L}}{\partial(\partial_\mu \varphi^k)} - \frac{\partial \mathcal{L}}{\partial \varphi^k}$$

with

$$\frac{\partial \mathcal{L}}{\partial \varphi^k} = \frac{1}{2} g^{\mu\nu}(\partial_k h_{lj}) \partial_\mu \varphi^l \partial_\nu \varphi^j,$$

$$\frac{\partial \mathcal{L}}{\partial(\partial_\mu \varphi^k)} = g^{\mu\nu} h_{kj} \partial_\nu \varphi^j$$

leading to the equations of motion

$$g^{\mu\nu}(\nabla_\mu \partial_\nu \varphi^i + \Gamma^i_{jl} \partial_\mu \varphi^l \partial_\nu \varphi^j) = 0,$$

where Γ^i_{jl} are the Christoffel symbols on the target. We here used Levi-Civita connection on TM. The fields satisfying this equation is known as harmonic maps, and are the field configurations in Maps(Σ, M) interesting for a classical theory. Note that the second nonlinear term stem from the fact that in both the ways we constructed the action, pullbacks along the field were involved.

A useful picture to have in mind is the following. We can look at the map φ in the sigma model simply as a tool to embed Σ in M. Then the quantum theory in some sense studies deformations of Σ in M.

Note that in the trivial case where $\Sigma = \mathbb{R}^n$ and $\mathrm{M} = \mathbb{R}$ the equations of motion reduce to the massless Klein-Gordon equations $\partial^\mu \partial_\mu \varphi = 0$. Of course, for any theory with base manifold M and target the real line, the sigma models reduce to theories of a single scalar field on spacetime M.

3.4.2 Worldsheet models and B-field terms

A special class of sigma models that we will discuss again later are the world sheet models. Here the base manifold is a Riemann surface, while the target can be arbitrary. We start with a discussion of sigma models in one dimension before moving on to the two dimensional case. Consider the sigma model $x : \mathrm{W} \to \mathrm{M}$

where W is a 1-manifold, i.e. either a line or a circle . For now we take the base space to be a finite interval of the real line. The action of this theory reads

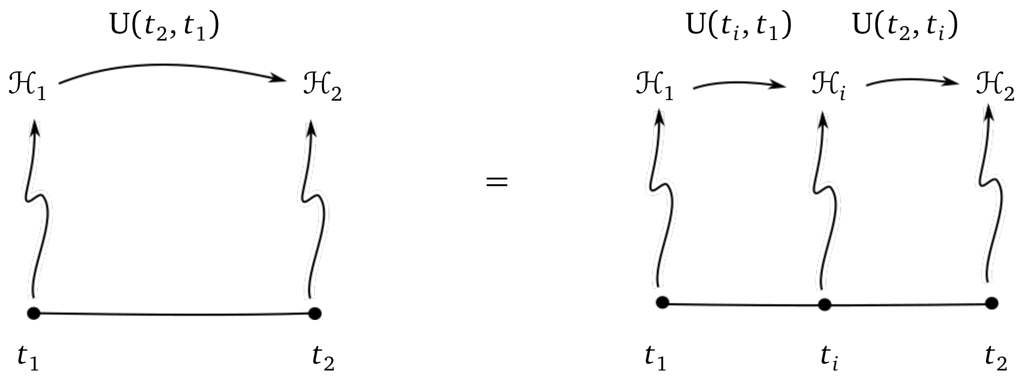
$$\mathrm{S} = \int (\dot{x}_i \dot{x}^i - \mathrm{V}(x)) dt,$$

where we added a potential term V : M → $\mathbb{R}$. If we interpret the base space interval $[t_1, t_b]$ as a section of time this theory is nothing but a quantum theory of trajectories trough M. The partition function of the theory reads

$$\mathcal{Z} = \int \mathcal{D}x e^{i\mathrm{S}}.$$

This sum may heuristically be read as the sum over paths weighted by the exponential factor. This is the path integral approach to quantum mechanics [37]. The corresponding Euclidian theory where paths are weighted by $\exp(-\mathrm{S}_\mathrm{E})$ has a natural interpretation as statistical mechanics.

In this sense quantum mechanics is a 0+1 dimensional field theory. To find the relevant Hilbert space of the theory we proceed as described in the in the above discussions. We can view the worldline of the particle as a 1-cobordism in the category 1Cob, and then consider the functor to the Hilbert space category. The operator $\mathrm{U} = \exp(i(t_2 - t_1)\mathrm{H})$, associated with the 1-cobordim, is the time evolution operator acting on the Hilbert spaces living on the ends of the worldline. This of course satisfies the wanted decomposition rules.



The Hilbert spaces at different times are isomorphic and we denote them by $\mathcal{H}$. By Taylor expanding this time evolution operator to first order, we can find a equation for a Hilbert space vector $|\psi\rangle$ at time $t+\delta t$, and hence also an expression for the time derivative of the vector. The resulting equation

$$-i\hbar\partial_t |\psi\rangle = \mathrm{H} |\psi\rangle$$

is of course the Schrodinger equation governing time evolution of the quantum system. An operator Q commuting with the time evolution operator is considered a symmetry of the problem. Since $[\exp(i\mathrm{TH}),\mathrm{Q}]=[1+i\mathrm{TH}+...,\mathrm{Q}]$ this simply means that the Hamiltonian commutes with Q. From this one often constructs the Hilbert space as eigenspace of the maximal set of commuting operators. This operator approach to quantum mechanics is discussed in the appendix.

If the particle moves in a electromagnetic field in n-dimensional Minkowski space, the Lagrangian has the additional term

$$-q(\rho-\mathbf{A}\cdot\dot{\mathbf{x}}).$$

This term can be written $\mathrm{A}_i\dot{x}^i$, with $\mathrm{A}_0=q\rho$. This we recognize as the component of the pullback one-form $x^*\mathrm{A}$ from a 1-form $\mathrm{A}_i dx^i$ on M. The gauge field term can then be written

$$\int_{\mathrm{W}} x^*\mathrm{A}=\int dt\mathrm{A}_i(x)\dot{x}^i.$$

Note that we could only pull back a one-form, as the base space is one dimensional, and hence only integration of one-forms is defined. However, from this we could try to construct similar pullback terms for the general sigma model. In fact, we have the general expression for the pullback of a n-form from our discussion of pullback bundles, which would fit nicely into such an action.

A similar interpretation is possible in two dimensions. For concreteness we let $\Sigma=\mathbb{R}\times\mathrm{S}^1$. Hence the sigma model in this case studies embedding of the cylinder in M, and can similarly to the d=1 case be interpreted as the motion of a object trough the target. In this case the object is a string, where the coaxial direction of the cylinder represents time and the circle the parameter on the string. By analogy with the pullback gauge field we can now pull back a 2-form B from the target [79], resulting in a addition to the action of the form

$$\mathrm{S}(\varphi)\to\mathrm{S}(\varphi)+\int_{\Sigma=\mathbb{R}\times\mathrm{S}^1}\varphi^*\mathrm{B}=\int_{\Sigma}\mathrm{B}_{ij}(\varphi)\epsilon^{\mu\nu}\partial_\mu\varphi^i\partial_\nu\varphi^j d^2x.$$

When one adds such a term to a two dimensional sigma model in a more general setting it is often called a Wess-Zumino-Witten term, 2-form term or a B-field term, and the total action would be

$$\mathrm{S}(\varphi)=\int_\Sigma d^2x\sqrt{g}g^{-1}(\varphi^*h)+\int_\Sigma\varphi^*\mathrm{B}$$

$$=\int_\Sigma d^2x\left[\sqrt{g}g^{\mu\nu}h_{ij}(\varphi)+\epsilon^{\mu\nu}\mathrm{B}_{ij}(\varphi)\right]\partial_\mu\varphi^i\partial_\nu\varphi^j.$$

Note that no metric is needed to integrate this second term, and in this sense it is topological. Note also that from Stokes theorem we can rewrite this additional term if we interpret Σ as the boundary $\partial \mathrm{N}$ of a 3-manifold

$$\int_{\Sigma} \varphi^* \mathrm{B} = \int_{\mathrm{N}} d(\varphi^* \mathrm{B}) = \int_{\mathrm{N}} \varphi^*(d\mathrm{B}).$$

We see that only the cohomology class of B in $\mathrm{H}^2(\mathrm{M})$ matters to this sigma model. The dimension of the moduli space of this model is determined by the number of parameters needed to define the relevant tensors, e.g. two metrics and a 2-form. While Σ is two dimensional and its metric requires three parameters, the target can be of any dimension m. Counting the independent parameters one can easily show that

$$\dim\mathcal{M} = \binom{m}{2} + \frac{m(m+1)+6}{2}.$$

So for two dimensional targets the dimension is 7, in three dimensions 12 and in four dimension the moduli space is 19 dimensional. The case relevant for us later will be when $\Sigma = \mathbb{R}^2$, and the extra three dimensions coming from the metric on Σ has to be subtracted. The dimensions of the sigma model moduli space then reads $\dim\mathcal{M} = \binom{m}{2} + \frac{m(m+1)}{2}$.

We would like to briefly discuss how D-branes appear in the geometric formulation of sigma models. Consider a model of maps $\varphi : \Sigma \to \mathrm{M}$ where Σ is the rectangle $[t_1, t_2] \times [s_1, s_2]$ and M is the manifold in which the string moves. This model would describe an open string, while the former example, where the world sheet was a cylinder, described the motion of closed strings. In the open case, the world sheet has some boundaries which we must handle carefully. The action is the standard sigma action

$$\mathrm{S} = \frac{1}{2}\int dtds\partial_\mu \varphi^i \partial^\mu \varphi_i$$

and the variation with respect to the fields read

$$\begin{aligned} \delta \mathrm{S} &= \int dsdt\partial_\mu \varphi^i \partial^\mu \delta\varphi_i && (3.1)\\ &= \int dtds\left\{\partial^\mu[(\partial_\mu \varphi^i)\delta\varphi_i] - \partial^\mu\partial_\mu\varphi^i\delta\varphi_i\right\}. && (3.2)\end{aligned}$$

In this case, we can not throw away the total derivative term. Writing out the contraction over the world sheet, this is evaluated to

$$\int_{s_1}^{s_2} ds[\partial_t \varphi^i \delta\varphi_i]_{t_1}^{t_2} - \int_{t_1}^{t_2} dt[\partial_s \varphi^i \delta\varphi_i]_{s_1}^{s_2}.$$

By assumption the field variations vanish at t_1, t_2 so the first terms simply vanishes. For the second term to vanish, i.e. $[\partial_s \varphi^i \delta \varphi_i]_{s_1}^{s_2} = 0$, we must demand that either the derivative vanishes, or there is no field variation. We consider the case

$$\partial_s \varphi^i|_{s_{1,2}} = 0, i = 1, ..., m\,,$$

$$\delta \varphi^i|_{s_{1,2}} = 0, i = m+1, ..., n = \mathrm{dimM}.$$

This means that the end of the string is attached to a m dimensional sub manifold B ⊂ Mof the target space. These geometric objects are called Dm-branes. For example, a D0-brane would simply be a point in the target, a D1-brane a string, or D2-brane a membrane, and so on. If we want to interpret these D-branes in a more general setting, not necessarily as part of string theory, they simply appear as boundary information in the sigma models.

4

Universality and the renormalization group

This chapter is intended to bridge the gap between the previous discussion on quantum field theories and the theory of phases and phase transitions. In particular we will discuss the notion of emergence and universality as they shed light not only on how to use field theories in condensed matter physics but also on constructing physical models in general. Large parts of this chapter will be focused on field theories and the renormalization group, which gives a very clear presentation of universality.

4.1 Emergence

The simplicity of complexity of a system depends crucially on its size. A single-particle system is often solvable, at least under some set of reasonable assumptions. Adding more particles increases the complexity considerably as interactions have to be included. While few-particle systems may be solved, systems with tens or hundreds of particles are more or less impossible to solve without turning to numerical methods. However, as the system grows larger still the very thing that made the few-particle problem harder is making the problem tractable in the infinite limit. In this way we pass from simplicity to complexity back to simplicity. What saves the day is the principle of emergence.

Emergent phenomena are roughly speaking those phenomena that can not be understood from models of independent particles, but rather are determined by interactions in the limit of large systems. In systems with interactions, the constituents can join forces to produce collective behavior. This collective motion of many particles is highly diverse and in many cases surprising. The effective degrees of freedom in the system can be rather exotic, which is the beauty of research in condensed matter. There is a qualitative difference between a system with a finite number of constituents and the limit where the system grows very large.

As we consider the large distance limit of some system, the microscopic details are replaced with averages. For example, the random motion and collisions in a material due to thermal motion is averaged out when we look at some large slab of the material in the lab. In stead the microscopic motion is reflected in material properties like density. In this sense, there is a certain amount of autonomy in the large scale behavior.

To summarize, the concept of emergence contains not only the fact that microscopic behavior can lead to effective collective behavior of large scales, but also that this behavior is independent of many of the microscopic details. This is, in the fewest possible words, universality: the observation that many systems can exhibit the same large scale properties. The class of systems with the same large-distance properties define an universality class.

4.2 Phases of matter as equivalence classes

Condensed matter physics roughly speaking studies the properties of materials and many-particle systems at low energies. The most important amongst these large-distance properties are the different phases and their transitions. We first explore the idea of phase transitions within a classical framework. For a system with a large number of particles, we let the phase space be denoted $\mathcal{P}$. The state of the system is determined by a point $x = (q_1, ..., q_n, p_1, ..., p_n)$ and it time evolution as determined by the Hamiltonian equations. Assuming energy to be conserved, we only consider the $2n - 1$ dimensional surfaces of constant energy E as phase space $\mathcal{P}$. The physical observables are $\mathrm{C}^\infty(\mathcal{P})$ functions. The physically observed quantity is often the temporal average

$$\langle \mathrm{A} \rangle_{\text{time}} = \lim_{t \to \infty} \int_0^t \mathrm{A}(x(t'))dt'.$$

The ergodic hypothesis states that in the asymptotic limit $t \to \infty$ all points of $\mathcal{P}$ will be explored. This means that the above temporal average can be exchanged with an average over phase space. This allows for the whole framework of statistical mechanics where such ensemble averages are calculated by differentiation of the partition function.

However, as this only is strictly true is the asymptotic limit, it may be that the ergodic hypothesis fails at the timescales involved in an actual experiment. In this case it may be that the phase space splits [91] into parts

$$\mathcal{P} = \cup_a \mathcal{P}_a.$$

In each region the ergodic hypothesis still holds, and the temporal average equals the (restricted) ensemble average.

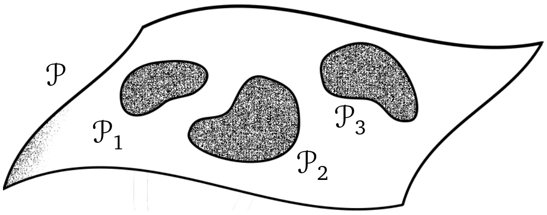


Figure 4.1: Sketch of the phase space $\mathcal{P}$ and its splitting into disjoint regions, e.g. phases. Each region corresponds to a class of collective behavior or organization of the many-particle system.

Symmetry breaking is a special example of ergodicity breaking. We assume that the microscopic model, for example as expressed by a Hamiltonian, is invariant under some transformations on the microscopic variables. If we can find observables that are not invariant under this transformation, and these have non-zero expectation value, we say that the symmetry is broken. Such a parameter is called a order parameter, as it reflects the microscopic organization of the systems constituents. The value of such a parameter will change under the broken symmetry transformations, and in some sense classify the classes of macroscopic behaviour. The non-zero expectation value is a consequence of the fact that the time average and (full) ensemble average will not agree - in this sense symmetry braking is a special case of ergodicity breaking.

In the classical picture, a phase is a collection of points in phase space (e.g. classical states) that share some physical property, like for example the order parameters we briefly mentioned. A definition similar in spirit also exist in the quantum picture of things. Phases can formally be seen as equivalence classes on the space of quantum mechanical systems. We imagine a space of quantum systems, where the points are labeled by (H, G). Here G is the symmetry group of the system from which the Hilbert space can be constructed as a representation, and H is the Hamiltonian compatible with this symmetry. For a fixed G, the Hamiltonian typically takes the form

$$\mathrm{H} = \mathrm{H}_0 + \sum_i \lambda_i \mathrm{H}_i$$

which we can think of as a deformation class of H_0. The space of parameters (moduli) span the a moduli space of quantum systems. Our general definition of a phase will be a equivalence class on $\mathcal{M}$ defined by an equivalence relation $\sim$. From the point of view of quantum states, this statement is very similar to the classical picture: a phase is simply a class of states that share some physical properties.

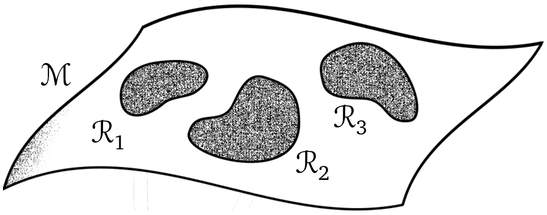


Figure 4.2: The moduli space $\mathcal{M}$ with regions $\mathcal{R}_i$ corresponding to equivalence classes under $\sim$.

What type of phase depends on the equivalence relation we choose. The standard picture of phases of matter are obtained when the equivalence relation is defined as path connected points in $\mathcal{M}$ for which the free energy $\log Z$ is non-singular [17]. Since the physical observables are usually given in terms of derivatives of the free energy or the partition function this simply means that the phases are defined as regions of analyticity of observables. When we discuss topological matter in upcoming chapters, the equivalence relation is of a topological nature. In this case the moduli space is partitioned into topological classes, each labeled by topological invariants.

4.3 Effective field theories

The justification of the application of field theory to condensed matter systems is universality and emergence. As we have discussed, the emergent large-distance physics depends only on certain aspects of the small-distance physics. Studying the condensed matter system at large scales is then tantamount to studying a field theory in the correct universality class. The concept of a universality class will become much more precise when we discuss renormalization group flow below.

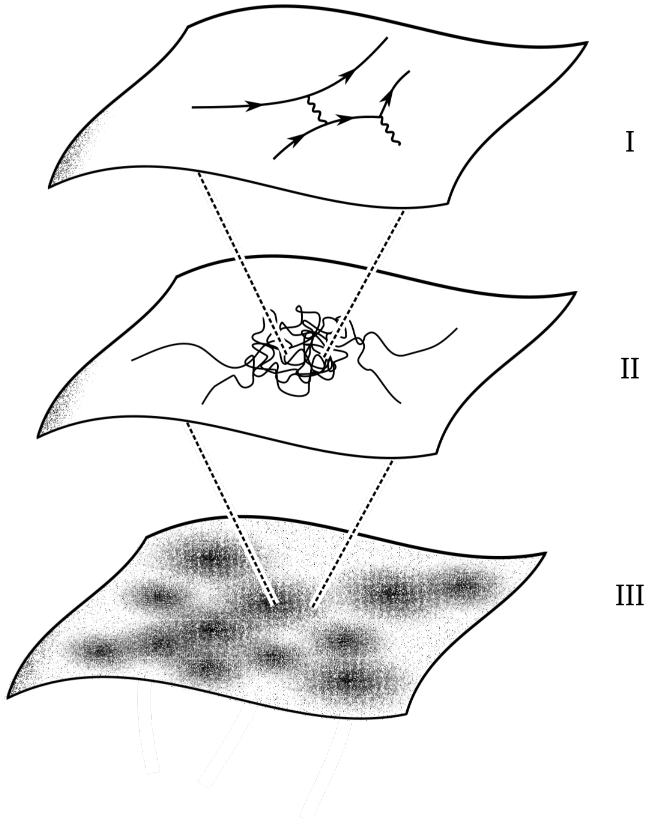


Figure 4.3: Simplicity at a microscopic level (I) turns into a complex intermediate range (II) before simplicity emerges (III) at large scales. At large scales an EFT describes universal properties of the original system.

We will represent the macroscopic (low energy) degrees of freedom of our system by a field φ which we take as a section of some appropriate bundle. The action is a function on the corresponding space of sections, and the (Euclidian) theory is represented by the partition function

$$\mathcal{Z} = \int \mathscr{D}\varphi e^{-S(\varphi)}$$

as always. We want an effective field theory to satisfy the following.

(i) Manageability: We should profit in going from the microscopic model to a EFT description of the system. If we can solve a microscopic model exactly, there is no more information to be obtained through the EFT other that a possible change of perspective.

(ii) Emergence: The effective theory should in principle emerge from some limiting procedure of the microscopic model. Given a micro-model the effective theory is obtained from some scheme of averaging over small scale effects, leaving only the relevant large-scale degrees of freedom.

(iii) Symmetries: If a microscopic model is not known and the procedure schematically outlined in (ii) is not possible, we use the characteristic symmetries expected of the system to study a class of candidate actions S. It is then a matter of (educated) guesswork to pick the right model for the given universality class.

Note that prom the point of view of (ii) effective theories are more than approximations. We do not simply forget information regarding the short-distance physics but rather include them by averaging over them. Somewhat schematically we can write

$$\mathcal{Z} = \int \mathcal{D}\varphi e^{-S} = \int \mathcal{D}\varphi_L \mathcal{D}\varphi_H e^{-S} = \int \mathcal{D}\varphi_L e^{-S_{eff}}$$

where we divided the fields into high and low energy modes. Here the effective action $S_{eff} \sim \log \int \mathcal{D}\varphi_H \exp(-S)$ describes the large-distance physics and is in every way equivalent to the full theory as long as one stays n the low-energy regime.

Note that while effective field theories obtained along the lines of (iii) performs no such averaging over high energy modes, the resulting field theory will describe the effective low-energy degrees of freedom. Indeed, the symmetries of the system are observed in macroscopic experiments, and a field theory based on this symmetry should contain information about the observed emergent phenomena.

This latter observation more or less summaries the phenomenological approach to effective theories, where the main idea is the following. We treat our many-particle system as a mysterious black box that we know nothing about, and construct a theory based only on observations. This almost sounds obvious, as this is simply the description of science. However, one often spends large amounts of time trying to derive a effective theory from some microscopic model in the belief that the microscopic theory is in some sense more fundamental than the effective one. In light of (iii), condensed matter can be treated in a manner very similar to high energy particle physics, where Lagrangians are constructed based on symmetries and couplings. In fact, we could see condensed matter and particle physics as the same class of system, where the only difference is that particle physics has less complicated vacuum [87]. In condensed matter systems the vacuum often corresponds to some ground state of a material. In this sense, we could view for example the QED vacuum as a "material" consisting of electrons and photons in which one studies for example scattering. Later we will see that this vacuum in fact corresponds to an insulator.

4.4 Geometry of the renormalization group

Renormalization and universality are two of the main pillars of modern physics, and sheds light on science in general. In a very broad sense it is simply the observation that physics at large scales is not dependent on all the microscopic details on smaller scales. If it was, doing science would be quite impossible. Consider for example a gas flowing around in 3-space. On scales larger than a couple of centimeters we use continuum mechanics to describe the relevant physics. However, at smaller scales of $10^{-8} - 10^{-11}$ cm we need atomic physics to describe the atoms in the gas and their electron cloud distribution. At even lower scales comparable to 10^{-15} cm quantum chromodynamics (QCD) rules, which is the physics of quarks and gluons. In this way, different scales correspond to different physics.

Luckily the physics at large scales more or less decouples from the physics at smaller scales. For example, we do not need to know about QCD to discuss either atomic physics or the classical fluid dynamics of the gas in the above example. This decoupling of length scales in nature is what enables us to do physics at all. If this was not the case, we would have to know everything before we can do anything. Of course, the small scale physics in some ways leaks through to the larger scale physics as a sort of average. Consider again the gas above. To do the classical continuum mechanics we need to know properties of the gas like density. This we often calculate using statistical models based on atomic physics. In this way the concept of an density can be defined by calculating the average number of particles in a finite volume V. In this way some of the information on small scales flow to the large scale physics. Intuitively it should be clear that such information flow in irreversible; As we forget details regarding the microscopic physics and only keep information regarding averages, there is no way to uniquely restore the microscopic information.

4.4.1 The renormalization group flow

This irreversibility is formally equivalent to the statement that the operations of averaging over high energy degrees of freedom, called a renormalization group transformation, constitutes a commutative monoid [1]. The general statement of the RG transform is the following [20]. Assume we are dealing with a theory parametrized by the moduli $(c^1, ..., c^m)$, which we interpret as local coordinates on a moduli space $\mathcal{M}$. We imagine that the theory is defined (in k-space) up to a cutoff Λ. The RG transform where we average over physics at large energies can

[1] Recall that a monoid is a group without inverse elements. A monoid without identity is a semi-group. These structures are discussed in the appendix.

be seen as a map

$$\begin{aligned} \mathrm{R}_{\Lambda,\mu} &: \mathcal{M} \to \mathcal{M} \\ &: \mathcal{L} \to \mathrm{R}_{\Lambda,\mu}\mathcal{L} \end{aligned}$$

where $\mu < \Lambda$ are energy scales. Here $\mathrm{R}_{\Lambda,\mu}\mathcal{L}$ is the effective theory at μ for the high energy theory at Λ. Clearly this transformation satisfies

$$\mathrm{R}_{\mu_2,\mu_3} \circ \mathrm{R}_{\mu_1,\mu_2} = \mathrm{R}_{\mu_1,\mu_3}.$$

However, as we have already mentioned, the renormalization group transformation is not reversible as microscopic information is lost in the averaging process. Hence the renormalization "group" is the action of a monoid on the space of QFTs

$$(\mathcal{L}, \Lambda) \to (\mathrm{R}_{\Lambda,\Lambda/s}\mathcal{L}, \Lambda/s).$$

where $s > 0$ is a parameter determining how much of k-space we integrate out. Formally this makes the renormalization "group" a monoid[2]. This RG transformation generates a flow on the space of moduli. Now consider a path $\mathrm{T} : [0,1] \to \mathcal{M}$, i.e. a one-parameter flow of theories $\mathrm{T}(0) \to \mathrm{T}(1)$ generated by

$$\frac{d}{dt} = -\beta^i \partial_i.$$

The vector field $\beta^i = -\partial c^i / \partial \log\mu$ on the moduli space contains information about the change of the parameters of the theory as we move along T. The vector field components are called the beta functions of the RG flow.

We consider a simple example of a scalar field. We write the scalar fields as a Fourier transform

$$\varphi(x) = \int \frac{d^n k}{(2\pi)^n} \tilde{\varphi}(k) e^{ikx}.$$

We also assume that there is some cutoff energy scale Λ that is the higher energy we think our field theory makes sense on. Let us say that we want to average over the physics in the range $\Lambda/s < |k| < \Lambda$, $s > 1$. We divide the fields into low energy modes φ_{L} with momentum below this region, and similarly high energy modes with momenta in the region. We express the integration measure over the high energy modes in momentum space as

$$\mathscr{D}\varphi_{\mathrm{H}} = \prod_{\Lambda/s<|k|<\Lambda} d\tilde{\varphi}(k).$$

[2] A monoid is a group without inverse elements, or equivalently a semi-group with identity. See the appendix for a more complete discussion on groups and monoids and their action on spaces.

The new effective action is obtained by performing this integral over high energy field modes

$$e^{-S'[\varphi;c_i]} = \int \prod_{\Lambda/s<|k|<\Lambda} d\tilde{\varphi}(k) e^{-S[\varphi;c_i]}.$$

This action in identical to the previous action as far as the large scale physic is concerned. To compare the two we rescale momenta and coordinates: $k \to sk$, $x \to x/s$. For the fields this means

$$\varphi(x) \to \varphi'(x/s) \equiv s^{\Delta}\varphi(x).$$

This relation defines the scaling dimension Δ of the fields. We assume that the action is written as a free part, plus a series of local operators

$$\sum_i c_i \int d^n x \mathcal{O}_i(x).$$

We denote the scaling of the local operators as $\mathcal{O}_i(x/s) = s^{\Delta_i}\mathcal{O}_i(x)$. Then, the additional terms in the action scale as

$$\sum_i c_i \int d^n x \mathcal{O}_i(x) \to \sum_i c'_i s^{\Delta_i - n} \int d^n x \mathcal{O}_i(x).$$

Hence we can identify $c_i = c'_i s^{\Delta_i - n}$ or equivalently $c'_i = c_i s^{n-\Delta_i}$. An iterative application of this operation produces the RG flow.

Note that the addition of a term in the action can be seen as deforming the original theory, for example we could consider a one-parameter deformation $S \to S + gS'$. However, depending on the scaling dimension of the corresponding operator, the deformed theory may represent the same large scale physics. Operators that does not contribute at large scales are called irrelevant, while those that survive the RG flow are called relevant. Operators that have scaling dimension 0 are called marginal. The same terminology applies to the parameters in front of the operator.

4.4.2 Universality classes

The fixed points of the RG flow correspond to theories invariant under change of scale. By studying the behavior of the beta function close to these fixed point we can learn about the different types of fixed points and how they relate to phases of matter. This discussion follows [14].

Consider the expansion of the beta function close to a fixed point

$$\beta^i = \dot{g}^i = \mathrm{B}^i_j(g-g_*)^j + ...$$

Working close to the fixed point, we keep only this term. Taking another derivative we get an differential equation for the beta function $\dot{\beta}^i = \mathrm{B}^i_j\beta^j$, or in a diagonal basis $\dot{\beta}^i = b_i\beta^i$. In this last expression we do not intend a sum to be taken. The corresponding equation for the coupling constant reads

$$\dot{g}^i = b_i(g-g_*)^i.$$

This can easily be solved by dividing by $(g-g_*)^i$ and integrating. We get

$$g^i(t) = g^i_* + [g^i_0 - g^i_*]e^{b_i(t-t_0)}.$$

For example, we see that if we start at the fixed point $g_0 = g_*$ we remain at the fixed point for all t. However, if we start slightly away from g_* the parameter b_i is a sort of measure of how fast we flow away from the fixed point. This parameter also controls the behavior of quantities with dimension length. Consider the scale transformation $\Lambda \to \lambda\Lambda$, $t = \log(\Lambda) \to t_0 + \log\lambda$ where $\lambda(t) = \exp(t-t_0)$. We want to know how this transformation affects a physical quantity ξ that transforms as a length

$$\xi(g(t)) = e^{-(t-t_0)}\xi(c_0).$$

Differentiating this equation and using the definition of the beta function we get the equation

$$\dot{\xi}(g(t)) = -e^{-(t-t_0)}\xi(c_0) = (\partial_i\xi(g(t)))\beta^i = (\partial_i\xi)\mathrm{B}^i_j(g-g_*)^j + ...$$

At t_0 this equation can, again in diagonal basis, be written

$$\partial_i\xi = -\frac{\xi}{b_i(g-g_*)^i}.$$

Assuming separated solution $\xi = \prod_i \xi^i$ this equation can be solved similarly to the equation for the coupling constants.

$$\xi^i(t) = \xi^i(t_0)\left[\frac{g^i(t)-g^i_*}{g^i_0-g^i_*}\right]^{-1/b_i}.$$

We see that if we start exactly at the fixed point, the correlation length diverges. However, we are more interested in the behavior close to g_*, which depends on what kind of fixed point we are dealing with. A generic renormalization group flow looks something like the below figure. A rough classification of fixed point can be made, separating them into three classes, stable ⊕, unstable ⊖ and saddle points ⊗.

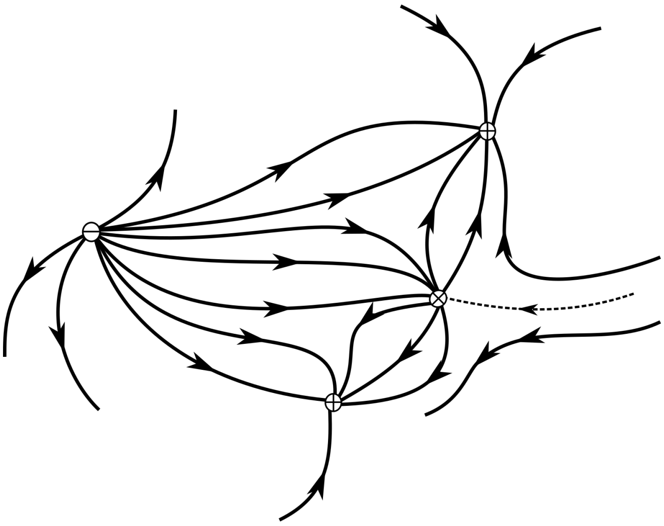


Figure 4.4

The unstable repulsive fixed points are not reached by a RG transformation, and often just play the role of sources in the flow diagram. Consider first the behavior of ξ close to an attractive stable fixed point. As we follow the RG flow, $g^i(t) - g^i_*$ becomes smaller and smaller and at the fixed point the length ξ^i vanishes. For the saddle points there are both stable and unstable directions. Imagine for example following the dotted line exactly. In this case the ξ^i vanishes as we follow the RG flow. However, this behavior is extremely unstable, as any slight departure from the dotted line will lead to a plunge into one of the attractive fixed points. In this case the correlations will grow again, before vanishing at the attractive point. A rough classification of the fixed points and their relation to phases of matter is the following.

(i) Stable fixed points $\oplus$. Theories flow to this attractive fixed point under the RG flow, and correlation lengths vanish. The theory at $\oplus$ describes a stable phase of matter, insensitive to small perturbations.

(ii) Unstable fixed points $\ominus$. Theories flow away from these repulsive points, and are not reached by RG flows.

(iii) Saddle fixed points $\otimes$. The saddle points, also called mixed points, have both stable and unstable directions. The points flowing into the saddle point span the critical surface (for example the dotted line in the figure). A theory at the critical surface will from small perturbations fall into one of the stable phases. Exactly at the critical surface, the theory flows into the saddle point, or critical point, which corresponds to critical behavior. Here the RG flow in the unstable directions start at the critical point, and the corresponding length diverges. Phase transitions correspond to crossing the critical surface.

How fast these lengths vanish or diverge is controlled again by the parameter b_i. The exponent $1/b_i = \nu_i$ is called an critical exponent, and is a part of a much larger story of exponents that we will not go into here. The basin of attraction of a given fixed point is called a universality class. Intuitively, each such class corresponds to systems that microscopically may differ substantially, but share mathematical description on large scales. In particular, near a unstable point they share critical exponents.

4.4.3 Gradient flows and the c-theorem

A famous theorem due to Zamolodchikov [89] makes the idea of information loss under RG flow a bit more precise. It states that there exists a function

$$c : \mathcal{M} \to \mathbb{R}$$

on the space of parameters that decreases monotonically under RG transforms on $\mathcal{M}$, and is a constant at the RG fixed points. This theorem is proven in the case of two dimensional QFTs. A consequence of this theorem in that the RG flow can't be completely arbitrary. For example, there can be no closed cycles in $\mathcal{M}$, which agrees with our intuition that when averaging over microscopic degrees of freedom to get a low energy theory, we should not suddenly end up at the same high energy theory. The c-function can be seen as a sort of information function on $\mathcal{M}$. In particular the value of c at a high energy fixed point is larger than at a low energy fixed point. In this sense, it is acts as a sort of ordering in this space $\mathcal{M}$ of QFTs that can be deformed into each other.

This should be compared with the so-called gradient flows. Consider a function $\Phi : \mathcal{M} \to \mathbb{R}$ and assume the space of moduli comes with a Riemannian metric G_{ij}. The RG flow is gradient if we can write a gradient formula

$$\frac{\partial}{\partial g^i}\Phi = -\mathrm{G}_{ij}\beta^j.$$

In particular, this implies that

$$\mu\frac{d\Phi}{d\mu} = -\langle \beta, \beta \rangle \leq 0.$$

Equality is here only achieved at fixed points $\beta = 0$. In this sense, the c-theorem is a weaker version of gradient flow, as gradient flows implies monotonicity while the reverse it not true. However, the intuitive picture one should have in mind is the same for Φ as it is for c.

4.4.4 Scale invariance

If we think of the renormalization group flow as a procedure where energy scales are averaged out layer by layer, it should be clear that the fixed points of the renormalization group flow corresponds to theories that are scale invariant. The quantum field theories with this symmetry are the conformal field theories [28] mentioned in the pervious chapter.

The conformal field theories should depend only on the conformal class of the Riemannian metric on spacetime. This has interesting consequences for the stress-energy tensor of the theory. Consider the variation of the action under some change in metric

$$\delta \mathrm{S} = \int d\mathrm{vol}_{\mathrm{M}} \frac{\delta \mathrm{S}}{\delta g_{\mu\nu}} \delta g_{\mu\nu}.$$

We recall that $\delta \mathrm{S}/\delta g_{\mu\nu}$ was proportional to the stress-energy tensor. For a conformal transformation the matrix transforms infinitesimally as $g_{\mu\nu} \to e^{\omega} g_{\mu\nu} = g_{\mu\nu} + \omega g_{\mu\nu} + \ldots$ with $\delta g_{\mu\nu} = \omega g_{\mu\nu}$. The change in action is then, up to constants, given by

$$\delta \mathrm{S} \sim \omega \int \mathrm{T}^{\mu\nu} g_{\mu\nu} = \omega \int \mathrm{T}^{\mu}_{\mu}.$$

This has to vanish if the theory is conformally invariant, hence $\mathrm{T}^{\mu}_{\mu} = 0$.

5

Dualities and the space of quantum field theories

This chapter discusses some big-picture ideas regarding quantum field theories and their interrelations. We discuss dualities as an abstract equivalence relation on the space of quantum field theories. We also discuss how dualities can be used to probe universality classes.

5.1 The local understanding of theory space

Recall the basic intuition behind a moduli space. After identifying discrete invariants we may still have some structure on our objects that comes in continuous families. The moduli space is supposed to represent this continuous freedom. If we view a quantum field theory as a geometric construction, we can ask questions about the moduli space of such theories. Different points in this space would correspond to certain choices of geometric data (manifolds, space of fields etc) that defines a partition function.

This chapter will be very academic, and only later will we see actual examples of the topics we here introduce. We start by imagining a space of all QFTs, where each point would give us a partition function. This large space should be divided into subspaces corresponding to symmetries, field types, space-time dimensions etc. One of the reasons why this discussion is somewhat naive is that there are non-Lagrangian QFTs which do not fit into this ’list of actions’ perspective. This will however not present any problems for our further discussions, and we restrict our attention to Lagrangian QFTs. We denote the theory space for quantum field theories $\mathcal{T}_{\text{QFT}}$.

Quantum field theory is perhaps the most successful theory we have that we at the same time do not understand. This is obviously one of the reasons why the space of quantum field theories is poorly understood - to classify some objects we need a definition of these objects. Some field theories are however quite well understood, and can be seen as subspaces $\tilde{\mathcal{T}} \subset \mathcal{T}_{\text{QFT}}$ that we do understand. For example, the classification of conformal and topological quantum field theories in low dimensions is more or less understood. In a similar way we understand some representatives of classes. For example, many parts of the four dimensional gauge theories that goes into the standard model is understood. At the very least, they are sufficiently well understood so that we are comfortable with them.

There are many natural questions that arise regarding $\mathcal{T}_{\text{QFT}}$ that do not have clear answers. For example, theory space may not be connected. Another way to say this is that not every QFT can be obtained from another by deforming it. The number of connected components would then be interesting to know as it measures how many large classes of QFTs we need to understand. One could also ask other topological or geometric questions, like what the meaning of distance would be in theory space. We will not discuss these topics, but for great discussions of theory space in general see [24].

In spite of being poorly understood globally, we have some knowledge regarding the local properties of theory space. For a given Lagrangian with some set of parameters, we can vary these parameters and span the moduli space of that theory. This is where the renormalization group flow takes place. This picture is similar to the picture perturbation theory presents [24]. Recall that in perturbation theory we deform a quadratic field action with a series of local field operators parametrized by moduli c^i in $\mathcal{M}$. Quantum amplitudes are then calculated as a series expansion in these moduli, or couplings, which are assumed to be small. Formally, any such series is at best asymptotic to the true amplitude in the limit of zero coupling, but for this discussion we will imagine that the perturbative results and the true result coincide[1]. In this sense the perturbative series of amplitudes defines the theory and we again get the picture of a parameter space $\mathcal{M}$.

From this point of view theory space consists of many disjoint patches, each of which we understand reasonably well. This is somewhat reminiscent of how a manifold is constructed, by patching together a large complicated object by many simpler objects. As in the case of manifolds we needed transition functions to go from one local realization of the space to another. The analogue to these maps are dualities, which have a natural place in the discussion of theory space.

[1] An often told semi-joke is that since any physical measurement as a consequence of uncertainties defines an open region in $\mathbb{R}$, perturbative expansions and in particular their truncated versions are the most natural tools to use in physics since we never really need anything to converge to a particular value.

5.2 Dualities

Dualities can be seen as one of the signs that we do not properly understand quantum field theory. Dualities can relate theories that lie in completely different sectors of theory space in very non-trivial ways. This is most of the times seen as a gift from the gods of physics as it expands the set of tools we have to probe a given physical system. At the same time however, it may be seen as a sign that what we tend to think of as widely different theories should really be considered the same. In this sense there may be that there it some larger theory that produces our current picture of quantum field theories and explains the weird dualities. While this may be worth noting, we will not philosophize more about this. We will however sketch how dualities appear and how they act as the glue that binds theory space together.

Consider two widely different classical field theories with different field content. For example, they could be based on completely different symmetries. On a classical level, these are not related in any way. However, as we quantize the theory by defining integration over the field space, it may be that the fields only appear as dummy variables. In other words, there is no a priori reason to believe that only one choice of fields will produce a given partition function. For our purposes, we will call two theories dual to each other if their partition functions agree, even though they classically may have different field content.

Note that there exists classical dualities as well. For example, we could call two classical theories the same of the solution space to their equations of motion are the same. Often one has a classical duality, and want to see if the duality exists also after quantization. When nothing else is mentioned we will by a duality mean the quantum version, defined as equivalence of partition functions.

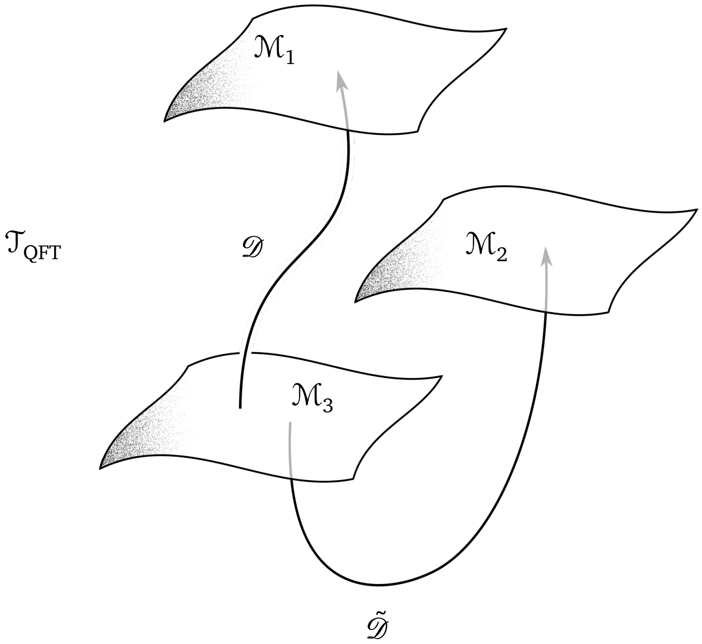


Figure 5.1: The picture of theory space provided by local moduli spaces and dualities. Here a given classical theory defines a point in theory space that is not unique. Roughly speaking, we can view the space of quantum field theories as the space of classical theories modulo these dualities.

So far our view of the QFT theory space has been the following. A point in this space corresponds to a particular theory with some moduli $(c^1, ..., c^n)$ in $\mathcal{M}$. By formally varying these parameters we trace out a connected local component of the full theory space. However, a duality transform may relate this theory to several other points in the full theory space. These theories have their own space of moduli $\mathcal{M}'$. In this way a rough outline of theory space emerges, where disjoint local components are well understood, possibly related to others by duality transformations. The important lesson is the following: A particular choice of quantized fields may not be the right way to think of quantum field theories [23].

5.3 A note on field theories and model building

Note that in some sense universality presents a kind of duality. Here different theories contain the same physics in the large distance limit, and hence are "dual" in this regime. In fact, the concepts of universality and duality presents a very powerful way of doing effective field theories phenomenologically[2].

[2] Phenomenological effective theories may should like a tautology, but we simply mean doing EFT without a formal derivation from microphysics.

Quite typically we do not want to deal with a detailed model of our system with large Hamiltonians with interacting terms as so on. Rather, we consider another theory that shares many of the central features that our original theory has, similar to how we peeled of unwanted parts of the many-body Hamiltonian in the chapter on topological matter. We then hope that these two models lie in the same universality class. In this way it is often sufficient to consider models that contain central and defining features of a given system if one is interested in universal properties. We could even pick some QFT that lies in the same universality class as the continuum limit of the microtheory. We can summarize the idea of universality and duality as it will concern us in the below figure.

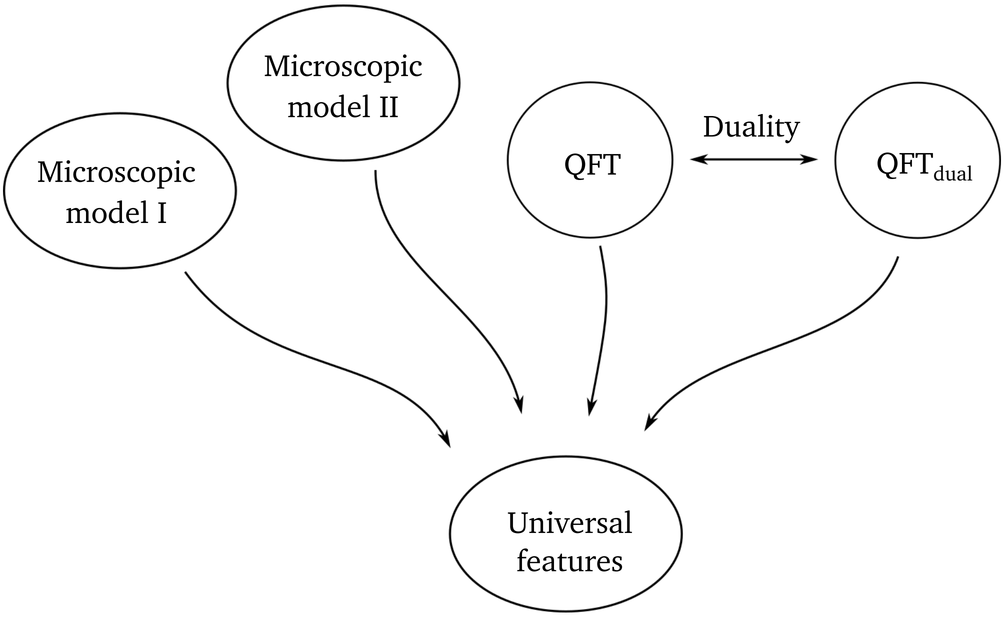


Figure 5.2

We will mainly be working in the right half of this figure, where we pick a QFT whose IR physics coincides with the original system (Microscopic model I). We should mention that by duality in this picture we mean that two QFTs are related by some non-trivial equivalence *and* the two theories lie in the same universality class. From a physical point of view this is maybe not hard to believe, as the two dual theories should describe the same physics and should do so at every scale. In this way dualities can be used to probe an universality class. At the end of the day we have a web of theories in $\mathcal{T}_{\mathrm{QFT}}$ each of which can be used to model the original system.

5.4 Target manifold dualities in sigma models

A particularly nice example of duality is that of target space dualities in nonlinear sigma models. Recall that a sigma model is defined by maps $\varphi : \Sigma \to \mathrm{M}$, where both manifolds are assumed to be Riemannian, possibly equipped with additional geometric structures. A target space duality is a map $\mathscr{D} : \mathrm{M} \to \tilde{\mathrm{M}}$ such that the sigma model remains the same. Is this way the field theory serves as an invariant that the two manifolds M and $\tilde{\mathrm{M}}$ have in common.

5.4.1 T-duality for the compact world sheet scalar

The simplest example of a target space duality is found in the sigma model $\varphi' : \Sigma \to \mathrm{S}^1_\mathrm{R}$. This is a standard textbook example and can be found it for example [37], [79] or [20].

By introducing normalized fields $\varphi = \varphi'/\mathrm{R}$ with period 2π the action can be written

$$\mathrm{S} = \mathrm{R}^2 \int d\mathrm{vol} \partial_\mu \varphi \partial^\mu \varphi .$$

As this action only depends on the derivative of the fields, it is trivially invariant under the shifts $\varphi(x) \to \varphi(x) + \omega$. If we interpret the fields as the angular coordinate on the target circle this is simply a constant shift in the angle. We will now gauge this symmetry, making $\omega \to \omega(x)$. This kind of gauge symmetry is often referred to as a non-compact U(1). As we discussed earlier, two models with different field content can often be realized to be equivalent by introducing auxiliary fields and integrating out the old fields. We will do something similar in this case. Having gauged the symmetry, the fields should be considered sections of a circle bundle over the Riemann surface Σ. We then have to introduce a connection A_μ and use the covariant derivatives $\mathrm{D}_\mu \varphi = \partial_\mu \varphi + \mathrm{A}_\mu$ [20]. Note that this derivative does not transform covariantly but rather invariantly under gauge transformations. The gauged action now reads

$$\mathrm{S} = \mathrm{R}^2 \int d\mathrm{vol} \mathrm{D}_\mu \varphi \mathrm{D}^\mu \varphi .$$

We have now introduced some additional freedom in the theory that originally was uninteresting. By introducing a Lagrange multiplier term to the action we can write the original theory as the gauged theory with an additional term

$$\mathrm{S} = \mathrm{R}^2 \int d\mathrm{vol} \mathrm{D}_\mu \varphi \mathrm{D}^\mu \varphi + \int d\mathrm{vol} f(x) \epsilon^{\mu\nu} \partial_\mu \mathrm{A}_\nu .$$

By the equation for the Fourier transform of the delta function, the field integral over this multiplier term yields

$$
\begin{aligned}
\mathcal{Z} &= \int \mathscr{D}\varphi \mathscr{D}[\mathrm{A}]\mathscr{D}f\, e^{-\mathrm{R}^2 \int d\mathrm{vol} \mathrm{D}_\mu \varphi \mathrm{D}^\mu \varphi - \int d\mathrm{vol} f(x) \epsilon^{\mu\nu} \partial_\mu \mathrm{A}_\nu} \\
&= \int \mathscr{D}\varphi \mathscr{D}[\mathrm{A}] e^{-\mathrm{R}^2 \int d\mathrm{vol} \mathrm{D}_\mu \varphi \mathrm{D}^\mu \varphi} \delta(\epsilon^{\mu\nu} \partial_\mu \mathrm{A}_\nu).
\end{aligned}
$$

This constraint is solved by writing $\mathrm{A}_\mu = \partial_\mu \theta(x)$, since then $\partial_1 \partial_2 \theta = \partial_2 \partial_1 \theta$ as desired. In other words, the multiplier implies pure gauge $\mathrm{A} = d\theta$. Then, as a classical field theory at least, the gauged action is equivalent to the original action since the gauge fields are gauge equivalent to the 0 configuration.

What we have achieved by this is a new way of writing the action of our sigma model. However, in this new form we have not one but three fields. We can then integrate out two of these fields and be left with some new action. We start by fixing the gauge to $\omega(x) = -\varphi(x)$ so that the original fields disappear. By performing a integration by parts on the multiplier term and completing the square the action can be written

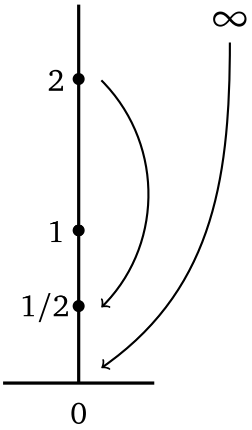
$$
\begin{aligned}
\mathrm{S} &= \int d\mathrm{vol}\ \mathrm{R}^2(\partial_\mu \varphi \partial^\mu \varphi + \mathrm{A}_\mu \mathrm{A}^\mu + 2\partial_\mu \varphi \mathrm{A}^\mu) + f \epsilon^{\mu\nu} \partial_\mu \mathrm{A}_\nu \\
&= \int d\mathrm{vol}\ \mathrm{R}^2 \left(\mathrm{A}_\mu + \frac{1}{2\mathrm{R}^2} \epsilon_{\mu\nu} \partial^\nu f \right)^2 - \int d\mathrm{vol}\ \frac{\epsilon^{\mu\nu} \epsilon_{\rho\nu} \partial^\rho f \partial_\mu f}{2\mathrm{R}^2}.
\end{aligned}
$$

Here we have ignored all overall factors, for example the determinant associated to the Faddeev-Popov procedure for fixing the gauge. See [20] for more details. We recall that Gaussian actions can be written as a determinant, so up to an overall scale factor coming from the gauge fields, the partition function for the sigma model can be written

$$
\mathcal{Z} \sim \int \mathscr{D}f\, e^{-\frac{1}{2\mathrm{R}^2} \int d\mathrm{vol}\ \partial_\mu f \partial^\mu f}.
$$

This is nothing but the original sigma model on a target circle with radius $\tilde{\mathrm{R}} = 1/\mathrm{R}$. Hence, T-duality in this case identifies a two dimensional CFT with background S^1_R with a CFT on $\mathrm{S}^1_{1/\mathrm{R}}$.

Note that the action of this theory has four parameters. The metric needs three parameters, since it is symmetric, and the target space is parametrized by only one parameter R. Consider for the sake of argument only the dimension associated with the radius, for example we could define the theory on $\mathbb{R}^2$. This is just a line $\mathbb{R}_+$ where points are reflected through the fixed self-dual point $\mathrm{R} = 1$.



This dualities act like $\mathbb{Z}_2$ transformations on the moduli space.

5.4.2 General Buscher dualities

The sigma models we consider here are the worldsheet models with B-field term. Target space dualities for these models have been studied by for example Buscher [15] and have been known for some time. The arguments are very similar to the previous example. Writing out the action we have

$$\begin{aligned} S &= \int d^2x\sqrt{g}g^{\mu\nu}h_{ij}\partial_\mu\varphi^i\partial_\nu\varphi^j + \int d^2x\epsilon^{\mu\nu}B_{ij}\partial_\mu\varphi^i\partial_\nu\varphi^j \\ &= \int d^2x\sqrt{g}g^{\mu\nu}\left\{h_{11}\partial_\mu\varphi^1\partial_\nu\varphi^1 + 2h_{1i}\partial_\mu\varphi^1\partial_\nu\varphi^j + h_{ij}\partial_\mu\varphi^i\partial_\nu\varphi^j\right\} \\ &+ \int d^2x\epsilon^{\mu\nu}\left\{B_{1i}\partial_\mu\varphi^1\partial_\nu\varphi^i + B_{ij}\partial_\mu\varphi^i\partial_\nu\varphi^j\right\}. \end{aligned}$$

Note that the latin indices now run over $2,\ldots,\dim M$. Now replace $\partial_\mu\varphi^1$ with some arbitrary vector field V_μ. This would change the theory. However, we can also add the multiplier term

$$\int d^2x\epsilon^{\mu\nu}\Phi\partial_\mu V_\nu.$$

Just as in the case with the gauge field, this would force $V_\mu = \partial_\mu\varphi^1$ and we would get back our original theory. By standard duality logic we could now in principle integrate out fields in different orders, to get dual theories. The dual sigma model for the scalar field Φ is [15] given by

$$S = \int d^2x\sqrt{\tilde{g}}\tilde{g}^{\mu\nu}\tilde{h}_{ij}\partial_\mu\Phi^i\partial_\nu\Phi^j + \int d^2x\epsilon^{\mu\nu}\tilde{B}_{ij}\partial_\mu\Phi^i\partial_\nu\Phi^j \tag{5.1}$$

and the geometric data on the target manifold is transformed to

$$\tilde{h}_{11} = \frac{1}{g_{11}}\,,\ \tilde{h}_{1i} = \frac{\mathrm{B}_{1i}}{h_{11}}\,,\ \tilde{\mathrm{B}}_{1i} = \frac{h_{1i}}{h_{11}} \tag{5.2}$$

$$\tilde{h}_{ij} = h_{ij} - \frac{h_{1i}h_{1j} - \mathrm{B}_{1i}\mathrm{B}_{1j}}{h_{11}}, \tag{5.3}$$

$$\tilde{\mathrm{B}}_{ij} = \mathrm{B}_{ij} - \frac{h_{1i}\mathrm{B}_{1j} - \mathrm{B}_{1i}h_{1j}}{h_{11}}. \tag{5.4}$$

This Buscher duality can be seen as a transformation on a manifold with Riemannian metric and 2-form

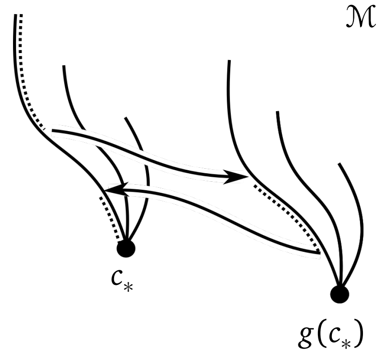
$$\mathscr{B} : \mathrm{M} \to \tilde{\mathrm{M}}$$

to yield a new but equivalent sigma model.

5.5 Duality groups and spaces of automorphic forms

A duality can be seen as a generalization of a symmetry. A normal symmetry is a transformation of the fields $\varphi \to \varphi'$ such that $\mathcal{Z}[\varphi; c^i] = \mathcal{Z}[\varphi'; c^i]$. The set of these transformations form the symmetry group G. In a similar way, a two theories are dual if $\mathcal{Z}[\varphi; c^i] = \mathcal{Z}[\varphi'; c'^i]$. This section studies a subset of these dualities where we in stead have $\mathcal{Z}[\varphi; c^i] = \mathcal{Z}[\varphi; c'^i]$. In some sense, these dualities serve as symmetries in the moduli space of a theory. In stead of relating widely different classical theories this duality maps the moduli space $\mathcal{M}$ to itself. We will assume that the set of transformations $c \to c'$ form a duality group G. When a duality group acts of the moduli space of a QFT, a seemingly trivial transformation law satisfied by the beta function opens the door to rather non-trivial mathematical theory. We here discuss this theoretical result which we later use to study universal properties of the Hall system as first discussed in [48].

Recall that the renormalization group could be seen as a monoid action on the moduli space $\mathcal{R} : \mathcal{M} \to \mathcal{M}$. Physically this represents the fact that coarse graining is a one-way street, where only information important on a given scale is conserved. In this sense, a RG transform can be seen as a special kind of "non-invertible duality" that leaves the partition function invariant as far as low-energy physics is concerned. This begs the question of how these maps on the moduli space play together. Since the duality group G presents an equivalence relation on $\mathcal{M}$ we can think of the moduli space as a quotient. Note however that different points in $\mathcal{M}$ still represents physically different points, so the quotient construction is only a mathematical trick. The map f that gets lifted to $\mathcal{R}$ by the projection $\mathcal{M} \to \mathcal{M}/\mathrm{G}$ is defined by $f([x]) = [\mathcal{R}(x)]$. We should think of this in the following way. Every stage in the renormalization group flow defines an equivalence class of points in $\mathcal{M}$ due to the duality group. In this way the road from the UV to the IR is not unique in $\mathcal{M}$. In particular, we can in stead of following a single flowline move to the G-transformed flowline and follow this for a while before moving back to the original flowline.



This implies that the RG transform and the duality commutes $\mathcal{R} = g \circ \mathcal{R} \circ g^{-1}$. Note also that fixed points are mapped to fixed points, and hence have the same universal features [14]. This is often referred to as superuniversality, as it is a sort of universality of universality classes. Again we emphasize that the points c_* and $g(c_*)$ represents physically distinct points, and should only be considered equivalent as far the RG flow in concerned. In particular, this means that the beta functions must satisfy an automorphy relation under the G action

$$\beta^i \partial_i \rightarrow \tilde{\beta}^i \frac{\partial c^j}{\partial g(c^i)} \partial_j$$

$$\therefore \;\; \tilde{\beta}^i = \beta^j \frac{\partial g(c^i)}{\partial c^j}. \tag{5.5}$$

Note also that if we see the partition function as a function of the moduli it should be a G-invariant. These are strong constraints on the functions defining a QFT, which can be of enormous help when doing "top-down" phenomenology. Note that in the case of gradient flows $\beta^i = -\mathrm{G}^{ij}\partial_j \Phi(c)$ the metric and derivative carries in total a contravariant index, and the RG potential Φ therefore has to be G-invariant. This also makes sense if we want to interpret the RG potential as a sort of information function - any two points related under G should lie on the same height in this landscape.

We will assume that an metric exists on the moduli space. Since there is an isomorphism between vectors and 1-forms in this case, we can think of the beta function as a G-invariant 1-form on $\mathcal{M}$. Equivalently, we can say that the beta function lies in the space of automorphic forms

$$\mathcal{A}(\mathrm{G}) = \Gamma(\mathrm{T}^*\mathcal{M}/\mathrm{G}).$$

The main motivation for thinking of the beta function as a covector is the connection that can be made with mathematical literature. The modular forms discussed earlier in this thesis presents the most studied case of automorphic forms. In this way we can borrow a lot of work from mathematics and directly apply it to the beta functions of our QFTs. In particular, we hope that the duality G is sufficiently strong to make the dimension of $\mathcal{A}(\mathrm{G})$ small as in the case of modular forms. If this is the case, we can write the beta functions as a linear combination of a basis this space.

Part III

Modular dualities and topological matter

6

Topological matter and the Hall phases

As we have discussed at length in previous chapters, there is not always a clear path from a microscopic model to a continuum effective theory. The main tool we will use to study the Hall effect in upcoming discussions is the automorphic constrain dualities put on the EFT beta functions. However, there are more key features of the quantum Hall system that an EFT should replicate that can be learned from studying the microphysics. The most important maybe, is the robustness of the stable phases. This is theoretically understood as a result of topological protection. This chapter aims at a clear discussion of the geometrical and topological nature of these types of topological matter.

6.1 Lattices and Brillouin zones

In the world of condensed matter and many-particle physics, most microscopic models can be cast in the schematic form

$$\mathrm{H} = \sum_{i=1}^{\mathrm{N}_e} \frac{p_i^2}{2m} + \sum_{a=1}^{\mathrm{N}_n} \frac{p_a^2}{2m_a} + \mathrm{H}_{e,n} + \mathrm{H}_{e,e} + \mathrm{H}_{n,n} + \mathrm{V}_{\mathrm{dis}}(x)$$

where the four last terms describe the electron-nucleon, electron-electron, nucleon-nucleon interactions and material disorder respectively. To deal with the wealth of systems in which condensed matter theory concerns itself, one relies heavily on the principle of emergence. As we have discussed at length, the emergent phenomena have the trait that they do not depend on all the microscopic details, i.e. they do not really depend on all the information stored in the above Hamiltonian. We know from our discussion of universality that we only need a microscopic model in the correct universality class to capture the right physics at large scales. This universality class will not depend on all the microscopic details but rather on a handful of central and defining features.

To access these classes however, we still need to trim of unnecessary parts of the above Hamiltonian. We are interested in electron systems in materials. Here the nucleons have a fixed lattice structure, and the Hamiltonian reduces to

$$\mathrm{H} = \sum_{i=1}^{\mathrm{N}_e} \frac{p_i^2}{2m} + \mathrm{H}_{e,e} + \mathrm{V}(\{x_i\}).$$

Here only the electron-electron interactions remain, as well as a potential meant to represent the electron-lattice interactions. This potential will have the same symmetry as the lattice structure. We have also ignored effects due to disorder. We will study the non-interacting electron gas moving in the potential landscape V, for which the Hamiltonian can be written

$$\mathrm{H} = \mathrm{H}_1 \otimes 1 \otimes \ldots \otimes 1 + 1 \otimes \mathrm{H}_2 \otimes 1 \ldots \otimes 1 + \ldots + 1 \otimes \ldots \otimes 1 \otimes \mathrm{H}_{\mathrm{N}_e},$$

$$\mathrm{H}_i = \frac{p_i^2}{2m} + \mathrm{V}(x_i).$$

The Hilbert space on which this Hamiltonian acts is the tensor product space $\otimes_i \mathcal{H}_i$ where each single-particle Hilbert space can be seen as $\mathrm{L}^2(\mathrm{M})$, where M is the classical configuration manifold[1]. In other words, this many-particle system is completely determined by the single-particle dynamics in the free theory. The topological phases we will discuss in this chapter can be understood by classifying these one-particle Hamiltonians according to some notion of topological equivalence. Later we will also discuss the interacting counterparts by the means of effective theories and superuniversality.

After eliminating the uninteresting parts of the original Hamiltonian, we are left with a single-particle Hamiltonian with a lattice symmetry. A Lattice is formally a discrete subgroup $\Lambda \subset \mathbb{R}^d$ that is isomorphic to $\mathbb{Z}^{\times d}$. The lattice consists of vectors $\{m^i v_i\}$ where m^i are integers and the vectors v_i are linearly independent. We will refer to the orbit of the lattice as a lattice as well. Alternatively, we can define a crystal $\mathcal{C}$ to be a subset of Euclidian space that is invariant under the lattice as a group Λ and is stationary in time. When the vectors of the lattice span the whole Euclidian space, the lattice is called a Bravais lattice [84]. There are several types of Bravais lattices in two dimensions, two of which we will meet later. These are the hexagonal and oblique lattices shown in parts a) and b) of the below figure.

[1] Although it is not the correct type of topology for topological phases, the homotopy groups of the classical configuration space M affects the quantization procedure. This is discussed in the appendix in some detail.

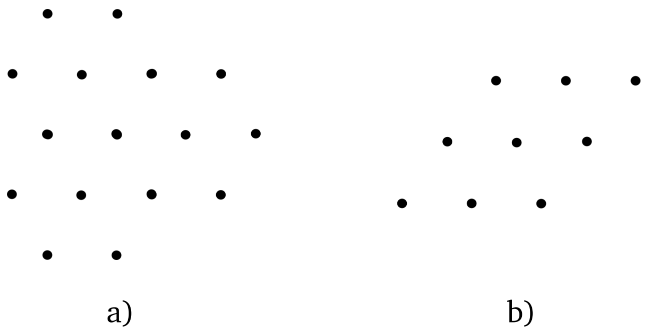


Figure 6.1

There are different ways of choosing a basis for the lattices. One standard choice for the oblique lattice is to start at, say, the lower left corner and let the vectors pointing to the two closest points be the basis. This choice of basis also affects the so called reciprocal or dual lattice. Let $w \in \Lambda$ and V be some vector in $\mathbb{R}^d$. The set of vectors {V} such that

$$\langle w, \mathrm{V} \rangle \in 2\pi\mathbb{Z}$$

span the so-called reciprocal lattice Λ^*. The fundamental region in reciprocal space is called the Brillouin zone $\mathcal{BZ}$, which is homeomorphic to a d-torus. It is worth noting that both these lattices, indeed all bravais lattices in d=2, have Brillouin zones homeomorphic to a torus when the edges are identified. The fact that several real-space lattices have similar mathematical description in the reciprocal space can be seen as a kind of universality.

Consider a particle moving in $\mathbb{R}^d$ with the symmetry of a Bravais lattice. The lattice translations are unitary representations

$$\rho : \Lambda \to \mathcal{U}(\mathrm{L}^2(\mathbb{R}^d))$$

on the Hilbert space of the system. We can realize these translation operators by $\mathrm{T_R} = \exp(i \langle \mathrm{R}, \mathrm{P} \rangle / \hbar)$ for $\mathrm{R} \in \Lambda$. To have lattice symmetry is then equivalent with the statement that $[\mathrm{H}, \mathrm{T_R}] = 0$. This is often reflected in a Λ-invariant potential landscape $\mathrm{V}(x)$. Note that as we are working in flat space $\mathrm{P}_\mu = -i\hbar\partial_\mu$.

The translational symmetry of the problem identifies points in $\mathbb{R}^d$. Effectively we can therefore take the configuration space topologically to be the d-torus

$$\mathbb{R}^d / \Lambda.$$

Clearly this space is topologically non-trivial, and has first homotopy group $\pi_1(\mathbb{T}^d) = \mathbb{Z}^{\times d}$. By the result on inequivalent first quantizations discussed in the appendix, we should construct unitary scalar representations of this group. For $\mathbb{Z}$ we can do this by $m \to \exp(imk)$ for some real number k. Generalizing to $\mathbb{Z}^{\times d}$ we must have wavefunctions satisfying

$$\mathrm{T}_\mathrm{R}\psi_k(x) = e^{i\langle k,\mathrm{R}\rangle}\psi_k(x).$$

These are called Bloch wavefunctions [84] and are sections of a complex line bundle over the real-space torus. Now pick some element $b \in \Lambda^*$ and note that

$$\langle k+b,\mathrm{R}\rangle = \langle k,\mathrm{R}\rangle + m2\pi.$$

This last factor does not change the phase, and hence k is only defined modulo the dual lattice and can be considered to lie in the Brillouin zone. A convenient parametrization of these states is $\psi_k(x) = e^{i\langle k,x\rangle}u_k(x)$ where $\mathrm{T}_\mathrm{R}u_k(x) = u_k(x)$. Often one wants to deal only with these periodic functions, and one transforms operators $\mathcal{O} \to e^{-i\langle k,x\rangle}\mathcal{O}e^{i\langle k,x\rangle}$. For example, the momentum operator transforms to $\mathrm{P} \to \mathrm{P} + \hbar k$, so the Brillouin torus consists of some kind of momentum coming from the translational symmetry of the lattice.

In the continuum limit, or at very large distance scales, the lattice points will effectively be dense in Euclidian space. In this case the k-space is simply $\mathbb{R}^d$, which we often compactify to a d-sphere $\mathbb{S}^d$. When we refer to the Brillouin zone $\mathcal{BZ}$ we mean either a d-torus or a d-sphere.

6.2 Topological insulators

6.2.1 The moduli space of gapped fermionic matter

The topology relevant for a discussion of topological matter is the topology of the space of Hamiltonians subject to certain constraints. We have already discussed the space of quantum mechanical systems when we discussed phases of matter. Here we saw that a phase could be seen as a equivalence class of points in a space parametrizing a continuous family of Hamiltonians. If observables could be evaluated everywhere along a path connecting two points, without having any singular behavior, the two points were in the same phase. In the case of topological matter, one divides the space into similar regions, where the equivalence relation now is of a topological nature.

The topological phases we are interested in are mainly the topological insulators, as these can be seen as a generalization of the simple 2-dimensional integer Hall effect. Insulators are characterized by having gaps in their Hamiltonian spectrum. Generally the spectrum obtained by $\mathrm{H}\psi_\mathrm{E} = \mathrm{E}\psi_\mathrm{E}$ can be a combination of discrete and continuous. Let E_0 denote the maximum energy of the filled states, which can either be an isolated point in the spectrum or a value in a continuous band. For energies $\mathrm{E} \geq \mathrm{E}_0$ the a spectral gap is defined by

$$\delta\mathrm{E} = \inf_{\psi \in \mathcal{H} - \mathcal{H}_{\mathrm{E}<\mathrm{E}_0}} \langle \psi, \mathrm{H}\psi \rangle - \mathrm{E}_0 = \inf_{c(\mathrm{E})} \int d\mathrm{E} |c(\mathrm{E})|^2 \mathrm{E} - \mathrm{E}_0 \geq 0$$

where $c(\mathrm{E})$ is the probability amplitude associated with ψ_E. Often we have a discrete series of eigenvalues where $c(\mathrm{E})$ is a series of Dirac delta functions at values E_n. In this case E_0 corresponds to the maximum energy level of the filled states, say E_n, and the energy gap is simply $\delta\mathrm{E} = \mathrm{E}_{n+1} - \mathrm{E}_n$. We will often assume that we can ignore very high energy levels as these will rarely be filled, and simplify the problem to the case of n filled levels and m empty levels. By the spectral theorem this means that the Hamiltonian, and correspondingly the Hilbert space, can be written as a direct sum of the filled and empty parts. An insulator corresponds to the case where the gap is not closed when an external perturbation is applied. For example, when the system is exposed to an electric field, the energy gap is not small enough so that filled levels can be excited into the empty ones. In the case of topological insulators, we do not really care about the particular energy levels of the filled and empty subspaces, and often deform these levels into single energy bands. What we do care about is the behavior of the energy gap when the Hamiltonian is deformed. As long as this gap stays open, we are studying the same (topological) type of system even though the filled or empty states get shifted around a little bit. This is the intuitive picture that will be made more presence in this chapter.

Imagine again a space where the points correspond to a choice of Hamiltonian and a symmetry group. Now consider only the Hamiltonians for gapped fermionic systems. This partitions the full moduli space into disjoint regions where one can continuously move without closing the energy gap. This space of gapped systems will be denoted $\mathcal{M}$. These regions corresponds to different topological phases. In fact, just counting the number of these regions is an interesting question. This corresponds to studying the zeroth homotopy group $\pi_0(\mathcal{M})$ of the moduli space. There are three main classes of topological insulators. These are sketched in figure 6.2, where the white areas correspond to value of parameters for which the Hamiltonian is gapped. The gray areas correspond to gapless systems.

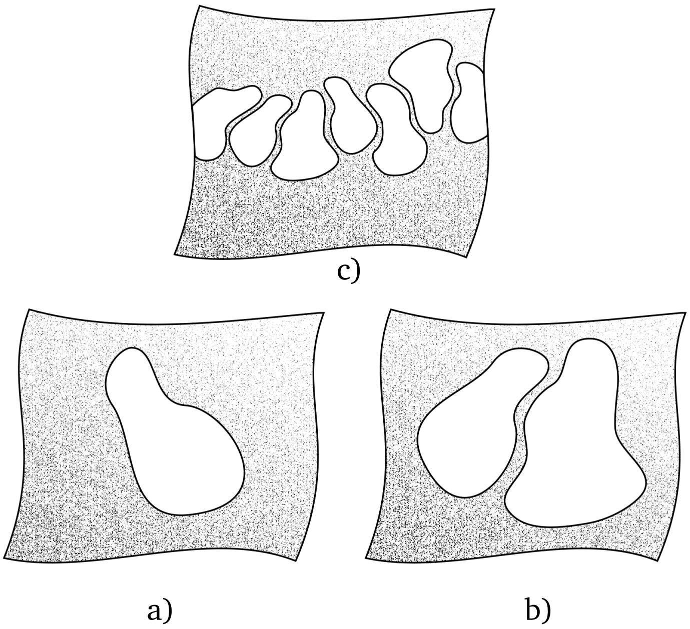
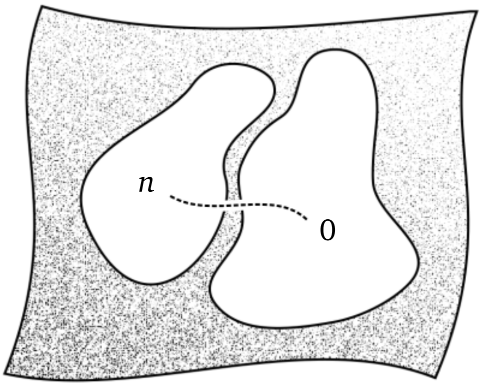


Figure 6.2: a) A single topological phase. b) Two topologically distinct phases. c) An infinity of topologically distinct phases.

The simplest case we can imagine is when the moduli space only has one connected component corresponding to gapped Hamiltonians. This is often called the trivial insulator or the vacuum. If we think of the QED vacuum, this can be seen as a simple insulator with two energy bands corresponding to electrons and positrons. The gap is associated with pair production. In the second case b) there are two such regions that can only be crossed by entering a gapless region of the parameter space. This corresponds to a topological phase transition. These are the so called $\mathbb{Z}/2\mathbb{Z}$ topological insulators, labeled by two numbers $0, 1$. In some sense, there is one topological phase distinct from the vacuum. The case most relevant to us is the $\mathbb{Z}$-type insulators where there is an infinity of connected regions in the parameter space. These phases can be labeled by integers, which for example will be the same integers as in the integer Hall effect.

Topological insulators turns out to be system with some intrinsic holography, in the sense that the information we need is contained on the surface of the system. This is called the bulk-boundary correspondence or the bulk/edge duality. We here briefly sketch why such a duality exists. Imagine placing two topological insulators on top of another. For simplicity we consider an trivial insulator (vacuum) on top of a topological insulator with phases labeled by integers n. Now imagining moving continuously from the non-trivial insulator to the trivial one. At the boundary between the two, something strange has to happen. As we can not, by definition, interpolate continuously between the two topological classes a topological transition occurs.



In the space of Hamiltonians this corresponds to moving from one region of gapped Hamiltonians to another, passing trough a region of gapless Hamiltonians. When the energy gap closes, electrons can move from the occupied insulating bands to the empty conducting bands. In this sense, the number n can be seen as a way of counting the number of conducting states moving on the (d-1)-dimensional edge of the d-dimensional topological insulator.

6.2.2 Homotopy approach and the 10-fold way

While we will meet different types of topological insulators as we go along, the class that contains the quantum Hall system will be of most focus. We here briefly discuss all classes to gain an overview. We consider a (single particle) Hamiltonian H describing a topological insulator in d spatial dimensions. In the case of a periodic system the Brillouin zone is a torus $\mathbb{T}^d$, or in the case of a system in the continuum a sphere $\mathbb{S}^d$. The Brillouin zone $\mathcal{BZ}$ acts as a parameter space in the sense that the Hamiltonian is given by maps

$$\mathcal{BZ} \ni k \to \mathrm{H}(k).$$

In this case it is the homotopy of the maps from $\mathcal{BZ}$ to the space of Hamiltonians that is of interest. Note that as we are interested only in the topology, we can deform the Hamiltonian to a flat-band insulator where the filled and empty bands are collapsed into single bands of energy $+1$ above and -1 below the energy gap. If P is the projector onto filled states, we can define the spectral flattened Hamiltonian [73]

$$\mathcal{Q} = (+1)(1-\mathrm{P}) + (-1)\mathrm{P} = 1 - 2\mathrm{P}.$$

In the topological classification we can equivalently study the homotopy of maps $k \to \mathcal{Q}(k)$, i.e. the homotopy classes of $\mathcal{Q}$'s.

If we wanted to carry out the classification scheme of topological insulators, we would need the spaces to which the spectral flattened Hamiltonians $\mathcal{Q}(k)$ belongs. These space can be found by considering what constraints generic symmetries put of the Hamiltonian [69]. A symmetry is associated with a unitary or anti unitary representation of a group G. For the unitary representations $\rho : \mathrm{G} \to \mathcal{U}(\mathcal{H})$ we can decompose the representation into irreducible representations. This Block diagonalizes the Hamiltonian, where each block corresponds to a particular irreducible representation. These symmetries put no interesting constraint on the Hamiltonian, as we can work in each block individually. The constraints we want come from the anti-unitary symmetries of time reversal symmetry and charge conjugation [69].

Time reversal is represented by the operator $\mathcal{T} = \mathrm{T}\mathcal{K}$ where $\mathcal{K}$ is complex conjugation and T is unitary [73]. Invariance of the Hamiltonian under $\mathcal{T}$ means $\mathrm{TH}^*\mathrm{T}^{-1} = \mathrm{H}$. Since time reversal maps $x \to x$, $k \to -k$, this can be written for Bloch Hamiltonians as $\mathrm{TH}^*(k)\mathrm{T}^{-1} = \mathrm{H}(-k)$. Similarly charge conjugation $\mathcal{C} = \mathrm{C}\mathcal{K}$ leads to the constraint $\mathrm{CH}(k)\mathrm{C}^{-1} = -\mathrm{H}(-k)$ [73]. Both these operators have the property that their square should not change the quantum state. Hence $\mathcal{T}^2 = \mathrm{TT}^* = e^{i\alpha}$. Since T is unitary, we can write $\mathrm{T} = e^{i\phi}\mathrm{T}^{\mathrm{T}}$. However, $\mathrm{T}^{\mathrm{T}} = (e^{i\phi}\mathrm{T}^{\mathrm{T}})^{\mathrm{T}}$. Combining these equations yields $\mathrm{T} = e^{i2\phi}\mathrm{T}$, so $e^{i\phi} = \pm 1$. Similar arguments can be given for charge conjugation. The combined symmetry $\mathcal{S} = \mathcal{T}\mathcal{C} = \mathrm{TC}^*$ is also a possibility we must consider. This symmetry is either realized $(+1)$ or not (0). Since time reversal and charge conjugation can be either absent or realized in two different ways, we write $+1, -1, 0$ for the three possibilities. Under time reversal and charge conjugation there are 9 ways the Hamiltonian can respond. In the $\mathrm{T} = \mathrm{C} = 0$ case we there is still a possibility of having $\mathcal{S}$ symmetry realized in one of two ways. This yields $(9-1)+2 = 10$ ways the Hamiltonian can transform. These 10 symmetry classes are often called Altland-Zirnbauer classes, and are given different names according to their connection with the so-called Cartan symmetric spaces [69]. For example, the case where no symmetries are present is called the A class and the case with time reversal realized by $\mathcal{T}^2 = -1$ is called AII. These are the 10 different classes of spectral flattened Hamiltonians $\mathcal{Q}(k)$ we should consider.

Consider the case where there are no symmetry constraints. With n filled and m empty energy bands, the effective Hamiltonian $\mathcal{Q}(k)$ is a $\mathrm{U}(n+m)$ matrix. However, there is a residual freedom in the form of a $\mathrm{U}(n)$ rotation amongst the filled states, and $\mathrm{U}(m)$ amongst the empty states. Hence we should consider $\mathcal{Q}(k)$ as a part of the Grassmannian

$$\mathrm{Gr}_{n,n+m}(\mathbb{C}) = \mathrm{U}(m+n)/\mathrm{U}(m) \times \mathrm{U}(n).$$

The different topological phases are then characterized by the homotopy of the maps $\mathcal{BZ} \to \mathrm{Gr}_{n,n+m}(\mathbb{C})$. In the case of a continuum model when $\mathcal{BZ}$ is a sphere the homotopy types are simply $\pi_d(\mathrm{Gr}_{n,n+m}(\mathbb{C}))$. However, in the toroidal case some case must be takes as lower dimensional homotopy groups $\pi_{d-s}(\mathrm{Gr}_{n,n+m}(\mathbb{C}))$ for $s = 1, ..., d$ can contribute to the homotopy type of $\mathbb{T}^d \to \mathrm{Gr}_{n,n+m}(\mathbb{C})$. The second homotopy group of this Grassmannian is $\mathbb{Z}$, while the first homotopy group is trivial [73]. Hence in two spatial dimensions, the different topological phases are labeled by integers. This class of topological insulators is one we will study in some detail. In particular, when we discuss the vector bundle approach shortly we will see that in every even spatial dimension there exist such integer topological phases. The integer quantum Hall effect is a realization of this class of topological insulators in two dimensions. Here the (k-independent) Landau levels acts as flat bands. Strong magnetic fields correspond to a large energy gap separating the lowest Landau level from the higher. The holographic picture of topological insulators where charge carriers move along the edge is in the IQHE attributed to electrons skipping against the wall in a magnetic field.

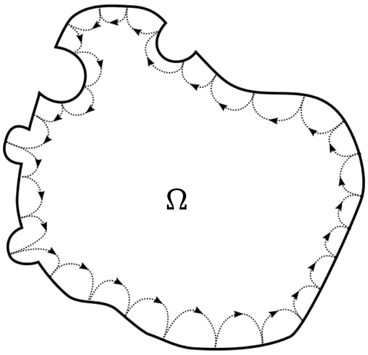


Figure 6.3: Topological Hall insulator in a sample Ω with electrons forced to skip along the edges due to the orthogonal magnetic field. These charge carriers have a fixed chirality and are insensitive to impurities.

When some of the discrete symmetries are present, additional constraints are put on the Hamiltonian and on the spectral flattened Hamiltonian $\mathcal{Q}(k)$. Since we will be dealing mainly with the Hall-like systems, i.e. the A class, we will not go into detail regarding the other classes of topological insulators. The constraints on $\mathcal{Q}$ can be found for example in table III of [70]. By computing the homotopy group of maps from the Brillouin zone into the different spaces of spectral flattened Hamiltonians, one can find a periodic table of topological insulators [73].

AZ class	$\mathcal{T}$	$\mathcal{C}$	$\mathcal{S}$	d = 0	d=1	d = 2	d = 3
A	0	0	1	$\mathbb{Z}$	0	$\mathbb{Z}$	0
AIII	0	0	1	0	$\mathbb{Z}$	0	$\mathbb{Z}$
AI	+1	0	0	$\mathbb{Z}$	0	0	0
BDI	+1	+1	1	$\mathbb{Z}_2$	$\mathbb{Z}$	0	0
D	0	+1	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}$	0
DIII	-1	+1	1	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}$
AII	-1	0	0	$\mathbb{Z}$	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$
CII	-1	-1	1	0	$\mathbb{Z}$	0	$\mathbb{Z}_2$
C	0	-1	0	0	0	$\mathbb{Z}$	0
CI	+1	-1	1	0	0	0	$\mathbb{Z}$

There are various approaches to understanding this table and its patterns. Perhaps best understood is the K-theoretic approach [43]. We will discuss the K-theory approach to the A class.

6.3 Vector bundle approach

The above approach uses homotopy theory to classify maps from the Brillouin zone into the correct space of effective Hamiltonians with the right symmetry properties. Here each topological phase is associated with a homotopy class as these enumerate the deformation classes of $\mathcal{Q}(k)$'s. An alternate approach is based on vector bundles. For every $k \in \mathcal{BZ}$ we have a Hamiltonian with eigenspace $\mathcal{H}$. By the spectral theorem we can decompose the Hamiltonian and the associated Hilbert space into a direct sum $\mathcal{H} = \mathcal{H}_- \oplus \mathcal{H}_+$ of filled and empty energy levels. Since this makes sense all over $\mathcal{BZ}$ by the gap hypothesis, the Hilbert spaces constitute a vector bundle $\mathcal{E}$ over the Brillouin zone which has the decomposition into subbundles

$$\mathcal{E} = \mathcal{E}_- \oplus \mathcal{E}_+$$

corresponding to filled and empty bands. Since the Brillouin zone generally needs to be covered by more than one coordinate chart, the vector bundle of filled states may very well be non-trivial. Since the bundle is ultimately constructed using eigenspaces of the Hamiltonian, we may expect that the deformation classes of Hamiltonians coincide with the topological classes of vector bundles. This motivates the definition of a topological phase in the vector bundle approach.

DEFINITION: *A topological phase of non-interacting gapped fermionic matter in $d+1$ spacetime dimensions is a phase that is labeled by the topological classes of a Hilbert bundle of filled states $\mathcal{E}_-$ over the d dimensional Brillouin zone $\mathcal{BZ}$.*

We will see that the classification based on this definition coincides with the homotopy classification we discussed for the A-type topological insulators.

6.3.1 A note on stable equivalence and topological K-theory

The periodic table of topological insulators is as mentioned perhaps best understood in the K-theoretic language. We will not go into too much detail regarding this , but we would like to show why K-theory is the right tool for the job. The classification of vector bundles that goes into the above table is in general a very hard problem. However, one can solve an easier problem where one relaxes the notion of equivalence. As it turns out we can solve the harder problem in the case of A-type insulators, but there is still some nice physics in the more K-theoretic approach.

Let us denote by $\mathbb{K}$ either $\mathbb{R}$ or $\mathbb{C}$, and let $\mathrm{M}\times\mathcal{T}^{\ell}$ denote the rank ℓ trivial bundle over M. If two bundles $\mathcal{E}$, $\mathcal{F}$ are isomorphic after they have been augmented by trivial bundles

$$\mathcal{E}\oplus\mathcal{T}^{i}\approx\mathcal{F}\oplus\mathcal{T}^{j}$$

the two bundles are called stably equivalent [61]. This equivalence is denoted $\mathcal{E}\sim_{s}\mathcal{F}$. Now, let $\mathcal{G}'$ be a bundle so that $\mathcal{G}\oplus\mathcal{G}'=\mathcal{T}^{k}$. Then if we have $\mathcal{E}\oplus\mathcal{G}\approx\mathcal{F}\oplus\mathcal{G}$ we get by taking the direct sum with $\mathcal{G}'$ on both sides

$$\mathcal{E}\oplus\mathcal{T}^{k}\approx\mathcal{F}\oplus\mathcal{T}^{k}.$$

Hence $\mathcal{E}\sim_{s}\mathcal{F}$. This idea of stable equivalence, being less strict that homeomorphisms, may seem to give a weaker classification of topological insulators. However, this augmentation by trivial bundles makes the stable classification more fitting. We should be able to add trivial bands to our insulator Hamiltonian without changing the topological type, as the two coincide after a spectral flattening. This corresponds exactly to a direct sum by trivial bundles [73].

Let $\mathrm{Vec}_{k}(\mathrm{M},\mathbb{K})$ denote the set of all $\mathbb{K}$ vector bundles of rank k over M. The sum

$$\mathrm{Vec}(\mathrm{M},\mathbb{K})=\oplus_{k}\mathrm{Vec}_{k}(\mathrm{M},\mathbb{K})$$

consists of all vector bundles over a manifold M. Note that the direct sum operation

$$\oplus:\mathrm{Vec}_{k}(\mathrm{M},\mathbb{K})\times\mathrm{Vec}_{\ell}(\mathrm{M},\mathbb{K})\rightarrow\mathrm{Vec}_{k+\ell}(\mathrm{M},\mathbb{K})$$

maps the set of vector bundles into itself. Hence $\mathrm{Vec}(\mathrm{M},\mathbb{K})$ is a commutative monoid, i.e. a semi-group with identity. The identity can be seen as taking the direct sum with $\mathrm{Vec}_{0}(\mathrm{M},\mathbb{K})$. In general, such monoids can be extended into abelian groups. Let $\mathcal{M}$ be such a monoid, and consider on the product space $\mathcal{M} \times \mathcal{M}$ an equivalence relation defined by

$$(m_1, n_1) \sim (m_2, n_2)$$

$$\text{if } \; m_1 + n_2 + k = m_2 + n_1 + k$$

for some $k \in \mathcal{M}$. If we imagine for a moment that subtraction would be defined, this would read $m_1 - n_1 = m_2 - n_2$. Hence we should think of the element (m, n) and a formal version of $m - n$ [61].

The natural numbers (including 0) under addition is an example of a commutative monoid. The above construction then entails thinking of equivalence classes in $\mathbb{N} \times \mathbb{N}$. Two pairs of natural numbers (a, b) and (c, d) are equivalent if $a + d = c + b$. For example $(1, 3)$ would be equivalent to (a, b) only for $(a, b) = (a, a + 2)$. Hence we have an equivalence class

$$[(1, 3)] = \{(0, 2), (1, 3), (2, 4), \ldots\}.$$

Since we think of (a, b) as $a - b$ the pair $(1, 3)$ should correspond to -2 in some formal sense. This element is the additive inverse of $(3, 1)$, which would correspond to 2. Similarly we can define the other positive and negative numbers. In this way, we have constructed $\mathbb{Z}$ as an abelian group under addition.

An analogous procedure as the one just carried out for the natural numbers for vector bundles yields the so-called K-theory for M [61]. The quotient

$$\mathrm{K}(\mathrm{M}) = \mathrm{Vec}(\mathrm{M}, \mathbb{K}) \times \mathrm{Vec}(\mathrm{M}, \mathbb{K}) / \sim$$

is called the K-group of M, and the equivalence is again of the form

$$(\mathcal{E}_1, \mathcal{F}_1) \sim (\mathcal{E}_2, \mathcal{F}_2)$$

$$\text{if } \; \mathcal{E}_1 \oplus \mathcal{F}_2 \oplus \mathcal{G} \approx \mathcal{E}_2 \oplus \mathcal{F}_1 \oplus \mathcal{G}.$$

An element in K(M) is denoted $[\mathcal{E} - \mathcal{F}]$.

The virtual dimension of an element of K(M) is defined as $d_v[\mathcal{E} - \mathcal{F}] = \mathrm{rk}\mathcal{E} - \mathrm{rk}\mathcal{F}$. The elements of K(M) with $d_v = 0$ defines the subgroup $\tilde{\mathrm{K}}(\mathrm{M})$, called the restricted K-theory of M [61]. This is the part of K-theory relevant for topological insulators. Let us define a map

$$\varphi : \mathrm{Vec}_k(\mathrm{M}, \mathbb{K}) \to \tilde{\mathrm{K}}(\mathrm{M})$$

$$: \mathcal{E}_k \to [\mathcal{E}_k - \mathcal{T}^k].$$

Now assume that $\varphi(\mathcal{E}_k)$ and $\varphi(\mathcal{F}_\ell)$ are equal in K(M), i.e.

$$\mathcal{E}_k \oplus \mathcal{T}^\ell \oplus \mathcal{G} \approx \mathcal{F}_\ell \oplus \mathcal{T}^k \oplus \mathcal{G}.$$

From our above discussions we know that this implies the stable equivalence $\mathcal{E}_k \oplus \mathcal{T}^\ell \sim_s \mathcal{F}_\ell \oplus \mathcal{T}^k$. By the definition of stable equivalence this simply means that $\mathcal{E}_k$ and $\mathcal{F}_\ell$ are stably isomorphic. One can also show the converse, that given two stably isomorphic bundles over M one gets a single element of $\tilde{\mathrm{K}}(\mathrm{M})$. A rather simple proof of this can be found in [61]. In this way, we have that the stable isomorphism classes of bundles correspond to an element of the restricted K-group of M. Since we identify topological insulators with stable equivalence classes of bundles, the different topological phases are, in the continuum case, described by $\tilde{\mathrm{K}}(\mathbb{S}^d)$.

6.3.2 Class A topological insulators in even dimensions

The filled bundle $\mathcal{E}_-$ belonging to the class A topological insulators has $\mathrm{U}(n)$ structure group. To classify these topological insulators we need the isomorphism classes of unitary vector bundles over $\mathcal{BZ}$, or rather stable equivalence classes if we do not care about trivial bundles. The resticted K-group for unitary vector bundles is denoted $\tilde{\mathrm{K}}\mathrm{U}(\mathrm{M})$, and can be shown to be related to the homotopy theory of $\mathrm{U}(r)$ when M is a sphere. Specifically [61]

$$\tilde{\mathrm{K}}\mathrm{U}(\mathbb{S}^d) \approx \pi_{d-1}(\mathrm{U}(r))$$

when $\mathrm{rk}\mathcal{E}_- = r \geq \frac{d}{2} + 1$. It is known [61] that these (restricted) K-groups have the isomorphisms

$$\tilde{\mathrm{K}}\mathrm{U}(\mathbb{S}^d) = \mathbb{Z} \ , \ d \in 2\mathbb{N},$$
$$\tilde{\mathrm{K}}\mathrm{U}(\mathbb{S}^d) = 0 \ , \ d \in 2\mathbb{N} + 1.$$

Hence the class A topological insulators are labeled by integers in every even spatial dimension.

As mentioned, we can solve the stronger classification problem in the case of A-type insulators, i.e. for unitary bundles. This can be done by associating a topological invariant, namely the Chern numbers, to each topological class of bundles. We recall from our discussion of Chern forms and Chern characters that we need a connection on the bundles. Quite generally, when we have a quantum system defined over a parameter space $\mathcal{M}$ the natural connection to consider on the associated bundle over $\mathcal{M}$ is the Berry connection.

Formally, the situation we will discuss is equivalent to a classical Yang-Mills theory on the parameter space $\mathcal{M}$, with U(r) gauge group. Although the physics is somewhat different, we will use the same terminology. For more detailed discussions on Berry connection and Berry curvature see for example [18].

Consider a r-fold degenerate quantum mechanical system with the moduli in the Hamiltonian spanning the space $\mathcal{M}$. As before, the Hamiltonian is assumed to be a smooth function of these parameters. We denote the degenerate states as φ^a, where a runs over $1,\ldots,r$. We can view these states as the basis sections of a rank r vector bundle $\mathcal{E}$ over the moduli space. We would like to know how the system evolves when the variations in time stem only from variations in the external parameters, i.e. following continuous paths in $\mathcal{M}$. At t=0 we set $\psi^a(0) = \varphi^a(\lambda(0))$, and by the adiabatic theorem we will stay in the degenerate subspace [18] as time progresses. At a later time we can then write

$$\psi^a = (\mathrm{U}^{-1})^a_b \varphi^b.$$

Note that by demanding that $\partial_t \langle \psi^a, \psi_b \rangle = 0$ the matrix U must be in U(r). Inserting the above rotated state into the Schrodinger equation, we can find an equation for the matrix U, which will evolve in time through the variations of the parameters. This yields a first order differential equation [18]

$$(\dot{\mathrm{U}}^{-1})^a_b = -(\mathrm{U}^{-1})^a_b \dot{\lambda}^k \left\langle \partial_k \varphi^b, \varphi_b \right\rangle - i\mathrm{E}(\lambda)(\mathrm{U}^{-1})^a_b.$$

We define $(\mathrm{A}_k)^b_c = -\left\langle \partial_k \varphi^b, \varphi_b \right\rangle$ as a U(r)-valued 1-form over the moduli space. This is the connection needed to perform covariant differentiation with respect to these parameters. Since we assumed that we started in one of the degenerate states, the unitary matrix has to start as the identity matrix. The above equation can then be solved to get [18]

$$\mathrm{U}^{-1}(t) = e^{\int \mathrm{A} - i\int_0^t ds \mathrm{E}(\lambda(s))}.$$

The second factor is of course only the dynamical phase factor, while the first is a geometric one. Here we are integrating the 1-form over the one dimensional subspace that is the path in $\mathcal{M}$.

The geometric Berry phase by itself will not be important to us, but rather its geometric content. First of all, note that if we reparametrize the degenerate states $\tilde{\varphi}^a = \mathrm{U}^a_b \varphi^b$, the connection transforms as

$$\begin{aligned}
(\tilde{\mathrm{A}})^a_c &= -\left\langle \partial_k [\mathrm{U}^a_b \varphi^b], \mathrm{U}^d_c \varphi^c \right\rangle && (6.1)\\
&= \left\langle \mathrm{U}^a_b \varphi^b, [\partial_k \mathrm{U}^d_c] \varphi_d \right\rangle && (6.2)\\
&= (\mathrm{U}^{-1})^a_b (\mathrm{A}_k)^b_d \mathrm{U}^d_c + (\mathrm{U}^{-1})^a_b \partial_k \mathrm{U}^b_c. && (6.3)
\end{aligned}$$

Hence the connection transforms a a proper Yang-Mills field over the moduli space. Note that in the non-degenerate case, we are dealing with a line bundle over $\mathcal{M}$ with U(1) gauge group. This would be analogous to a classical theory of electrodynamics. Just as in the field theory setting, we can construct the curvature 2-form, or field strength, by $F = dA + A \wedge A$. In the line bundle case this is simply dA and the curvature in components read

$$(F_{ij})^a_b = \langle \partial_i \varphi^a, \partial_j \varphi_b \rangle - \langle \partial_j \varphi^a, \partial_i \varphi_b \rangle .$$

Using the Berry connection and curvature we can define the Chern forms and Chern characters for the Bloch bundle of filled states. The total Chern character reads $ch_0(\mathcal{E}) + ch_1(\mathcal{E}) = \mathrm{rk}(\mathcal{E}) + c_1(\mathcal{E})$. These are the two topological numbers that classify the bundle. For fixed rank, the corresponding Chern number

$$C_k(\mathcal{E}_-) = \int_{\mathcal{BZ}} ch_k(\mathcal{E}_-) \in \mathbb{Z}$$

is a topological invariant of this bundle, and classifies the A class topological insulators in $d = 2k$ spatial dimensions. Recall from our discussion on Chern cohomology that by deforming a bundle with a trivial bundle the Chern numbers remain the same. Hence the number that classifies a type A topological insulator in even spatial dimension is not sensitive to the addition of a trivial bundle.

If we place a class A topological insulator with Chern number C_k next to the trivial vacuum insulator with Chern number 0, we see that C_k necessarily measures the number of massless charge carriers on the $(d-1)$-dimensional boundary of the system. This is essentially the content of the famous TKNN formula for the 2-dimensional Hall effect [76]. We discuss this relation between vector bundle topology and conductivity next.

6.4 Topological conductivity

By now we have seen various ways to think of topological insulators. Either we can study the homotopy of Grassmannians or we can look at the topology of vector bundles over the k-space. In even dimensions the A-type topological insulators are classified by integers. In fact, this integer can be shown to correspond to the conductivity in the two dimensional case. This work is well known and yields the celebrated Thouless-Kohmoto-Nighingale-den Nijs (TKNN) formula for conductivity [76], where the Chern number of the bundle of filled states is identified with the off diagonal part of the conductivity tensor. There are two main elements of this story. First is the Kubo formula, a result from linear response theory where the change in an observable due to external perturbations is related to statistical 2-point functions. Second is the Chern cohomology just mentioned.

6.4.1 Linear response and the Kubo formula

In linear response theory one studies first order responses of a system to external perturbations. We briefly discuss the result known as the Kubo formula in a general setting as this plays a key role in relating the Hall conductivity with the Chern numbers. We follow [78] rather closely. Consider a quantum system with Hamiltonian H_0 and a set of observables $\mathcal{O}^i$. We want to study the response of the system to an external perturbation of the form

$$\mathrm{H}_0 \to \mathrm{H} = \mathrm{H}_0 + \delta\mathrm{H}(t) = \mathrm{H}_0 + f_i\mathcal{O}^i.$$

The main assumption of linear response is that the change in any expectation value of an observable is linear in the sources f_i. Explicitly

$$\delta\left\langle \mathcal{O}^i(t)\right\rangle = \int d\tau \chi_{ij}(t,\tau) f^j.$$

The function χ_{ij} is contains the information regarding the change in expectation value in the presence of a source, and is called the response function. The Kubo formula is a general formula relating the response function to statistical 2-point correlation functions.

We consider a state $|\mathcal{S}_0\rangle$ in the asymptotic past $t \to -\infty$, which in the interaction picture of quantum mechanics evolves in time as

$$|\mathcal{S}(t)\rangle = \mathrm{U}\,|\mathcal{S}_0\rangle = e^{\frac{i}{\hbar}\int_{t_0}^{t}\delta\mathrm{H}(\tau)d\tau}\,|\mathcal{S}_0\rangle\,.$$

We pick an observable $\mathcal{O}_j$. To first order in the perturbation $\delta\mathrm{H}$, the expectation value of this observable is

$$\left\langle \mathcal{O}_j(t)\right\rangle = \left\langle \mathcal{O}_j(t)\right\rangle_0 + \frac{i}{\hbar}\int_{-\infty}^{t} d\tau \left\langle [\mathcal{O}_j(t), \delta\mathrm{H}(\tau)]\right\rangle$$

which can easily be obtained by Taylor expanding the time evolution operator. Hence the change in expectation value as a result of the external perturbation is

$$\delta\left\langle \mathcal{O}_j(t)\right\rangle = \frac{i}{\hbar}\int d\tau \left\langle [\mathcal{O}_j(t), \mathcal{O}_i(\tau)]\right\rangle f^i(\tau).$$

This is the Kubo formula for the linear response. By comparison with our assumption of linear response, the response function can be identified.

For a system of charged particles, there will be a collective response when a electric field is applied. The electric field creates a current J_μ that describes the collective motion of charged particles trough space. For a single particle with charge q we can take the current to be simply $J^\mu = qv^\mu$. The response of such systems to electromagnetic fields is contained in the conductivity tensor $\sigma^{\mu\nu}$. We consider the applied electric field as a perturbation $\delta H = -J^\mu A_\mu$. We also consider an alternating field $E_\mu \exp(-i\omega t)$, following [80]. The gauge field can then be written $A_\mu = (i\omega)^{-1} E_\mu \exp(-i\omega t)$. Using this in the Kubo formula, we find the response

$$\langle J^\mu \rangle = \frac{1}{\omega\hbar}\int_0^\infty dt' e^{i\omega t'} \left\langle [J^\mu(0), J^\rho(t')] \right\rangle E_\rho e^{-i\omega t}.$$

We have here introduced a new time coordinate $t' = t - \tau$. We have also assumed that the expectation value of the current in the unperturbed system vanishes. From this expression we can easily read of the response tensor

$$\sigma^{\mu\rho} = \frac{1}{\omega\hbar}\int_0^\infty dt'' e^{i\omega t''} \left\langle [J^\mu(0), J^\rho(t'')] \right\rangle$$

where current operator transforms in time as $J^\rho(t) = e^{iH_0 t/\hbar} J^\rho(0) e^{-iH_0 t/\hbar}$. We will now insert a complete set of energy eigenstates of the unperturbed Hamiltonian, which we call $|n\rangle$. The conductivity tensor becomes

$$\begin{aligned} \sigma^{\mu\rho} = \frac{1}{\omega\hbar}\sum_n \int_0^\infty & dt e^{i\omega t} [\langle \mathcal{S}|J^\mu(0)|n\rangle \left\langle n|e^{iH_0 t/\hbar} J^\rho(0) e^{-iH_0 t/\hbar}|\mathcal{S}\right\rangle && (6.4) \\ & - \left\langle \mathcal{S}|e^{iH_0 t/\hbar} J^\rho(0) e^{-iH_0 t/\hbar}|n\right\rangle \langle n|J^\mu(0)|\mathcal{S}\rangle]. && (6.5) \end{aligned}$$

The unperturbed Hamiltonian now acts on the eigenstates and gives simply the exponential of the energies. We assume also that the state in the asymptotic past is a energy state with energy $E_\mathcal{S}$. The time integral is then easily performed, and we find

$$\sigma^{\mu\rho} = -\frac{i}{\omega}\sum_n \left\{ \frac{\langle \mathcal{S}|J^\rho|n\rangle \langle n|J^\mu|\mathcal{S}\rangle}{\hbar\omega - E_n + E_\mathcal{S}} - \frac{\langle \mathcal{S}|J^\mu|n\rangle \langle n|J^\rho|\mathcal{S}\rangle}{\hbar\omega + E_n - E_\mathcal{S}} \right\}.$$

For ease of notation we introduce $a^{\rho\mu} = \langle \mathcal{S}|J^\rho|n\rangle \langle n|J^\mu|\mathcal{S}\rangle$ and $\delta E = E_n - E_\mathcal{S}$. In this notation:

$$\sigma^{\mu\rho} = \frac{i}{\omega}\sum_n \left\{ \frac{a^{\mu\rho}}{\hbar\omega + \delta E} - \frac{a^{\rho\mu}}{\hbar\omega - \delta E} \right\}. \tag{6.6}$$

In the presence of a magnetic field, the conductivity must be antisymmetric as a consequence of an Onsager relation. The response is in our case of the form

$$\int dt \left\langle [\mathcal{O}_j(0), \mathcal{O}_i(t)] \right\rangle.$$

Under time reversal we expect the microscopic physics to be invariant. By letting $t \to -t$ and then performing a time translation $+t$ in time, the response changes to

$$\int dt \left\langle [\mathcal{O}_i(0), \mathcal{O}_j(t)] \right\rangle,$$

so the response should be symmetric in i and j. This makes it seem like the conductivity should be symmetric. However, a magnetic field is also flipped by time reversal [66]. An electron close to the edge of the Hall sample will be forced to move along the edge in a direction dictated by the magnetic field. Hence, a flip of the magnetic field should yield an additional change of sign in the Hall conductivity. It is this anti-symmetric contribution we are interested in calculating.

We therefore anti-symmetrize the above expression to get

$$\sigma^{\mu\rho} = \frac{i}{2\omega}\sum_n \left\{ \frac{a^{\mu\rho} - a^{\rho\mu}}{\hbar\omega + \delta \mathrm{E}} - \frac{a^{\rho\mu} - a^{\mu\rho}}{\hbar\omega - \delta \mathrm{E}} \right\} = i \sum_n \hbar \frac{(a^{\mu\rho} - a^{\rho\mu})}{(\hbar\omega)^2 - \delta \mathrm{E}^2}.$$

Note that this expression now only holds for the off-diagonal Hall components. Remember that we are interested in the infinite period limit of this case, so the denominator simplifies. The final expression in the DC limit reads

$$\sigma^{\mu\rho} = i\hbar \sum_n \left\{ \frac{\langle \mathcal{S}|\mathrm{J}^\mu|n\rangle \, \langle n|\mathrm{J}^\rho|\mathcal{S}\rangle}{(\mathrm{E}_\mathcal{S} - \mathrm{E}_n)^2} - \frac{\langle \mathcal{S}|\mathrm{J}^\rho|n\rangle \, \langle n|\mathrm{J}^\mu|\mathcal{S}\rangle}{(\mathrm{E}_\mathcal{S} - \mathrm{E}_n)^2} \right\}.$$

This is the Kubo formula for conductivity. Note that the antisymetrization we did saved us for divergences when taking the $\omega \to 0$ limit.

To make use of this formal expression for the conductivity, we need an expression for the current. Recall that a particle moving freely in $\mathbb{R}^n$ has the general wave function

$$\psi(r,t) = \int \frac{d^n k}{(2\pi)^n} \tilde{\psi}(k,0) e^{i(\omega t - k_\mu r^\mu)}$$

where $\omega = \frac{\hbar}{2m} k_\mu k^\mu$. By Taylor expanding around an arbitrary point

$$\omega = \omega_0 + \partial_\mu \omega_0 (k - k_0)^\mu + \ldots$$

we can write the wavefunction

$$\psi(r,t) = e^{i(\omega t - \partial_\mu k_0^\mu t)} \int \frac{d^n k}{(2\pi)^n} \tilde{\psi}(k,0) e^{i(k_\mu r^\mu - \partial_\mu \omega_0 k^\mu t)}.$$

Hence, the distribution $|\psi|$ moves with a velocity

$$v_\mu = \frac{\partial \omega_0}{\partial k^\mu}.$$

Recalling the relation between ω and the Hamiltonian we can write the velocity as an operator

$$v_\mu = \frac{1}{\hbar} \frac{\partial \mathrm{H}}{\partial k^\mu}.$$

When the particle is charged we can introduce the single-particle current as this velocity multiplied by its charge.

6.4.2 Conductivity as the bundle slope

The Kubo response formula is the key to a geometric understanding of the conductivity tensor. We will consider the Kubo formula applied to the states in the bundle over the Brillouin zone, first with no degeneracy. We replace the asymptotic state $|\mathbb{S}\rangle$ with the states of filled bands $u_\alpha(k)$ and replace $|n\rangle$ with empty bands $u_\beta(k)$. The sum over energy levels will be replaced with a sum over the more general band index α, β. We write

$$\sigma_{xy} = \sigma_{\mathrm{H}} = i\hbar \sum_{\alpha,\beta} \int_{\mathcal{BZ}} \frac{d^2k}{(2\pi)^2} \left\{ \frac{\langle u_\alpha, \mathrm{J}_y u_\beta \rangle \langle u_\beta, \mathrm{J}_x u_\alpha \rangle}{(\mathrm{E}_\alpha - \mathrm{E}_\beta)^2} - \frac{\langle u_\alpha, \mathrm{J}_x u_\beta \rangle \langle u_\beta, \mathrm{J}_y u_\alpha \rangle}{(\mathrm{E}_\alpha - \mathrm{E}_\beta)^2} \right\}.$$

Here we identified the conductivity with the conductivity averaged over the torus, which turns out to be a crucial step. We make a comment on this shortly. Using the equation for current in terms of the Hamiltonian, and noting that

$$\left\langle u_\alpha, \frac{\partial \mathrm{H}'}{\partial k^y} u_\beta \right\rangle = (\mathrm{E}_\alpha - \mathrm{E}_\beta) \left\langle \frac{\partial}{\partial k^y} u_\alpha, u_\beta \right\rangle$$

the conductivity reads

$$\sigma_{\mathrm{H}} = i\frac{e^2}{h} \sum_{\mathrm{E}} \int_{\mathbb{T}^2_{\mathscr{B}}} \frac{d^2k}{2\pi} \Big\{ \langle \partial_y u_\alpha, u_\beta \rangle \langle u_\beta, \partial_x u_\alpha \rangle - \langle \partial_x u_\alpha, u_\beta \rangle \langle u_\beta, \partial_y u_\alpha \rangle \Big\}.$$

Inserting the completeness relation $\sum_\alpha |u_\alpha\rangle \langle u_\alpha| + \sum_\beta |u_\beta\rangle \langle u_\beta| = 1$ the formula reduces to

$$\sigma_{\mathrm{H}} = -i \sum_\alpha \int \frac{d^2k}{2\pi} \left(\langle \partial_y u_\alpha, \partial_x u_\alpha \rangle - \langle \partial_x u_\alpha, \partial_y u_\alpha \rangle \right)$$

where we work in units of e^2/h, and remove this overall constant. This is nothing but the integral over the curvature of the (line) bundle over the torus

$$\sigma_{\mathrm{H}} = \sum_\alpha i \int \frac{d^2k}{2\pi} \mathrm{F}^{(\alpha)}_{xy}$$

which corresponds to a sum of Chern numbers over the filled bands.

Why we should integrate over the torus is not obvious, but turned out to be crucial. We attempt a explanation following [88]. Consider a single filled band with Chern number

$$c_1 = \frac{i}{2\pi} \int d^2k \mathrm{F}.$$

Consider a finite system on a lattice of size $\ell_1 \times \ell_2$ with $\ell_1 = n_1 \omega_i$, where ω_i are the two lattice vectors. With periodic boundary conditions the momentum takes the usual discrete values $p_i = m_i 2\pi/\ell_i$ where i refers to one of the two spatial directions. When a U(1) gauge field is present we should use the gauge covariant derivative $\mathrm{P}_i = p_i + \mathrm{A}_i = -i\partial_x + \mathrm{A}_i$. When we translate a state around a loop ℓ_i it should at most change by a phase, so we have

$$e^{i\mathrm{P}_i \ell_i}\psi = e^{i\alpha_i}\psi$$

$$\therefore \mathrm{A}_i = \alpha_i/\ell_i$$

where α_i are defined modulo 2π. We can think of these parameters as parameters of the Hamiltonian and vary then adiabatically. This yields yet another torus $\tilde{\mathbb{T}}^2$ [88]. Note that setting α_i to 2π yields a momentum

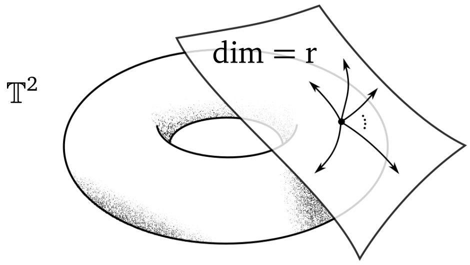
$$\mathrm{P}_i = m_i \frac{2\pi}{\ell_i} + \frac{2\pi}{\ell_i}$$

which is equivalent to a shift $m_i \rightarrow m_i + 1$. In the momentum space lattice, (α_1, α_2) parametrized a unit cell. Equivalently, for a pair of momentum numbers (m_1, m_2) the "angles" (α_1, α_2) parametrized a cell. Just as before we can define a connection $\tilde{\mathrm{A}}$ over $\tilde{\mathbb{T}}^2$ and define a Chern number

$$\tilde{c}_1 = \frac{i}{2\pi}\int_{\tilde{\mathbb{T}}^2} d^2\alpha \tilde{\mathrm{F}}.$$

However, since the total contribution for every allowed momentum state is equivalent to integrating over the full Brillouin torus $\mathbb{T}^2$, we have $\tilde{c}_1 = c_1$ [88]. From this point of view, the argument that we should integrate over the torus is more believable - we should average over α_i as all phases $e^{i\alpha_i}$ are equivalent.

We will now assume that we are dealing with a fixed band, or rather energy level, with r-fold degeneracy. This corresponds to a rank $\mathrm{rk}(\mathcal{E}) = r$ bundle over the torus.



In addition to averaging over the torus, we should now also average over the contributions from each of the r degenerate states [65]

$$\sigma_{\mathrm{H}} = \sum_{i=1}^{r} \frac{1}{r} \int \frac{d^2k}{2\pi} \mathrm{F}_{xy}^{(i)}.$$

From our above discussions of the Chern numbers of different types of bundles, it should be clear that this conductivity is (proportional to the) the Chern number of this rank r bundle. In fact, it is the ratio

$$\sigma_{\mathrm{H}} = \mu(\mathcal{E}) = \frac{c_1(\mathcal{E})}{\mathrm{rk}(\mathcal{E})} \in \mathbb{Q}.$$

This topological parameter is called the slope of the bundle. While the averaging over r-fold degenerate states leading to the fractional conductivity was mentioned in some of the original work of Thouless *et al* [65], it was discussed as the topological slope in Varnhagens work in the mid 90s [83]. A mathematical review of the Varnhagen paper [31] also contains some information regarding the slope. The slope plays an important role in the classification of stable vector bundles. A vector bundle is called stable if for a non-trivial subbundle $\mathcal{F}$ the slopes satisfy $\mu(\mathcal{F}) < \mu(\mathcal{E})$. In [31] it is found that the relevant vector bundle is stable for odd ranks. An interesting question that we will not go into is whether there is a relation between the geometric idea of stability and stability in a physical sense.

6.5 Effective topological quantum field theory

An alternative way of approaching the topological phases described above is by using topological quantum field theories. According to common lore, the low-energy effective field theory for topological phases is a topological QFT (TQFT). We recall our approach to effective field theories. Given a microscopic model, say a lattice system, we assume it has a continuum limit. An effective field theory we think of as a QFT that share the properties of this continuum field theory in the low-energy regime. More generally, the pick a QFT in the same universality class as our system, which contains the universal large-scale properties of our system. We here briefly discuss the role TQFTs play in the theory topological matter, before briefly discussing the simplest example, namely the Chern-Simons theory describing the class A insulators.

We have seen that at the end of RG flows sits conformal field theories. These are QFTs defined on a manifold M which are invariant under scale transformations of the metric. More precisely, in a classical field theory the scale invariance is in some sense enhanced to a conformal invariance. In this way, the effective low-energy QFT is sensitive only to a conformal equivalence class of metrics. Roughly speaking these theories care less about geometric structures and their partition function can be seen as a number labeling the conformal class of M .

In our present case we are dealing with systems with a finite energy gap in their spectrum, separating the ground states from the excited states. If we assume that this gap persists in the continuum limit, the low-energy effective theory at a scale below the gap has no dynamical content. These field theories are the topological field theories where the Hamiltonian vanishes. Under a change of the metric on spacetime M, the action of a QFT transforms

$$\delta S = \int_M \frac{\delta S}{\delta g_{\mu\nu}} \delta g_{\mu\nu} \sim \int_M T^{\mu\nu} \delta g_{\mu\nu}.$$

Hence theories that are invariant under any deformations of the metric have zero stress-energy tensor, and in particular a zero Hamiltonian. The partition functions of these theories are topological invariants of the manifold M. In this sense, the TQFT provides a topological invariant that can be used to label our system. The classification of topological phases should in this sense be related to the classification of TQFTs. It is in regards to this that the formal functorial definition of a QFT, and a TQFT in particular, is important - to classify TQFTs one needs a formal definition of what a TQFT is before one can define a notion of equivalence. We will not go into this any further, but the interested reader can see for example [30] [29].

The low-energy effective field theory for charge transport in the $d = 2$ A-type topological insulators is a U(1) Chern-Simons theory [69]. It can be shown that in this two dimensional case the level k of the theory is in fact the Hall conductivity. We will not go to much into the Chern-Simons description, but we quickly discuss the most important aspects, following [73].

Let us start with pure $d = 2 + 1$ U(1) Chern-Simons theory defined by the Lagrangian

$$\mathcal{L} = \frac{k}{2} \int d^3x \epsilon^{\mu\nu\rho} A_\mu \partial_\nu A_\rho.$$

Since the electromagnetic current and the gauge field couples like $A_\mu J^\mu$, the current can be obtained by a functional differentiation of the action with respect to the gauge field. This yields [80] the current

$$J^i = \frac{\delta S}{\delta A_i} = \frac{\partial \mathcal{L}}{\partial A_i} - \frac{\partial}{\partial A^\ell} \frac{\partial \mathcal{L}}{\partial \partial_\ell A_i},$$

$$J_1 = \sigma^{12} E_2 = k E_2,$$

so we have $\sigma_H = k \in \mathbb{Z}$. However, more interesting physics appear if we include a coupling between the current and gauge field. Consider a system of N electrons [73]. The position and velocity densities can be written

$$j_0 = \sum_{n=1}^{N} \delta(x - x_n),$$

$$j_i = \sum_{n=1}^{\mathrm{N}} v_n \delta(x - x_n).$$

These may be combined into the 3-vector $j^\mu = (j^o, j^i)$, which has to be conserved $\partial_\mu j^\mu = 0$ due to charge conservation. This conservation is automatic if we write

$$j^\mu = \frac{1}{2\pi} \epsilon^{\mu\nu\rho} \partial_\nu a_\rho$$

where a_ρ can be interpreted as a emergent U(1) gauge field. We can imagine having a field theory with spinors and electromagnetic gauge fields, and integrating out the spinors and high energy degrees of freedom until we get a effective theory for the a fields. The effective field theory Lagrangian takes the form [73]

$$\mathcal{L} = \frac{m}{4\pi} \epsilon^{\mu\nu\rho} a_\mu \partial_\nu a_\rho - \frac{e}{2\pi} \epsilon^{\mu\nu\rho} \mathrm{A}_\mu \partial_\nu a_\rho + \ldots$$

where the second term if the coupling $j^\mu \mathrm{A}_\mu$ between the current and electromagnetic gauge fields A_μ. The Euler-Lagrange equations for this theory can be shown to be

$$-e\mathrm{J}^\lambda = -\left(\frac{e^2}{2\pi m} \right) e^{\mu\sigma\lambda} \partial_\sigma \mathrm{A}_\mu.$$

Since the constant of proportionality between the electromagnetic current and the electric field is the conductivity, we can read of $\sigma_\mathrm{H} = e^2/2\pi\hbar m$ where we reintroduced Plancks constant. Since m appears in from of the Chern-Simons term in the effective Lagrangian we know that it takes integer values as discussed in earlier chapters. This reproduces the fractional $1/m$ phases.

Note first that since the Chern form is defined in any even dimensions, the Chern-Simons theory exists in odd dimensional space-times. In this way this field theoretic framework agrees with the first quantized result that topological insulators classified by integers exists in every even spatial (and hence odd space-time) dimension. In addition we have seen that the effective topological field theory holds information regarding the interacting fractional phases. In the first quantized approach the only way of obtaining fractional conductivities was to introduce a higher rank bundle over the Brillouin zone, corresponding to degenerate states.

7

Dirac matter and modular fixed points

Closely related to topological insulators are the phases referred to as Dirac matter. Here a finite number of crossing points where the energy gap closes exist in the Brillouin zone. The effective degrees of freedom in this type of system are Dirac fermions, hence the materials name. We here discuss these phases and their relation to the IR fixed points of field theories with modular dualities. We should note that some authors include topological insulators as Dirac matter since their edge modes are massless Dirac fermions. We will here use the name for systems with such massless modes in their bulk Brillouin zone.

7.1 Nielsen-Ninomiya theorem

We have seen that topological insulators correspond to different topological classes of the vector bundle $\mathcal{E}_- \to \mathcal{BZ}$ of filled states. This means that for every k in the Brillouin zone, we have to be able to identify a proper vector subspace $\mathcal{H}_-$ of the full Hilbert space $\mathcal{H}$. Essentially this means that the gap between the n and the $(n+1)$ level has to be well developed globally on the Brillouin zone $\mathcal{BZ}$. However, there can be special points in the Brillouin zone where the gap closes, rendering the filled sub bundle ill defined. We discuss the physics of these points following [88] quite closely.

Consider a simple 2-band model with Hamiltonian $\mathrm{H}(k) = h^i(k)\sigma_i + h_0(k)\mathbb{I}$. This is clearly determined by four parameters: four parameters goes into the four-vector $h^\mu = (h^0, h^i)$. If the parameters are adjusted so that the two eigenvalues coincide, the Hamiltonian is diagonal $\mathrm{H}(k) = \mathrm{E}(k)\mathbb{I}$ and is determined by a single parameter. Hence we need to adjust the three remaining parameters to make the energy levels cross. For a band model with more bands the number of parameters in the Hamiltonian is larger, and so more of these must be varied. In a general setting however, at least three parameters have to be varied [88]. This means that in 2+1 dimensions, when the $\mathcal{BZ}$ in two dimensional, we typically do not expect crossings since we do not have enough parameters. However, in the presence of certain symmetries such points can occur. To get a feeling for these special points in the Brillouin zone, we discuss the famous Nielsen-Ninomiya theorem in 3+1 space-time dimensions before discussing the analogous points in 2+1 dimensions.

We assume that the Bloch Hamiltonian $\mathrm{H}(k)$ is gapped in the three dimensional Brillouin zone $\mathcal{BZ}$ except at a collection of points $\{k_*^{(i)}\}$. For future reference we let D_i denote a small ball containing $k_*^{(i)}$. Removing these regions yields a "punctured" Brillouin zone $\mathcal{BZ}-\cup_i\mathrm{D}_i$. On this space, the filled sub bundle in well-defined. Consider again a two-band model with Hamiltonian $\mathrm{H}(k) = h^i(k)\sigma_i$. Let us on $\mathcal{BZ}-\cup_i\mathrm{D}_i$ define the map

$$\phi : k^i \to n^i(k) = \frac{h^i(k)}{|h|}.$$

Here the vector n^i spans a 2-sphere. In particular, we can restrict this map to a sphere S^2 enclosing a point $k_*^{(i)}$ in the Brillouin zone.

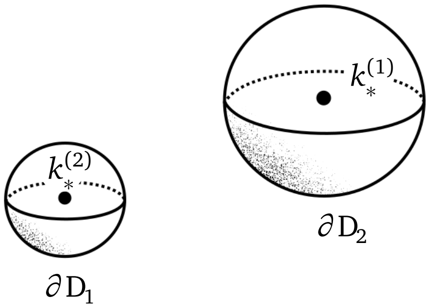


Figure 7.1

These spheres can be seen as the boundaries of the balls D_i. This restricted map is then a homotopy map $\phi : \mathrm{S}^2 \to \mathrm{S}^2$ where the target sphere is parametrized by the vector $n^i(k)$. The different homotopy sectors of this map is labeled by the integers $\pi_2(\mathrm{S}^2) = \mathbb{Z}$. This winding number can be expressed as a pullback [2][88] of the volume form on the target sphere

$$w = \int_{\mathrm{S}^2} \phi^* d\mathrm{vol}.$$

Since the volume form is a top dimensional form, it clearly vanishes under the exterior derivative. Using Stokes theorem and the fact that the pullback commutes with exterior derivatives we have

$$0 = \int_{\mathcal{BZ}-\cup_i\mathrm{D}_i} \phi^* d\, d\mathrm{vol} = \sum_i \int_{\mathrm{S}^2} \phi^* d\mathrm{vol} = \sum_i w_i$$

where we used that $\partial(\mathcal{BZ} - \cup_i D_i) = \cup_i S^2$. Hence the sum of all winding number must vanish. This is the geometric statement of the Nielsen-Ninomiya theorem [88].

As a brief example we can consider the Dirac Hamiltonian where $h^i = \pm k^i$. The eigenvalues are $\pm|k|$, and hence the gap is $2|k|$. The signs in the Hamiltonian corresponds to positive and negative chirality states. The vector parametrizing the target sphere for the map ϕ is now

$$n^i = \pm\frac{k^i}{|k|}.$$

The point where the Hamiltonian becomes gapless is of course $k = 0$. We take the unit sphere to enclose this point. The map $\phi : k^i \to \pm k^i$ then corresponds to the identity map with winding number ± 1. Clearly the sum of these winding numbers vanish. What we see here is quite generic [88]. Each point $k_*^{(i)}$ typically have winding numbers ± 1, corresponding to massless Weyl fermions of positive or negative chirality. The Nielsen-Ninomiya theorem then states that there are equally many fermion states of positive and negative chirality. Somewhat more loosely, one can say that the fermions come in pairs.

In 2+1 dimensions the story is slightly different, since we generically expect no energy crossing. Note that in general we can consider an effective theory close to the minimum gap points. Since we can deform the Hamiltonian within its topological class, we can treat the system as a two-band model.

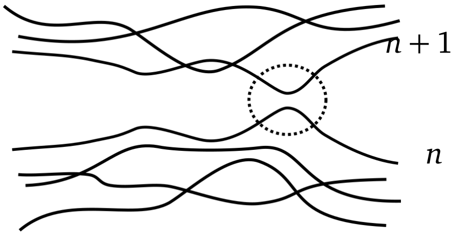


Figure 7.2

The Hamiltonian for the two relevant bands read $\mathrm{H} = h^i(k)\sigma_i$, where we dropped the constant energy shift $h^0\mathbb{I}$. Let k_* denote a Dirac point. Expanding the function $h^i(k)$ we get the effective Hamiltonian

$$\mathrm{H}(k) = \left.\frac{\partial h^i(k)}{\partial k^j}\right|_{k*}(k^j - k_*^j)\sigma_i = \partial_1 h_*^i \sigma_i \delta k^1 + \partial_2 h_*^i \sigma_i \delta k^2.$$

This is essentially a massless Dirac Hamiltonian. Using the third Pauli matrix a mass term $m\sigma_3$ can be added. However, this mass term can be shown to break time reversal and reversal symmetries [88]. On a generic Dirac Hamiltonian $\mathrm{H} = k^i \sigma_i$ the time reversal operator acts by [27]

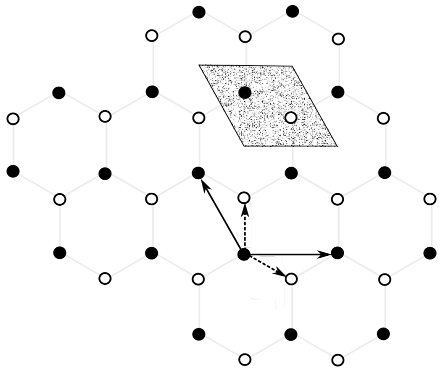
$$\mathrm{TH}(k)\mathrm{T}^{-1} = i\sigma_2 \mathrm{H}^*(-k) i\sigma_2 = k^1\sigma_1 + k^2\sigma_2 - m\sigma_3.$$

Hence the mass term breaks T-symmetry. Similarly we have under parity that [27]

$$\mathrm{RH}(k)\mathrm{R}^{-1} = \sigma_1(k^1\sigma_1 - k^2\sigma_2 + m\sigma_3)\sigma_1 = k^1\sigma_1 + k^2\sigma_2 - m\sigma_3.$$

Hence the mass term also breaks this symmetry. Another way to state these observations is that in the presence of discrete symmetries massless Dirac fermions can appear. In this way points in k-space where the gap closes can be protected by discrete symmetries. This happens for example in graphene. For more detailed discussions on Dirac Hamiltonians and the relation to topological phases see [88].

Graphene is what one sometimes refers to as a Bravais lattice with basis. The Bravais lattice is a hexagonal lattice, where each point has associated with it two neighboring points. In physical terms we can think of the hexagonal Bravais lattice as the discrete symmetry group of a crystal that translates a whole neighborhood.



The hexagonal lattice is spaces by the black dots in the figure. For every black dot there are two white dots. The Bravais lattice has two basis vectors (black solid), and the dotted vectors takes you from the Bravais lattice to the neighboring points. Note that there are several ways of choosing such bases. The fundamental cell is shown in gray, and contains two atoms. Alternatively, we can view graphene as the combination of two triangular lattices, one black and one white.

The fact that graphene can be seen as two triangular lattices, often referred to as the A and B sublattices, is also clear on the group-theoretic level. Consider one of the hexagons in graphene spanned by three black and three white points. This polygon has some discrete symmetries. Any n-polygon in the plane has discrete rotational symmetries $\mathscr{R}_m = \exp(2\pi i m/n)$. These are generated by $\mathscr{R} = \exp(2\pi i/n)$ and constitutes a cyclic group of order n, isomorphic to $\mathbb{Z}^n$. Similarly, there are n lines of reflection. The reflections constitute n cyclic groups of order 2. These groups of rotations and reflections are called Dihedral groups $\mathfrak{Dih}_n$. In the case of the hexagon in graphene, the symmetry group is the order 12 group

$$\mathfrak{Dih}_6 = \langle e, \mathscr{R}, \mathscr{R}^2, ..., \mathscr{R}^5, \sigma_1, ..., \sigma_6 \rangle$$

where the σ_j's are reflections satisfying $\sigma_j^2 = 1$. This group is known to be isomorphic to $\mathfrak{Dih}_3 \times \mathbb{Z}_2$. Here $\mathbb{Z}_2$ flips between the black and white triangular sublattices, while the order 6 Dihedral group is the symmetry group of the 3-gons.

A simple model for graphene is a 2-band model, analogous to our discussion in the section on the Nielsen-Ninomiya theorem. For graphene we can take as basis vectors $a_1 = (\sqrt{3}, 1)$ and $a_2 = \frac{1}{2}(\sqrt{3}, 3)$, and for the k-space $b_1 = \frac{2\pi}{3}(\sqrt{3}, -1)$ and $b_2 = \frac{4\pi}{3}(0,1)$ [73]. The Brillouin zone is a hexagon in k-space. The 2-band model with Hamiltonian $\mathrm{H}(k) = h^i(k)\sigma_i$ often considered citestanescu has components

$$h^1(k) + ih^2(k) = -t(1 + e^{ik^i a_{1i}} + e^{ik^i a_{2i}})$$

while the third component vanishes in the presence of time reversal and reflection symmetry [73]. The components of $h(k)$ can be shown to vanish at the points

$$\mathrm{K} = \frac{4\pi}{3}(1/\sqrt{3}, 0)$$

$$\mathrm{K}' = \frac{2\pi}{3}(1/\sqrt{3}, 1)$$

which lie at the corners of the Brillouin zone [73]. These points are also called valleys. The discrete reflection symmetry we discussed earlier maps these Dirac points into each other. Hence the massless fermions close to the points K, K' remain massless as long as this symmetry is not broken [88]. Recall also that the energies associated with the Hamiltonian $h^i\sigma_i$ are $\pm|h|$, and hence are invariant under rotations of the vector h. By the discrete rotations $e^{2\pi i/6}$ the point K is mapped to K', and hence the two lie on the same energy circle in k-space.

7.2 Conductivity of Dirac matter in $d = 2 + 1$

We would here like to consider the charge transport properties of a system with n_f flavors of fermions corresponding to n_f Dirac points. With a mix of phenomenology and theory we will end up with a formula for the Hall conductivity which we will relate to the RG fixed points of an effective QFT based on modular dualities.

We imagine a 2+1 dimensional system with points $k_*^{(i)}$, $i = 1, ..., n_f$ in the Brillouin zone $\mathcal{BZ}$ where the Hamiltonian $\mathrm{H}(k)$ vanishes. We also imagine these points to be protected by symmetries as in the case of graphene. We know that since the dispersion relation is linear close to the points $k_*^{(i)}$ the low-energy effective theory is a massless Dirac theory

$$\mathrm{S} = \int d^3x \overline{\psi}^a (i\gamma^\mu \partial_\mu)\psi_a$$

where $a = 1, ..., n_f$ runs over the flavors associated with the different special points $k_*^{(i)}$. By the Nielsen-Nanomiya theorem these points come in pairs, and n_f should be an even number. Now imagine breaking T-symmetry. This allows the fermions to become massive, adding a mass term $m_a \overline{\psi}^a \psi_a$ to the above Lagrangian. For example, the Dirac fermions in graphene are known to become massive when the graphene layer is grown on a substrate [27] [4]. In this case the mass term breaks also the chiral (sub lattice) symmetry. While the dispersion is no longer perfectly conical, there is still an even number of fermion flavors. Coupling this theory to an external electromagnetic field, we get the action [27]

$$\mathrm{S} = \int d^3x \overline{\psi}^a (i\gamma^\mu \partial_\mu - m_a - e\gamma^\mu \mathrm{A}_\mu)\psi_a.$$

In the low-energy limit, this theory is described by a gauge theory without fermions (as they have been "integrated out") with a Chern-Simons term [27]. For each fixed flavor index a the contribution to the conductivity is, in dimensionless units,

$$\sigma_{\mathrm{H}}^a = \frac{1}{2}\mathrm{sgn}(m_a).$$

In this way, the low-energy Dirac modes contributes with $1/2$ to the Chern number[1]. Due to the Nielsen-Ninomiya theorem there are an even number of such contributions which agrees with our previous discussions that the Chern number should be integer.

Recall that in the case of a higher rank vector bundle $\mathcal{E}$ over the Brillouin torus, the Kubo formula lead to fractional conductivity $\sigma_{\mathrm{H}} = c_1(\mathcal{E})/\mathrm{rk}(\mathcal{E})$. In order to describe the observed Hall phases, the denominator should be an odd number $2m + 1$. In the presence of a magnetic field, the low-energy fermions can occupy relativistic Landau levels. In contrast to the non-relativistic Landau

[1] There is much that should be said regarding materials with Dirac points, e.g. generalizations of graphene. For example, the pairs of Dirac points have opposite mass sign but also apposite Chern number so that the conductivities add. Since our discussions will become more and more phenomenological we will not discuss this, but the interested reader can see for example [27] or [73]

levels there is now a zero mode present in the massless case [4]. We imagine adiabatically turning on a magnetic field, so that this lowest level is adiabatically connected to the fermion state contributing $1/2$ to the conductivity [4]. We write the numerator of the Hall conductivity n or $n+1/2$ depending on whether or not we have Dirac points. In a first approximation we can also include spin just by taking degeneracy into account, which contributes with an overall factor of 2. Generically we will then be dealing with two classes of conductivities. With no Dirac points, we have a conductivity of the form

$$\sigma_{\mathrm{H}}(d_s) = \frac{d_s n}{2m+1}, \tag{7.1}$$

where d_s is either 1 or 2 depending on the presence or absence of spin degeneracy. When the low-energy degrees of freedom are Dirac fermions we have the conductivity

$$\sigma_{\mathrm{H}}(n_f, d_s) = \frac{n_f d_s(n+1/2)}{2m+1}. \tag{7.2}$$

Here n_f should be an even integer according to Nielsen-Ninomiya. For example, for 2 Dirac points and spin degeneracy taken into account, we get the series $4(n+1/2)/(2m+1)$. The integer series (i.e. m=0 series) $4(n+1/2)$ is the well known Hall effect in graphene. Similarly the case with no Dirac points and no spin degeneracy corresponds to the standard spin-polarzed hall effect. It could be interesting to note that since the IQHE corresponds to a subset of 2-dimensional topological insulators the above formula partitions this set into classes labeled by (d_s, n_f). As we discuss next, this is equivalent to a partition labeled by certain $\mathrm{SL}_2(\mathbb{Q})$ subgroups. Note that we do not claim this partition to be exhaustive.

7.3 The 2-parameter modular field theories

Recall our discussion on effective field theories and universality. If we can find a field theory where the low-energy physics coincides with our system, we think of them as being in the same universality class. The stable phases of a system we imagined as the attractive fixed points of a renormalization group flow. In the present case these attractive fixed points should be the Hall conductivities derived above, as these are the topological invariants (Chern numbers) that label each phase.

In turns out that the presence of a duality group conjugate to level 2 modular subgroups is a sufficient property of a field theory for the RG fixed points to coincide with the above conductivities (7.1) and (7.2). This approach to the RG flow and its connection with quantum Hall effects is discussed in for example [48] and is reviewed in [56]. The discussion in this section in similar to what is presented there but has a slightly different perspective on the duality groups that we use in the appended paper. In this way the group theoretic foundation of the dualities are put on a more natural level.

7.3.1 Parameter space and renormalization

The transport properties of the Hall system is governed by the conductivities (σ_D, σ_H), which will serve as the parameters in the field theory. Consider therefore the 2-parameter quantum field theories with classical action $S(\varphi;\sigma)$ where $\sigma = \sigma_H + i\sigma_D$ is a complexified coupling constant in the upper half plane $\mathfrak{H}$. We are interested in the theories where a modular duality group Γ_X acts on the moduli space. Here Γ_X is one of the level 2 modular subgroups we discussed in earlier chapters.

As we discussed in some detail in the chapter on dualities, the existence of a duality group will put strong constraints on the theory. In particular, if one thinks of the beta function as a 1-form on the moduli space this can be expanded in the basis of automorphic forms for the duality group. In the present case this corresponds to a weight 2 modular form for Γ_X. The physical beta function is the corresponding vector obtained by raising the covariant index.

Consider for example the case where the duality is Γ_T. The corresponding space of weight 2 forms is one dimensional and is spanned by $\mathcal{E}_T$. From earlier chapters we saw that this form could be written $\mathcal{E}_T = \frac{12}{\pi i}\partial\varphi_T(\sigma) = \frac{12}{\pi i}\partial \log\frac{\eta(2\sigma)}{\eta(\sigma)}$. More generally we can imagine a duality Γ_X so that the beta functions can be written in terms of φ_X, which we now naturally can interpret as a RG potential. On the upper half plane we use the hyperbolic metric. In particular we have $G^{\sigma\overline{\sigma}} = \mathfrak{I}(\sigma)^2$. The physical (vectorial) beta function can under the assumption of a sub modular duality be uniquely written [56]

$$\begin{aligned}\beta_X^{\overline{\sigma}} = G^{\sigma\overline{\sigma}}\beta_\sigma &= \mathfrak{I}(\sigma)^2\mathcal{N}\partial_\sigma\varphi_X, \\ \varphi_T &= \log\frac{\eta(2\sigma)}{\eta(\sigma)}, \\ \varphi_S &= \log\frac{\eta(\sigma/2)\eta(2\sigma)}{\eta^2(\sigma)}, \\ \varphi_R &= \log\frac{\eta(\sigma/2)}{\eta(\sigma)}.\end{aligned} \tag{7.3}$$

Here $\mathcal{N}$ is a normalization constant that will not be important in this work. We recall that the groups Γ_R and Γ_T were conjugate in $GL_2(\mathbb{Q})$ by the operation $G(\sigma) = 2\sigma$. This is apparent in the renormalization group potentials as well. In fact, this G-conjugation presents a natural perspective on the modular subgroups that will mix nicely with the perspective we have on Hall conductivities in different materials. In stead of considering all level 2 subgroups we want to think of just the two groups Γ_T and Γ_S *as well as* their G-conjugates. As we have seen Γ_R is G-conjugate to Γ_T, but we have not yet met the conjugate partner of $\Gamma_S = \left\langle S, T^2 \right\rangle$. By conjugating the generators S,T^2 by G we get

$$GSG^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1/2 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & -2 \\ 1/2 & 0 \end{bmatrix} \equiv Q$$

which maps $Q(\sigma) = -4/\sigma$. Similarly the G-conjugate of T^2 is T^4. The resulting group $\Gamma_Q = \left\langle Q, T^4 \right\rangle$ is a subgroup of $SL_2(\mathbb{Q})$, since F has rational matrix entries. Since Γ_Q is obtained by conjugation by $G(\sigma) = 2\sigma$ we can think of the group action of Γ_Q as a scaled version of the group action of Γ_S. Notice also that just as we could obtain modular forms for Γ_R by scaling Γ_T forms, we can create forms for Γ_Q by scaling arguments of Γ_S forms.

This presents a slightly different picture of the relevant duality groups. The three subgroups $\Gamma_{T,R,S}$ are all conjugate in $SL_2(\mathbb{Z})$, while the fourth group Γ_Q is not. If we want to put all these groups on the same level we should think of them as conjugate subgroups of $GL_2(\mathbb{Q})$. From this perspective we are dealing with four isomorphic subgroups. In this way any statement made regarding the universality class defined by one model can, at least in principle, be translated into a statement about the other models. For a nice overview of the relevant groups see fig.1 of the appended paper.

By plotting the RG flow based on the above beta function the fixed points structure can be seen. The stable (attractive ⊕) fixed points are, in the condensed matter lingo, the stable low-energy phases. We will derive the stable fixed points next and compare them to the conductivities derived above.

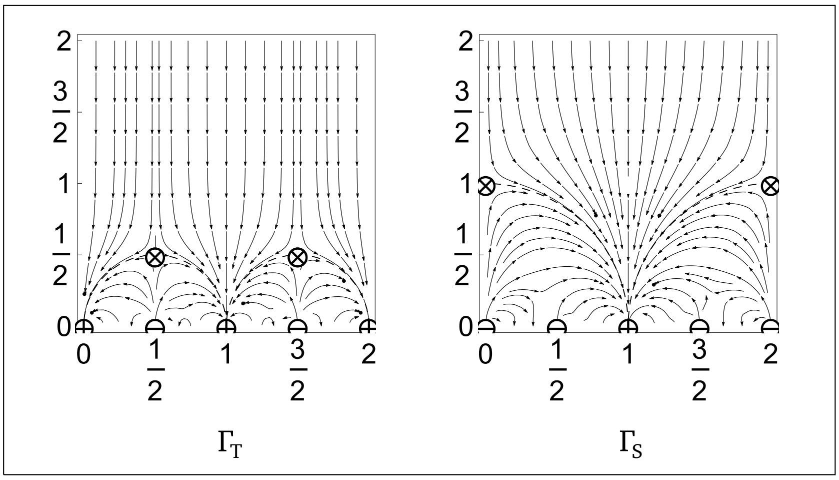


Figure 7.3: Renormalization group flow compatible with the Γ_T and Γ_R dualities. The G-conjugate groups have RG diagrams that are scaled overall by a factor 2.

7.3.2 Stable fixed points and 2-dimensional Dirac matter

The different RG flows obtained in this way will as we have seen have different stable fixed points [56]. To find these points we consider the transformation of fractions of certain parities under the different modular groups. Consider for example Γ_T acting on a even/odd fraction:

$$\frac{2\mathbb{Z}+1}{2\mathbb{Z}} \ni \frac{2p+1}{2q} \xrightarrow{\text{ST}^2\text{S}} \frac{2p+1}{2(q-2p-1)} \in \frac{2\mathbb{Z}+1}{2\mathbb{Z}}$$

$$\frac{2\mathbb{Z}+1}{2\mathbb{Z}} \ni \frac{2p+1}{2q} \xrightarrow{\text{T}} \frac{2p+2q+1}{2q} \in \frac{2\mathbb{Z}+1}{2\mathbb{Z}}$$

Hence Γ_T preserves the parity of such fractions. However, the even/odd fractions are not preserved:

$$\frac{2\mathbb{Z}}{2\mathbb{Z}+1} \ni \frac{2p}{2q+1} \xrightarrow{\text{ST}^2\text{S}} \frac{2p}{2q-4p+1} \in \frac{2\mathbb{Z}}{2\mathbb{Z}+1}$$

$$\frac{2\mathbb{Z}}{2\mathbb{Z}+1} \ni \frac{2p}{2q+1} \xrightarrow{\text{T}} \frac{2p+2q+1}{2q+1} \in \frac{2\mathbb{Z}+1}{2\mathbb{Z}+1}$$

In a similar manner, the odd/odd fractions are mapped into even/odd fractions by T. Note that in these phenomenological field theory approaches, the conductivity always from from somewhere in the upper half plane down to the rational numbers. Hence $i\infty$ should be considered a repulsive fixed point [56]. Since $\infty = 1/0$ is odd/even, the fractions with odd denominator correspond to stable phases in the EFTs. For Γ_T we can divide $\mathbb{Q}$ into three classes

$$\Gamma_T: \left\{\frac{2\mathbb{Z}}{2\mathbb{Z}+1}, \frac{2\mathbb{Z}+1}{2\mathbb{Z}+1}\right\}_{\oplus} \cup \left\{\frac{2\mathbb{Z}+1}{2\mathbb{Z}}\right\}_{\ominus}.$$

By the same arguments, we can find the classes corresponding to the Γ_{S} subgroup:

$$\Gamma_{\mathrm{S}}: \ \left\{\frac{2\mathbb{Z}+1}{2\mathbb{Z}+1}\right\}_{\oplus} \cup \left\{\frac{2\mathbb{Z}+1}{2\mathbb{Z}}, \frac{2\mathbb{Z}}{2\mathbb{Z}+1}\right\}_{\ominus}.$$

Since the Γ_{R} action can be obtained from the Γ_{T} case by a scaling by 2, the stable fixed points of Γ_{R} are $\frac{2\mathbb{Z}}{2\mathbb{Z}+1}$. Similar arguments hold for Γ_{Q}.

From this we can read off the general form of the fractions corresponding to stable phases. For Γ_{T}, the numerator can be both even and odd, while the denominator must be odd. Hence we have the fixed points $\oplus_{\mathrm{T}} = \frac{n}{2m+1}$. For the other subgroups we similarly have $\oplus_{\mathrm{R}} = \frac{2n}{2m+1}$, $\oplus_{\mathrm{S}} = \frac{2n+1}{2m+1}$ and $\oplus_{\mathrm{Q}} = \frac{4n+2}{2m+1}$. By comparing with equations (7.1) and (7.2) for conductivities above, we see that

$$\oplus_{\mathrm{T}} = \sigma_{\mathrm{H}}(d_s = 1), \tag{7.4}$$
$$\oplus_{\mathrm{R}} = \mathrm{G}(\oplus_{\mathrm{T}}) = \sigma_{\mathrm{H}}(d_s = 2), \tag{7.5}$$
$$\oplus_{\mathrm{S}} = \sigma_{\mathrm{H}}(n_f = 2, d_s = 1), \tag{7.6}$$
$$\oplus_{\mathrm{Q}} = \mathrm{G}(\oplus_{\mathrm{S}}) = \sigma_{\mathrm{H}}(n_f = 2, d_s = 2). \tag{7.7}$$

We see that the G-conjugation can be interpreted as a 2-fold spin degeneracy.

There is a nice geometric analogy in the integer case (m=0) that we should mention. Since we think of the conductivities as the first Chern numbers of bundles over the Brillouin zone, these four different classes can be seen as different classes of bundles. For example, if we let $\mathcal{L}_{\mathbb{K}}$ denote the line bundles with first Chern number belonging to $\mathbb{K} \subseteq \mathbb{Z}$, we can form the following correspondence:

$$\Gamma_{\mathrm{T}} \longleftrightarrow \mathcal{L}_{\mathbb{Z}}$$
$$\Gamma_{\mathrm{R}} \longleftrightarrow \mathcal{L}_{\mathbb{Z}} \oplus \mathcal{L}_{\mathbb{Z}} = \mathcal{L}_{2\mathbb{Z}}$$
$$\Gamma_{\mathrm{S}} \longleftrightarrow \mathcal{L}_{2\mathbb{Z}+1}$$
$$\Gamma_{\mathrm{Q}} \longleftrightarrow \mathcal{L}_{2\mathbb{Z}+1} \oplus \mathcal{L}_{2\mathbb{Z}+1} = \mathcal{L}_{4\mathbb{Z}+2}$$

We here mean that the group on the left hand side has integer fixed points corresponding to the first Chern number of the line bundle of the right hand side. In some sense, we see that the direct sum is the geometric analogy of the conjugation G. This makes sense since the direct sum of the bundles combines the two independent systems without mixing. One should also note that the S and Q class corresponds to classes of non-trivial bundles, since their Chern number is never zero.

In any case the model with n_f Dirac points can be identified with different effective QFTs with modular dualities. That is, they describe the same low energy fixed points. In this way the Hall effect in different materials can be placed into classes categorized by the modular groups. Before we discuss this we want to briefly discuss a field theory where the modular duality enters very naturally.

7.4 Double modularity in toroidal sigma models

We have not yet discuss any representative of the modular field theories. While these are hard to identify, there is a model with a seemingly too big parameter space suggested in [52]. The proposed field theory is a non-linear sigma model where the target space is a complex torus. The inclusion of complex toroidal geometry is very natural when we search for field theories with modular dualities since the moduli space of these tori are exactly the moduli space of the QFT. We here discuss the interplay between the moduli space of geometric structures for complex tori with spin structures and the dualities of this toroidal sigma model defined in flat Euclidian space. This is also discussed in Dijkgraafs les Houches lectures [23] as the simplest example of mirror symmetry in string theory.

7.4.1 Geometric structures and the parameter space

We now consider the setup of the toroidal model. We are considering a field theory where the field is a map

$$\varphi : \mathbb{R}^2 \to \mathrm{E}_{1,\tau} = \mathbb{C}/\Lambda_{1,\tau}$$

from flat two dimensional Euclidian space to a complex torus with complex structure parameter $\tau \in \mathfrak{H}$. Recall that such a two dimensional sigma model (the dimension referring to $\mathbb{R}^2$) admitted an additional term which was the pullback of a 2-form B_{ij} on the target manifold. We also let g be the metric tensor on the torus. The action for such theories, here written in local real coordinates, read

$$\mathrm{S} = \int d^2x\{\delta^{\mu\nu}g_{ij} + b\epsilon^{\mu\nu}\epsilon_{ij}\}\partial_\mu\varphi^i\partial_\nu\varphi^j, \tag{7.8}$$

which we recall from the section on worldsheet models. We have here written the anti-symmetric 2-form as $\mathrm{B}_{ij} = b\epsilon_{ij}$, since in our case the target is two dimensional, and the space of top-dimensional forms is one dimensional. From earlier discussions we also know that the parameter space of this theory is four dimensional. Hence it seems that it is too large to explain the 2-parameter scaling of the Hall systems. However, dualities come to the rescue.

First recall that by the uniformization theorem, the complex torus is the only genus one surface with zero curvature. The metric is inherited from the complex plane by the quotient, and takes the simple form $dz \otimes d\overline{z}$ in complex coordinates. We now introduce coordinates $a + \tau b$ on this torus, each ranging from 0 to 1. In these coordinates the metric takes the form

$$dz \otimes d\overline{z} = da \otimes da + |\tau|^2 da \otimes db + (\tau + \overline{\tau}) db \otimes db.$$

As a matrix we then write

$$g = \frac{\sqrt{g}}{\mathfrak{I}\tau} \begin{bmatrix} 1 & \mathfrak{R}\tau \\ \mathfrak{R}\tau & |\tau|^2 \end{bmatrix}.$$

We have here included the overall factor so that the torus has the proper area. Hence we seen that the toroidal model (7.8) has the following parameters in its moduli space

$$\{\mathfrak{R}\tau, \mathfrak{I}\tau, \sqrt{g}, b\}.$$

Of course, we can exchange these four real parameters for the two complex parameters

$$\{\tau, \sigma = b + i\sqrt{g}\}$$

as is done in for example [52]. To have a field theory where the moduli space in the modular curve $\mathfrak{H}/\Gamma_{\mathrm{X}}$ we should consider a non-linear sigma model where the target manifold is a complex torus with appropriate spin structure. The additional complex parameter σ (not related to the conductivity) can be interpreted as a complexified volume.

7.4.2 T-duality and modularity

We can now write down the theory dual to the toroidal sigma model (7.8) by using the Buscher rules (5.2) discussed in the chapter on dualities. Note that in (5.2) g is the metric on the base space, h is the target space metric and B is the 2-form on the target. In the toroidal model we used g to denote the *target* metric. Recall that $\mathfrak{R}\tau = \tau_1, \mathfrak{I}\tau = \tau_2$ and $\mathfrak{R}\sigma = b, \mathfrak{I}\sigma = \sqrt{g}$. The Buscher transformation for the toroidal case yields

$$g_{11} = \frac{\sqrt{g}}{\mathfrak{I}\tau} \quad , \quad g_{12} = \frac{\sqrt{g}}{\mathfrak{I}\tau}\mathfrak{R}\tau,$$

$$g_{21} = g_{12} \quad , \quad g_{22} = \frac{\sqrt{g}}{\mathfrak{I}\tau}|\tau|^2,$$

$$\mathrm{B}_{12} = -\mathrm{B}_{21} = b.$$

$$\Big\downarrow \mathscr{B}$$

$$\tilde{g}_{11} = \frac{\mathfrak{I}\tau}{\sqrt{g}} \quad , \quad \tilde{g}_{12} = \frac{b\mathfrak{I}\tau}{\sqrt{g}},$$

$$\tilde{g}_{22} = \sqrt{g}\,\Im\tau - b^2 \frac{\Im\tau}{\sqrt{g}},$$
$$\tilde{B}_{12} = \Re\tau.$$

These transformation become clearer if written in matrix form. The torus metric and 2-form are under the Buscher transformation mapped to

$$g = \frac{\sqrt{g}}{\Im\tau}\begin{bmatrix} 1 & \Re\tau \\ \Re\tau & |\tau|^2 \end{bmatrix} \to \tilde{g} = \frac{\Im\tau}{\sqrt{g}}\begin{bmatrix} 1 & \Re\sigma \\ \Re\sigma & |\sigma|^2 \end{bmatrix},$$

$$B = \begin{bmatrix} 0 & \Re\sigma \\ -\Re\sigma & 0 \end{bmatrix} \to \tilde{B} = \begin{bmatrix} 0 & \Re\tau \\ -\Re\tau & 0 \end{bmatrix}.$$

We see that these transformations simply interchange the two complex moduli of the theory $\tau \longleftrightarrow \sigma$. Hence the dual theory is still a sigma model on a torus, now with a different volume and shape. The modular group acts on both the parameters, and we can think of the moduli space of the toroidal model as two copies of the modular curve $\mathfrak{H}/\Gamma_X$.

7.5 Classification of materials by modular fixed points

We have seen that the modular subgroups Γ_X picks out particular classes of conductivities when acting on the rational numbers as the boundary of the extended upper half plane. This categorization of RG fixed points coincides with the rough classification of Hall effects in different materials with respect to their low-energy excitations. The modular symmetry is found to hold also away from the rational numbers, i.e. the full temperature driven RG flow of the quantum Hall system has a modular symmetry [54][49]. In our paper we explore this modular flow in several materials. Here we would like to use discuss some theoretical expectations using the modular fixed point formulas (7.4).

While it is true that once the degeneracies and number of Dirac points are identified we can use the above dictionary (7.4), it is not always easy to identify the number of low-energy Dirac fermions and the relevant degree of degeneracy. We will here consider the integer series (i.e. $m = 0$ in (7.4)) which for the different modular groups read

$$\oplus_T = \{..., -1, 0, +1, ...\},$$
$$\oplus_R = \{..., -2, 0, +2, ...\},$$
$$\oplus_S = \{..., -1, +1, ...\},$$
$$\oplus_Q = \{..., -2, +2, ...\}.$$

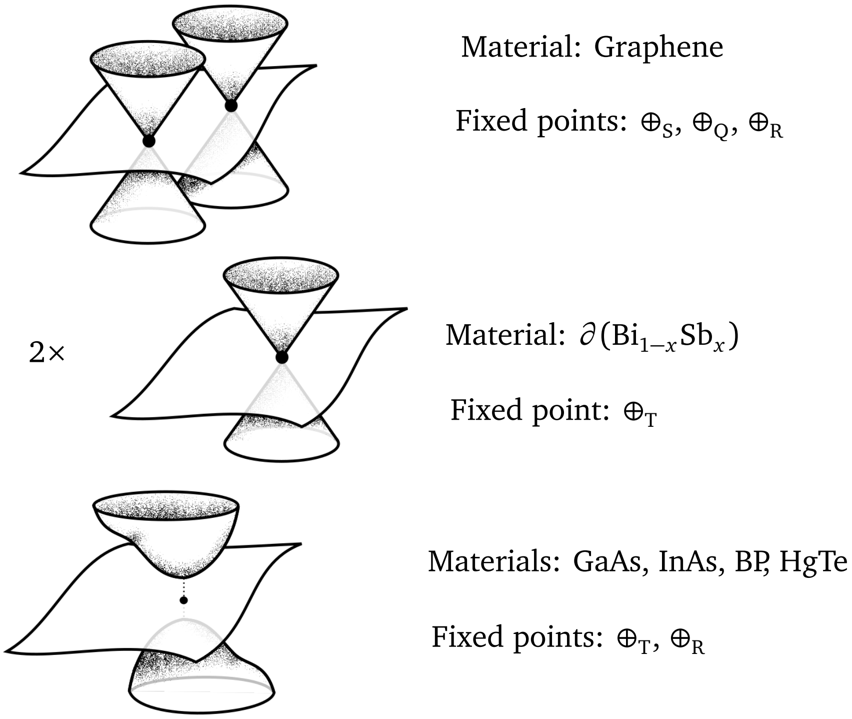
The vanishing of the $\oplus_{\mathrm{X}} = 0$ point is a consequence of the presence of Dirac points in $\mathcal{BZ}$, since the integer part of the conductivity (7.2) takes the form $\{\ldots, -\frac{d_s q}{2}, +\frac{d_s q}{2}, \ldots\}$ while (7.1) yields the series $\{\ldots, -d_s, 0, +d_s, \ldots\}$. Since the integer quantum Hall effect is a subclass of topological insulators, we think of the conductivities as the topological invariants labeling the phases. In this way, the $\oplus_{\mathrm{X}} = 0$ phase corresponds to a trivial insulator with no edge degrees of freedom. However, in the presence of a low-energy Dirac mode in the bulk there will always be some non-zero contribution to the Chern number making the phase non-trivial. Hence materials without Dirac points are not expected to fall into the Γ_{S}-class (or its G-conjugate). However, as we will see there may be phenomena that makes Dirac materials fall into the $\Gamma_{\mathrm{T,R}}$-class.

Often the materials in which the Hall effect takes place is a main material sandwiched between other materials or grown onto some substrate. The following discussion is somewhat naive in that it only considers the "main" materials and not the composed material as a whole. When the main material is sandwiched between two other materials one gets what is known as a quantum well that confines the electrons to the middle material. In this case it makes sense to consider the material properties of this main material. In the case where the material is grown onto some other material and a Hall effect takes place in the 2-dimensional interface it is not so obvious why only the "main" material should be considered. The largest class of materials in which the Hall effect has been observed is the Γ_{T}-class corresponding to spin-polarized materials without Dirac points in the Brillouin zone. Examples of materials that fall into this category that we will meet in the paper are gallium arsenide (GaAs), black phosphorous (BP) and mercury telluride (HgTe) which are semiconductors. All these gapped materials will most likely fall into the Γ_{T} or Γ_{R}-class depending on spin degeneracy and needs no special attention.

The first material that *does* deserve some special attention is graphene. Graphene has two Dirac points in its Brillouin zone, making it a Dirac material. Hence we expect to find $\oplus_{\mathrm{S}}$ or $\oplus_{\mathrm{Q}}$ fixed points according to (7.2). However, as we will see in the paper, we also observe $\oplus_{\mathrm{R}}$ as fixed points. This can be explained by taking into account the observed fine structure of the lowest Landau level in graphene which in high magnetic fields splits in two [90]. If we imagine this to happen adiabatically, there will be two states connected to the lowest Landau level each contributing $1/2$ to the conductivity. Hence we should in this case expect the fixed point $2(n + 1/2 + 1/2) = \oplus_{\mathrm{R}}$ in the polarized case or perhaps the G-conjugate of R in the unpolarized case. In this way the trivial insulator phase $\oplus = 0$ reappears.

Another example relevant to the upcoming paper is the Hall effect of the surface of a 3-dimensional topological insulator in the presence of ferromagnetism

[85]. The edges of this effective 2-dimensional system are ferromagnetic domain walls along which the charge carriers move. This surface has a single Dirac point in its Brillouin zone, with Hall conductivity $\sigma_{\mathrm{H}}(n_f = 1, d_s = 1)$. However, there is a partner particle at the opposite surface of the topological insulator, meaning the Hall conductivity is $\sigma_{\mathrm{H}}(n_f = 1, d_s = 1)^{\mathrm{top}} + \sigma_{\mathrm{H}}(n_f = 1, d_s = 1)^{\mathrm{bottom}} = n_t + n_b + 1 = \oplus_{\mathrm{T}}$. It is interesting that widely different materials, for example GaAs and the 3-dimensional topological insulator material $\mathrm{Bi}_{1-x}\mathrm{Sb}_x$ can both fall into the Γ_{T}-class. The only thing they seem to have in common is that their low-energy excitations conspire to give the same conductivity. We explore which materials belong to which modular class by studying the modular symmetry in the full parameter space $(\sigma_{\mathrm{H}}, \sigma_{\mathrm{D}})$ in the appended paper. A brief summary of the fixed point expectations is presented in the below figure which indicates both Brillouin zone and modular fixed point.


Material: Graphene
Fixed points: $\oplus_{\mathrm{S}}$, $\oplus_{\mathrm{Q}}$, $\oplus_{\mathrm{R}}$
2×
Material: $\partial(\mathrm{Bi}_{1-x}\mathrm{Sb}_x)$
Fixed point: $\oplus_{\mathrm{T}}$
Materials: GaAs, InAs, BP, HgTe
Fixed points: $\oplus_{\mathrm{T}}$, $\oplus_{\mathrm{R}}$

Part IV

Conclusion and outlook

Summary of results

The purpose of thesis has been twofold, with a series of smaller goals. The main motivation has been the exploration of modular symmetries both from a mathematical and a physical perspective. These two studies were joined together by the notions of universality and duality.

In the introduction we mentioned four main goals. The first of these was to explore the geometric nature of modular forms. We showed that several results took a simple form when expressed geometrically, which we believe may be more pedagogical. Particularly noteworthy perhaps was the construction of connections and covariant derivatives on the space of modular forms by means of logarithmic derivatives of eta functions. This lead to a simple geometric understanding of the Ramanujan identities.

The second goal was concerned with the ideas of universality and duality and was in large parts a review. We showed that by thinking of a quantum field theory as a geometric construction, a theory space started to take form. A rough partition of this space was in terms of the spacetime category of a theory and its spatial dimensions. We emphasized that a proper understanding of such a theory space is not possible until a definition of quantum field theory is agreed upon. In other words, just as different axioms of QFT emphasizes different aspects of the theories, they could also possibly lead to different perspectives on theory space. However, we saw that locally theory space could be understood as a series of patches corresponding to the parameter spaces of classical theories, with dualities relating different patches. If we imagine the set of dualities to form a group G we argued that a theory should really be considered the set of G-equivalences in theory space. We argued that it was this class of theories that should be used to probe a given universality class.

The last two goals had to do with placing the modular symmetries in a physical context. We reviewed different aspects of non-interacting topological matter like the Hall effect and other topological insulators, as well as Dirac materials like graphene. By a simple game of counting we saw that the conductivities of the Hall effect in different materials could be found. Besides a contribution from the Landau levels common to all Hall effects we saw that the number of low-energy Dirac fermions and spin degeneracy left unique fingerprints on the conductivities. If we viewed these conductivities as the low-energy fixed points in the RG flow of a 2-parameter QFT with appropriate duality we saw that different degrees of degeneracy and number of Dirac fermions corresponded exactly to different modular subgroups. If we picture the integer Hall effects as a subclass of all two-dimensional topological insulators, this result is a division into four classes labeled by different sets of Chern numbers. Two of these corresponded to materials with no Dirac cones and two to materials with Dirac cones. The fixed points of materials considered in our paper could be explained using these simple rules. In particular we saw that the surface of 3-dimensional insulators acted like effective Dirac materials in which the quantum Hall effect could take place.

There is however no theoretical reason why the modular duality of the 2-parameter field theories should coincide with experiments in the *entire* parameter space. An exploration of this global modular symmetry is the main purpose of our paper where we observe modular symmetries in all but one case. The cause of this anomalous case may be either experimental errors or possibly a symmetry breaking down to a smaller group than the ones considered. This modular symmetry is a new type of symmetry in Nature, in that it usually only appears in mathematical frameworks of a theory. For example, it appears as consistency conditions in 2-dimensional conformal field theories or string theories. In the present case the modular symmetry was directly measurable simply by plotting the data. There is no theoretical bias in any way. We explore the robustness and universality of this modular symmetric flow by comparing experimental data in a wide range of materials to the theoretical RG flow.

Outlook

There are several natural extensions of the work done in this thesis. Several of these have to do with the connection between the microscopic understanding of the Hall effect and the modular symmetries.

As was discussed in the paper it is possible that it will become necessary to study a larger class of modular subgroups to explain the data. While the four subgroups between the full modular group $\Gamma(1)$ and $\Gamma(2)$ are well understood, there are many more groups between, say, $\Gamma(2)$ and $\Gamma(4)$. In addition to understanding *why* a modular symmetry should appear in the first place, it would be nice to take a step back to see which modular groups should be considered. Given these groups one should try to find the basis for the weight 2 forms so that beta functions can be constructed.

A more theoretical question would be to explore the automorphic forms for groups other than the modular subgroups. An overview of different groups with low-dimensional space of automorphic forms could be valuable if these symmetries were to appear in Nature. This is likely related to the mathematical theory of Shimura varieties (higher dimensional analogues of the moduli space of complex tori) and maybe even the Langlands program.

Another interesting question is how the different notions of stability for the Hall phases are related. From the point of view of vector bundle topology the Hall phases are associated with topological classes of a bundle over the Brillouin torus, i.e. a Chern class $c_1(\mathcal{E})$. We also saw that for a higher rank bundle the conductivity took the form $c_1(\mathcal{E})/\mathrm{rk}(\mathcal{E})$. From the point of view of modular effective field theory the same numbers that label the phases are placed along the edge of the extended upper half plane $\partial\overline{\mathfrak{H}} = \mathbb{Q}$. However, from this modular approach there seem to be nothing topological about the phases. In this case stability is represented by the fact that the conductivities are attractive fixed points of a renormalization group flow, and hence not sensitive to small perturbations. It would be interesting to see if there are any mathematical connections between the two subjects, i.e. bundle topology over tori and edges of the upper half plane. One piece of circumstantial evidence for this connection is the fact that if we view the Brillouin torus as a complex surface, its moduli space is exactly the upper half plane modulo modular transformations. Another type of stability is also involved here since the slope $\mu = c_1(\mathcal{E})/\mathrm{rk}(\mathcal{E})$ is used to answer mathematical stability question of bundles. These different notions of stability and their interplay would be interesting to study.

There is also more work that can be done experimentally. A lot is happening in material science and many new 2-dimensional materials are appearing. To enlarge the class of materials in which the modular symmetry is observed would strengthen the hypothesis of a modular symmetry in the effective theory. Experimental data with fractional phases would also be valuable as it is here the modular group really shows its strengths.

Paper

(In preparation)

DRAFT2 May 15, 2017

On the universality of modular symmetries in two-dimensional magneto-transport

K.S. Olsen, H.S. Limseth and C.A. Lütken
Department of Physics, University of Oslo

We analyze experimental quantum Hall data from a wide range of different materials, including semiconducting heterojunctions, thin films, surface layers, graphene, mercury telluride, bismuth antimonide, and black phosphorus. The fact that these materials have little in common, except that they all are effectively two-dimensional, shows how robust and universal the quantum Hall phenomenon is. The scaling and fixed point data we analyze appear to show that magneto-transport in two dimensions is governed by a small number of universality classes that are classified by infinite discrete symmetries not previously seen in Nature. These are so rigid that they fix the global geometry of the scaling flow, and therefore predict the exact location of quantum critical points, as well as the shape of flowlines anywhere in the phase-diagram. The Hall plateaux are (infrared) stable fixed points of the scaling-flow, and quantum critical points (where the wavefunction is delocalized) are unstable fixed points of scaling, which in some cases has been observed over several decades in temperature. We show that most available experimental quantum Hall scaling data is in good agreement with these predictions.

I. INTRODUCTION

The continuous and discrete symmetries observed in Nature may be exact or approximate. The continuous case includes exact symmetries like Lorentz and gauge invariance, which severly constrains possible dynamical models, while discrete symmetries usuallly are finite and approximate. For example, parity P, charge conjugation C, time reversal T and CP are all broken, leaving CPT as the only exact discrete space-time symmetry. We shall here investigate a class of experimental data that appear to respect a new type of symmetry that is called *modular*. Although these are finitely generated approximate (emergent) discrete symmetries, because they are non-abelian and infinite they provide unusually strong constraints on low-energy model building.

Infinite discrete groups, including modular symmetries, play an important role in modern mathematics, but because they are extremely rigid it is not clear if they can exist in the real world of experimental physics. Indeed, it is only in bespoke physical systems ("designer universes") engineered to be effectively (i.e., for all practical purposes) 2-dimensional that modular symmetries have been found.

The quantum Hall effect (QHE) appears in materials where charge carriers are forced to move in a single atomic plane, for example on the surface of a crystal or in a sheet of graphene. Experiments measuring the electromagnetic properties (magneto-transport) of Hall-systems produce what at first sight appears to be an impenetrable morass of data. But first appearances can be misleading, and if the quantum Hall data is viewed from a particularly advantageous vantage point a hidden pattern of great beauty and utility is revealed.[1–4] This rigid emergent order is encoded in a fractal phase-diagram tightly harnessed by a modular symmetry that allows it to teeter on the brink of chaos, without actually taking the leap.

Our purpose here is to explore the robustness and univerality of these new symmetries, by comparing and contrasting data from the most disparate materials available. We do this in the simplest possible way, by superimposing scaling data directly onto mathematical diagrams with modular symmetry. This "phenomenological" approach is unbiased, since no theoretical assumptions are invoked. We will not here discuss theoretical ideas that are needed in order to connect the well-known microphysics ("electrons in a dirty lattice") to the emergent macrophysics observed in transport experiments.

Since modular mathematics is unfamiliar to most physicists, in the next section we provide a brief introduction to modular symmetry in physics. In order to motivate this, there are two key observations that transmutes modular curiosities into a powerful tool in physics:

(i) scaling functions (β-functions in renormalization theory) must respect any symmetry of the parameter space on which they act (space of coupling constants or transport coefficients)

(ii) two-dimensional parameter spaces may be endowed with a natural complex structure.

If the emergent (parameter space) symmetry is modular, these circumstances conspire to give a very strong constraint on low-energy physics, and any model of this physics, as we now explain.

In our case the space of "coupling constants" is parametrized by the Hall- and magneto-resistivities ρ_H and ρ_D, or equivalently, the conductivities σ_H and σ_D. It is convenient to combine these into a complex quantity $\sigma = \sigma_H + i\sigma_D$ that takes values in the upper half of the complex plane: $\sigma \in \mathbb{C}^+$ (since $\sigma_D > 0$). This is useful because it reduces matrix operations to ordinary (complex) algebra, but it is much more than that. The flow generated by the scaling functions $\beta_H = d\sigma_H/dt$ and

$\beta_D = d\sigma_D/dt$ (t is the dominant scale parameter, usually determined by the temperature) is severly constrained by the following conjunction of favourable physical and mathematical circumstances.

(a) Since there are no sources ($\ominus$) or sinks ($\oplus$) for the scaling flow on $\mathbb{C}^+(\sigma)$, it is divergence free.

(b) Since a curl would render the physical interpretation of the β-functions meaningless, the flow must also be curl free.

(c) Since the β-functions vanish at quantum critical points $\otimes$, the flow has *simple* zeros on $\mathbb{C}^+(\sigma)$ iff $\sigma \in \otimes$.

(d) It is an empirical fact that the flow obeys a modular symmetry Γ.

In two dimensions (a) and (b) are equivalent to the Cauchy-Riemann equations, and it follows that $\beta = \beta_H + i\beta_D$ is a holomorphic function of σ. Together with (c) and (d) this means that β must be a complex analytic modular vector field with simple zeros. In mathematics this is called *a modular form of weight two.*

It is the paucity of such functions (forms) on large modular groups that gives modular symmetry extremely sharp teeth. The first useful result is that there are *no* such forms at all if Γ is the full modular group $\Gamma(1) = \mathrm{SL}(2,\mathbb{Z})$, and therefore no candidate β-functions with this symmetry. This provides a theoretical reason, independent of the experimental observation that this symmetry is too strong, for considering smaller groups. So we turn our attention to maximal subgroups of $\Gamma(1)$, where further surprises await us as we are forced to draw two improbable conclusions:

(A) for any of the largest viable symmetries (maximal subgroups of the modular group) *the β-function is unique*, up to an overall normalization.

(B) if the modular symmetry is reduced to the biggest subgroup (called $\Gamma(2)$) shared by the maximal subgroups of $\Gamma(1)$, then *there is a unique family of β-functions, parametrized by a single real number*, up to an overall normalization.

We shall see that this provides a host of rigid predictions that are eminently falsifiable. The most surprising consequence of a modular symmetry is perhaps that the plateaux *must* be rational. This follows from the remarkable fact that in order for a modular symmetry to act "properly" on the real line (in a strict mathematical sense[5]), which by definition is excluded from the upper half plane, $\mathbb{C}^+(\sigma)$ is compactified by adding *only* rational numbers: $\overline{\mathbb{C}}^+(\sigma) = \mathbb{C}^+(\sigma) \cup \mathbb{Q}$. It is also appealing that the integer (IQHE) and fractional (FQHE) quantum Hall effects are automatically and inextricably unified by any modular symmetry.

The mathematical primer in Sect. II is followed by an equally brief introduction in Sect. III to the novel materials that have yielded most of the new data discussed in the following sections. They give a fairly comprehensive overview of the current experimental status of the modular hypothesis, including all scaling experiments we have found to be of sufficient quality to enable us to extract a partial flow diagram. Sects. IV-VIII provide what is essentially a catalogue of fixed point data and scaling diagrams, organized by the modular symmetry they exhibit. Within each of these universality classes the data are grouped according to the type of material used in the experiment.

Sect. IX summarizes the successes of the modular paradigm so far, as well as some of the outstanding problems and challenges to be addressed in future work.

II. MODULAR SYMMETRY

The nested hierarchical structure that is emerging in phase portraits of the QHE is the signature of an approximate global discrete symmetry, which, given some familiarity with modular groups, is surprisingly easy to identify by finding some of the fixed points.

The sources and sinks of the scaling (RG-) flow, i.e., the "trivial" ultraviolet (UV) fixed points ($\ominus$) and infrared (IR) fixed points ($\oplus$ = plateaux), all lie on the boundary of the parameter space. The quantum critical points $\otimes$ all lie in the interior of parameter space. This fixed point structure, which can be extracted directly from the geometry of the data without any theoretical bias, is the DNA of the symmetry, from which all else will follow.

The full modular group $\mathrm{SL}(2,\mathbb{Z}) = \langle T, S\rangle$ can be represented by fractional linear (Möbius) transformations, generated by *translations* $T(z) = z + 1$ and *duality transformations* $S(z) = -1/z$, acting on the upper half of the complex plane $\mathbb{C}^+(z)$. It is the fact that T and S do not commute that makes this group infinite, and interesting. Any "word" in T and S is a fractional linear (Möbius) transformation $\gamma(z) = (az + b)/(cz + d)$, with integer coefficients and unit determinant ($ad - bc = 1$). Words can only be simplified using the "grammatical" rules $S^2 = 1 = (ST)^3$ that define the abstract group.

As far as the full modular group is concerned, all fractions (plateaux values) are equivalent, so if this were a physically viable symmetry we should observe all possible fractional plateaux. However, we never observe the full set of fractions in any given quantum Hall experiment, but only plateaux (fractions) that satisfy certain constraints on the parities of the numerator, or denominator, or both. These *parity rules*, which depend on the 2-dimensional material under consideration, are the key to identifying any would-be modular symmetry. They link microphysics to macrophysics, because the observed spectrum of integer fixed points follows directly from the spectrum of charge carriers supported by the system in the non-interacting limit ("Landau level spectroscopy").

The resistivity $\rho = S(\sigma) = -1/\sigma$ is conveniently given by the modular duality transformation S, since this is equivalent to taking the matrix inverse of the conductivity tensor. Note that it is conventional to choose $\sigma_H = \sigma_{12}$ and $\rho_H = \rho_{21}$ in order to eliminate a minus sign.

A. Soupçon of group theory

So the full modular symmetry is too strong for the QHE, but the largest subgroups of $\mathrm{SL}(2,\mathbb{Z})$ are not. A map showing the tip of the modular iceberg, including all the groups we need, is presented in Fig. 1.

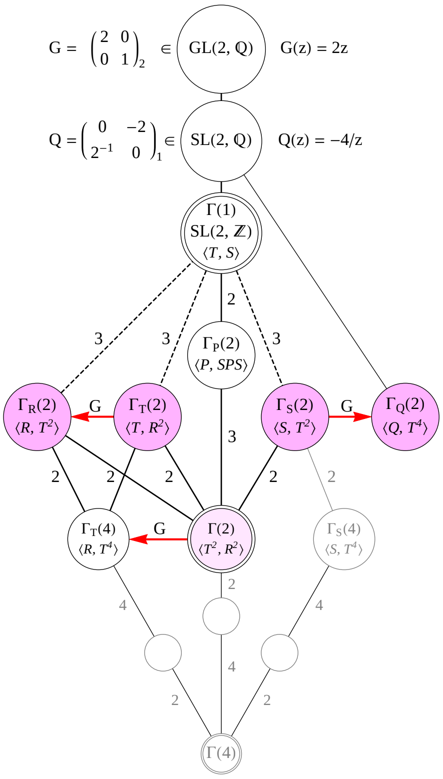


Figure 1. Some of the groups between $\mathrm{GL}(2,\mathbb{Q})$ and $\Gamma(4)$. $P = ST$ ($P^3 = 1$). There are another twenty groups between $\Gamma(1) = \mathrm{SL}(2,\mathbb{Z})$ and $\Gamma(4)$ that are not shown here.[7,8] A thick solid line means that the subgroup is normal, and the index of the subgroup labels the line. The red arrow is a *modular correspondence* obtained by conjugating with $G \in \mathrm{GL}(2,\mathbb{Q})$, where $G(z) = 2z$. The relation $\Gamma_{\mathrm{T}}(4) = G\,\Gamma(2)\,G^{-1}$ is important in the theory of theta-functions (modular forms of weight $w = 1/2$).[9] Conjugating the polarized group $\Gamma_{\mathrm{T}}(2)$ gives the familiar unpolarized group $\Gamma_{\mathrm{R}}(2) = G\,\Gamma_{\mathrm{T}}(2)\,G^{-1}$, but $\Gamma_{\mathrm{Q}}(2) = G\,\Gamma_{\mathrm{S}}(2)\,G^{-1}$ is new. The G-conjugate $Q = G\,S\,G^{-1}(z) = -4/z$ of the duality generator S is called a *Fricke involution.* The pink level two groups are relevant for the QHE, and the four groups $\Gamma_{\mathrm{X}}(2)$ (X = Q, R, S, T) are the symmetries usually observed in experiments (compare next figure). Since only level two appears to be physically relevant, we usually simplify notation by dropping the level ($\Gamma_{\mathrm{T}} = \Gamma_{\mathrm{T}}(2)$, etc.).

Subgroups of the modular group are obtained by relaxing the translation symmetry ($T \to T^n$), or the duality symmetry ($S \to R^n$, where $R(z) = TST(z) = z/(1+z)$), or both. Three of these so-called "congruence subgroups at level two" preserve parities, which means that each of them groups the fractions into two equivalence classes. Because p and q in $\sigma_{\oplus} = p/q$ are relatively prime, there are only three types of fractions with well defined parities. With "o" representing odd integers and "e" representing even integers, we have $p/q \in o/o$, o/e or e/o, and it is easy to verify that the equivalence classes are[6]

$$\Gamma_{\mathrm{T}} = \langle T, R^2\rangle : \quad \left\{\frac{e}{o}, \frac{o}{o}\right\}_{\oplus} \cup \left\{\frac{o}{e}\right\}_{\ominus}$$

$$\Gamma_{\mathrm{R}} = \langle R, T^2\rangle : \quad \left\{\frac{e}{o}\right\}_{\oplus} \cup \left\{\frac{o}{o}, \frac{o}{e}\right\}_{\ominus}$$

$$\Gamma_{\mathrm{S}} = \langle S, T^2\rangle : \quad \left\{\frac{o}{o}\right\}_{\oplus} \cup \left\{\frac{o}{e}, \frac{e}{o}\right\}_{\ominus} .$$

A class is indexed by $\oplus$ if the fractions are sinks (attractive fixed points) for the scaling flow in the σ-plane, and by $\ominus$ if they are sources (repulsive fixed points). This assignment follows from the requirement that the direction of the flow is downward at the top of the conductivity plane, which is a result that can be obtained in a perturbative analysis of localisation in the weak coupling limit $\sigma \to i\infty$. The fixed point at vanishing coupling must therefore be repulsive, $i\infty = \ominus$. Since $\infty = 1/0 \in o/e$, and all fixed points in a given class are mapped into each other by the symmetry, all fractions in the class containing o/e must be repulsive. Notice that the denominators of attractors are always odd. Following Laughlin, this is a consequence of the Fermi statistics obeyed by electrons.

B. Modular phase-diagrams

Because the duality transformation S swaps e/o and o/e, leaving o/o unchanged, the direction of the flow in the $\rho = S(\sigma)$-plane is reversed if the symmetry acting on σ is Γ_{T} or Γ_{R}, but not if the symmetry is Γ_{S} (since it contains S). This dichotomy is a persistent theme.

The fixed point at the origin of the σ-plane (at $i\infty$ in the ρ-plane) has a special significance. If it is attractive this means that the system has an insulating phase, which we call the quantum Hall insulator (QHI) and assign the special symbol $\circledast$. Since $0 = 0/1 \in e/o$, we conclude that a model with Γ_{T}- or Γ_{R}-symmetry in the σ-plane does have this insulator phase, but that a Γ_{S}-symmetric model does not.

Observe also that Γ_{R} and Γ_{T} are conjugate inside the parent group $GL(2,\mathbb{Q})$ under the rescaling $G(z) = 2z$ by a factor of two (compare Fig. 1). This means that flow diagrams with these two symmetries are identical, up to a doubling of all coordinates. A similar rescaling of Γ_{S} gives a conjugate group Γ_{Q} that is not strictly speaking modular (compare Fig. 1), but its flow diagram is just a doubling of the Γ_{S}-symmetric flow.

In summary, there are just two types of flow diagrams with maximal admissible (both Γ_{P} and $\Gamma(1)$ are too large) modular symmetry: $\Gamma_{\mathrm{T}}(\sigma)$ (and its G-conjugate $\Gamma_{\mathrm{R}}(\sigma)$), and $\Gamma_{\mathrm{S}}(\sigma)$ (and its G-conjugate $\Gamma_{\mathrm{Q}}(\sigma)$).

For convenience a chart of the $Q-$, $R-$, $S-$ and T-flows, in both σ and ρ, is provided in Fig. 2. In these cases the

IQHE	$\sigma = \sigma_H + i\sigma_D \in \overline{\mathbb{C}}^+(\sigma)$			$\rho = \rho_H + i\rho_D \in \overline{\mathbb{C}}^+(\rho)$		
Γ_{X}	$\oplus$	$\longleftarrow \otimes \longrightarrow$	$\oplus'$	$\oplus$	$\longleftarrow \otimes \longrightarrow$	$\oplus'$
Γ_{T}	n	$\frac{2n+1+i}{2}$	$n+1$	$\frac{1}{n+1}$	$\frac{2n+1+i}{2n^2+2n+1}$	$\frac{1}{n}$
Γ_{R}	$2n$	$2n+1+i$	$2n+2$	$\frac{1}{2n+2}$	$\frac{2n+1+i}{4n^2+4n+2}$	$\frac{1}{2n}$
Γ_{S}	$2n-1$	$2n+i$	$2n+1$	$\frac{1}{2n+1}$	$\frac{2n+i}{4n^2+1}$	$\frac{1}{2n-1}$
Γ_{Q}	$4n-2$	$4n+2i$	$4n+2$	$\frac{1}{4n+2}$	$\frac{2n+i}{8n^2+2}$	$\frac{1}{4n-2}$

Table I. Left half: Integer plateaux values $\oplus$ of the Hall conductivity σ_H constrained by a symmetry $\Gamma(2) \subset \Gamma_{\mathrm{X}} \subset \mathrm{SL}(2,\mathbb{Q})$, with X = Q, R, S or T, together with the location of semi-stable fixed points for transitions between these plateaux, i.e. the position of "integer" quantum critical points $\otimes$ in the complexified conductivity-plane. Right half: Corresponding values of the resistivity (see Sect. II for details).

shape of the flow lines (but not the flow rate) is completely fixed by the large symmetry. They are most easily derived as a gradient flow of RG-potentials with the requisite symmetry. We defer details to the discussion below of symmetry breaking.

For future reference we have also listed the integer fixed points for these cases in Tab. I. The complete spectrum of attractors (plateaux) for these symmetries may be found in Fig. 3.

Γ_{T} and Γ_{R} are the relevant groups for the ordinary spin-polarized and unpolarized QHE, respectively, where quasi-particles have the usual parabolic ("nonrelativistic") dispersion, i.e., the QHE that appears in materials without Dirac-modes. We will therefore call these the *nonrelativistic polarized and unpolarized* groups, respectively.

Graphene is different. Due to the peculiar topology of its Fermi surface, there are gapless (massless) excitations at half filling with linear dispersion, i.e., their energy is linear in momentum, and there is a doubling of degrees of freedom due to an additional "pseuodspin" or "valley" degeneracy. These modes therefore behave like relativistic (Dirac) fermions, with the Fermi velocity replacing the speed of light. The linear dispersion and unusual band structure leads to a different non-interacting spectrum, but that is all we need to identify the potential modular symmetry, and the phenomenological analysis of graphene is analogous to the parabolic case[10].

The relevant groups in this "relativistic" case with Dirac modes are Γ_{S} and Γ_{Q}, for the spin-polarized and unpolarized QHE respectively. We will therefore call these the *relativistic polarized and unpolarized* groups, respectively.

A phase is by definition the set of all points in $\overline{\mathbb{C}}^+$ that flow to a given plateau $\oplus$ (IR fixed point), and is uniquely

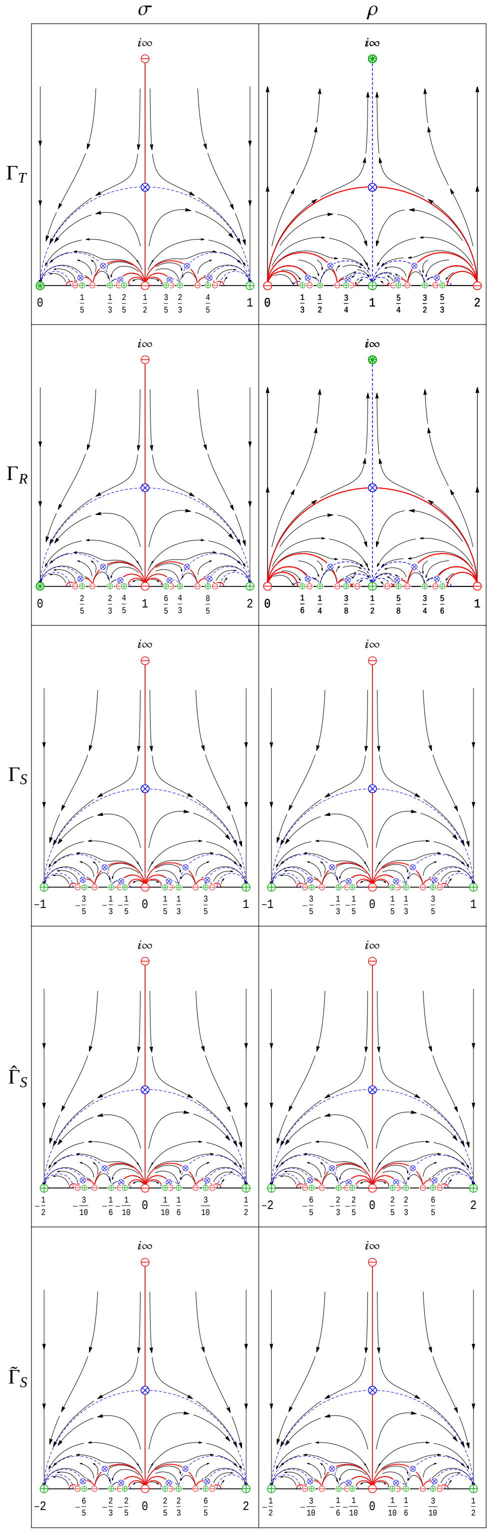


Figure 2. (Color online)(placeholder) Conductivity (right) and resistivity (left) phase-diagrams with symmetry Γ_{X} (X = Q, R, S, T). Only $\Gamma_{\mathrm{T}}(\sigma)$ and $\Gamma_{\mathrm{S}}(\sigma)$ are truly different, since $\Gamma_{\mathrm{R}}(\sigma)$ is a simply a doubling of $\Gamma_{\mathrm{T}}(\sigma)$, and likewise for $\Gamma_{\mathrm{S}}(\sigma)$ and $\Gamma_{\mathrm{Q}}(\sigma)$, and $\rho = S(\sigma)$.

(a)

$$\begin{tikzcd}
\boxed{\Gamma_{\mathrm{Q}}}^{\uparrow\downarrow}_{\Uparrow\Downarrow} \arrow[r, "\not\updownarrow"] \arrow[d, "\not\Updownarrow"'] \arrow[dr, phantom, "\sigma"] & \boxed{\Gamma_{\mathrm{S}}}^{\uparrow}_{\Uparrow\Downarrow} \arrow[d, "\not\Updownarrow"] \\
\boxed{\Gamma_{\mathrm{R}}}^{\uparrow\downarrow} \arrow[r, "\not\updownarrow"'] & \boxed{\Gamma_{\mathrm{T}}}^{\uparrow}
\end{tikzcd}$$

(b)

$$\begin{tikzcd}
\boxed{\oplus_{\mathrm{Q}} = \frac{4n+2}{2m+1}} & \boxed{\oplus_{\mathrm{S}} = \frac{2n+1}{2m+1}} \arrow[l, "\alpha"'] \\
\boxed{\oplus_{\mathrm{R}} = \frac{2n}{2m+1}} \arrow[u, "\beta"] \arrow[ur, phantom, "\begin{array}{c}\alpha:\sigma\to2\sigma\\ \beta:n\to2n+1\end{array}"] & \boxed{\oplus_{\mathrm{T}} = \frac{n}{2m+1}} \arrow[l, "\alpha"] \arrow[u, "\beta"']
\end{tikzcd}$$

Figure 3. (a) A possible pattern of symmetry breaking. Superscripts refer to real spin: ↑↓ means that spin-splitting is insignificant, so that spin-up and -down states belong to the same band, while ↑ means that spin-splitting has left only spin up states in the lowest band. With spin-splitting there are half as many charge carriers per band, and consequently the conductivity is half of what it is when the ↑- and ↓-bands are degenerate (α). Subscripts refer to pseudospin: ⇑⇓ means that there are two Dirac cones with linear ("relativistic") dispersion, and the absence of a subscript means that there are none, so the dispersion is parabolic ("non-relativistic"). Adding Dirac cones adds zero-modes (β). (b) Each symmetry group Γ_{X} (X = Q, R, S, T) leaves a distinct fingerprint on the spectrum $\oplus_{\mathrm{X}} = \oplus_{\mathrm{X}}(n \in \mathbb{Z})$ of attractive fixed points (rational plateaux values of the Hall-conductivity σ_H $[e^2/h]$).

labelled by this limit point on the real axis. A phase transition between two plateaux $\oplus$ and $\oplus'$ is permitted by the symmetry iff it has a fixed point $\otimes$ located on the semi-cricle in $\overline{\mathbb{C}}^+$ connecting $\oplus$ and $\oplus'$, which we write as $\oplus \leftarrow \otimes \rightarrow \oplus'$ or $\oplus \overset{\otimes}{\longleftrightarrow} \oplus'$. If one of the attractors is $i\infty$ the semi-circle has infinite radius, i.e., it is a vertical line. We also adopt the convention that $\oplus \overset{\otimes}{\longleftrightarrow} \oplus'$ refers to a transition in the conductivity plane, whence an integer plateau-value $\oplus = \sigma_{\oplus} = \sigma_H = n$ $[e^2/h] \in \mathbb{Z}$ refers to the IQHE ($\rho_H = 1/n$ $[h/e^2]$).

C. Symmetry-breaking

We discuss two types of symmetry-breaking that are of immediate relevance for the QHE.

Firstly, the spin-valley symmetry giving rise to Γ_{Q} can be broken by in-plane and other interactions, or possibly by external electric and magnetic fields. Independently of the microscopic mechanism, Fig. 3 (a) shows a possible pattern of symmetry breaking that transmutes the flow diagrams (and associated fixed point structures).

The broken arrow $\not\updownarrow$ means that spin degeneracy is broken, which changes an unpolarised spectrum (left column) to a polarised spectrum (right column), while $\not\Updownarrow$ means that the valley degeneracy between the two sublattices is broken and both Dirac cones have been destroyed. This changes a relativistic spectrum (top row) to a non-relativistic spectrum (bottom row). Each symmetry leaves a unique fingerprint on the plateau-spectrum, compare Fig. 3 (b).

Another, more severe type of symmetry breaking appears when the spins are neither fully polarized, nor fully degenerate. The maximal groups are no longer relevant, but it is conceivable that some smaller symmetry survives. The simplest situation would be if we have "minimal breaking", meaning that the largest common subgroup survives. From our map in Fig. 1 we see that this is $\Gamma(2)$, and our task is to find a $\Gamma(2)$-symmetric family of physically viable β-functions that interpolate between Γ_{R}, Γ_{T} and Γ_{S}. (Γ_{Q} is not in this family because it is not in the modular group $\Gamma(1)$.)

$\Gamma(2)$ admits a 2-dimensional space of weight two forms, which is spanned by two of Jacobi's theta-functions, for example θ_2^4 and θ_3^4. The third Jacobi function also gives a weight two form θ_4^4, but it is already included because it is a linear combination of the other two, $\theta_4^4 = \theta_3^4 - \theta_2^4$. Any $\Gamma(2)$-symmetric β-function must therefore be a linear combination of these,[11,12]

$$\beta_a \propto \theta_2^4 - a\theta_3^4 \propto \partial\varphi_a,$$

i.e., a gradient flow derived from the $\Gamma(2)$-invariant RG-potential

$$\varphi_a = \ln \lambda(\lambda - 1)^{a-1}, \quad \lambda = (\theta_2/\theta_3)^4 \qquad (a \in \mathbb{R}).$$

The phenomenological parameter a, which has an unknown and presumably complicated dependence on non-universal microscopic details (e.g. Zeeman-splitting), must be real for the flow to agree with perturbative localization theory at weak coupling.

For certain values of the parameter a the symmetry is enhanced to one of the maximal modular subgroups discussed above (to Γ_{R} when $a = -1$, Γ_{T} when $a = 1/2$, and Γ_{S} when $a = 2$), but never to the full modular group. These are the cases most often encountered in experiments. For example, very strong magnetic fields polarize the samples and yield a large Zeeman splitting that effective eliminates half the bands, leading to Γ_{T}-symmetry if the Fermi-surface does not give additional zero-modes (graphene).

Fig. 4 shows the complete family of $\Gamma(2)$-symmetric flow diagrams (for $a < -1$ the flows are similar to the flow at $a = -1$, and for $a > 2$ the flows are similar to the one at $a = 2$). Slicing this "family-plot" at any value of the symmetry-breaking parameter $a \neq 0, 1, \infty$ gives a "warped" but physicially sensible diagram, i.e. a scaling flow that is finite except for simple zeros. These are the

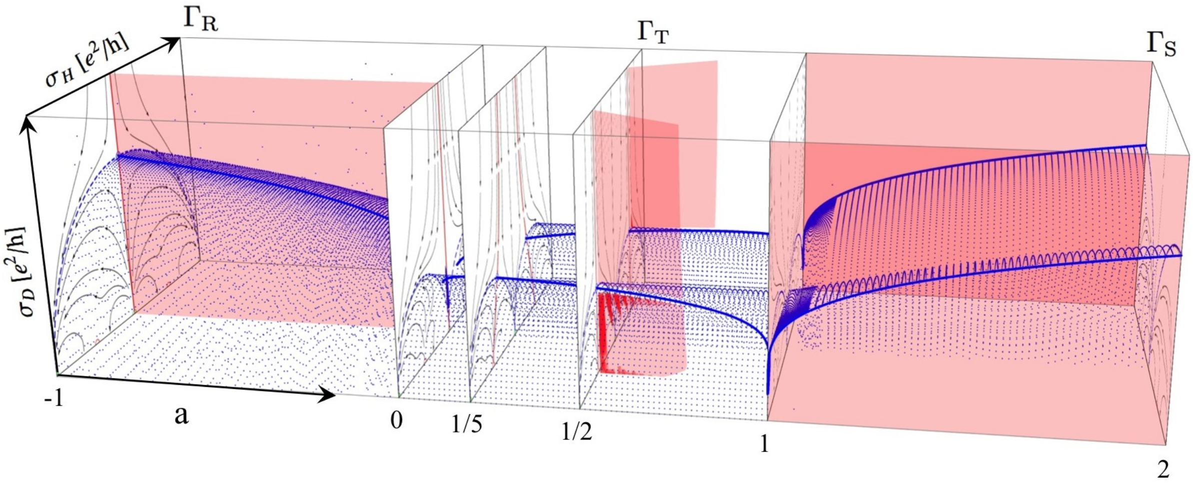


Figure 4. (Color online)(placeholder) One-parameter family of $\Gamma(2)$-symmetric RG-flows. For certain values of the parameter a the symmetry is enhanced to one of the maximal modular subgroups: $a = -1$ has Γ_R-symmetry, $a = 1/2$ has Γ_T-symmetry and $a = 2$ has Γ_S-symmetry, shown here by slicing up the family-plot, together with some of the scaling data discussed in more detail below.

quantum critical points (they are fixed under rescaling and therefore RG fixed points) and they are located at the $\Gamma(2)$-images of $\sigma_\otimes = iK'(a^{-1})/K(a^{-1})$, where K and K' are elliptic integrals of the first kind.[11] This family is sufficiently large to accomodate virtually all quantum Hall data we have examined.

III. NEW MATERIALS

We have argued that the remarkable convergence of modular mathematics and quantum Hall physics suggests that it would be unnatural to restrict attention to only one of the descendants of the modular group. We have also seen that there are very few viable candidates to choose from, and that these fit snugly into a simple and unique one-parameter family of $\Gamma(2)$-symmetric β-functions (up to normalization). In other words, while these infinite non-abelian symmetries are extremely constraining, they do leave enough flexibility that we can accomodate almost all experiments to date (but only barely so).

The discovery in recent years of new types of materials that support Dirac modes and "robust" topological edge states presents new opportunities for testing the modular paradigm sketched above. We will review a number of recent experiments that have explored large tracts of the modular landscape that were previously inaccessible.

These experiments have provided substantial evidence for those level two symmetries that until now have been beyond our reach, and aguably verified that the full complement of level two symmetries are present in Nature. In preparation for that discussion we give a brief summary of some of the most salient features of these materials.

A. Dirac matter

Dirac matter is a name used to commonly describe materials in which the low-energy excitations are Dirac fermions. In terms of Bloch theory these states appear as a consequence of a finite number of crossing points in the materials Brillouin zone where the Hamiltonian becomes gapless. Close to these points the energy dispersion is linear, similar to the relativistic dispersion in particle physics. This dispersion is often referred to as a Dirac cone. The effective Hamiltonian in the low-energy regime is a Dirac Hamiltonian, hence the name of the material. When a Hall effect takes place in such materials, each low-energy bulk Dirac mode contributes $1/2$ to the Hall conductivity.[13] A theorem due to Nielsen and Ninomiya states that Dirac cones come in pairs, ensuring an integral conductivity.

The most celebrated material with a relativistic spectrum is graphene, where two Dirac cones sit at corners of the Brillouin zone. In the presence of a magnetic field, each Dirac fermion contributes $n + 1/2$ to the Hall conductivity. Taking into account spin degeneracy the Hall conductivity in graphene reads $\sigma_H = 4n + 2$. The most notable property of these plateaux is the vanishing of the trivial insulator phase $\sigma_H = 0$. This is a consequence of the low-energy Dirac mode shifting the Hall spectrum.

B. Topological insulators

Topological insulators are special phases of matter characterized by a gapped bulk material with gapless edge or surface modes.[15] These gapless modes are topologically protected in the sense that they are robust to

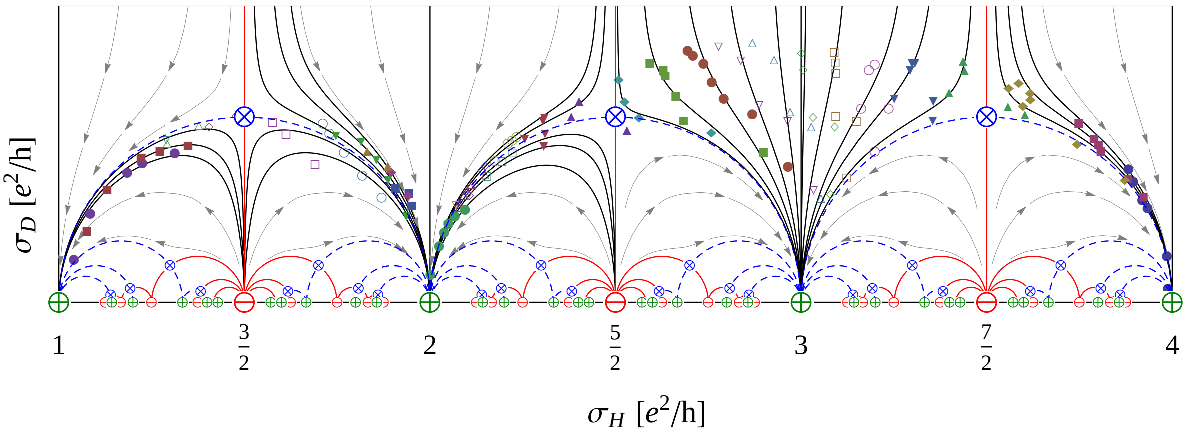


Figure 5. (Color online) Reconstruction of temperature-driven scaling data (discrete icons) exploring the plateaux transitions $1 = \oplus \overset{\otimes}{\longleftrightarrow} \oplus = 2 \overset{\otimes}{\longleftrightarrow} 3 = \oplus \overset{\otimes}{\longleftrightarrow} \oplus = 4$ in a semiconducting InGaAs/InP heterojunction.[14]

perturbations that preserve the symmetries of the system. The theory of topological insulators relies on Bloch theory as well as recent mathematical tools like Chern numbers and homotopy theory to characterize classes of Hamiltonians that preserve the bulk gap.

A normal insulator is said to be topologically trivial. The QED vacuum presents an insulator in this class. Here two bands are associated with electrons and positrons, while a large gap is associated with the pair production energy. The gapless surface modes of a topological insulator appear as a necessary consequence of a topologically non-trivial material ending on a trivial one (e.g. the vacuum). The only way a topological property can change across the interface is for the gap to close. This relation between bulk topology and edge modes are called the bulk-edge correspondence (bulk/edge duality).

The first topological insulator to be discovered was the IQHE itself. Here the Landau levels serve as energy bands, while a strong magnetic field induces a gap up to the first empty level. The bulk-boundary correspondence is in this case attributed to electrons skipping along the edges of the Hall sample due to the magnetic field. In this case it is not a material that is considered a topological insulator but the IQHE as a whole.

Depending on the material in which the Hall effect takes place, different imprints are seen on the Hall conductivity. Graphene, for example, has a unique Hall spectrum $\sigma_H = 4n + 2$ due to its two Dirac cones.

Another example is provided by the surface of a 3-dimensional topological insulator, which can serve as an effective 2-dimensional arena for the QHE. The bulk-boundary correspondence tells us that this surface has massless excitations. Depending on the what kind of bulk topology the surface Brillouin zone has, either an even or an odd number of Dirac cones are present,[15] and the effective 2-dimensional material can be seen as a Dirac material. In the case of an odd number of Dirac cones the Nielsen-Ninomiya theorem appears to be broken. This is solved by the existence of partner Dirac fermions at the opposite surface of the 3-dimensional topological insulator.[15] Under the assumption that the two sides are independent the Hall conductivity will be a sum of both contributions.

IV. UNIVERSALITY CLASS Γ_{T}

A. Plateaux transitions in InGaAs/InP

The result of the first scaling experiment in the context of the QHE, obtained in 1985 using a semiconducting heterojunction cooled below 4.2 K,[14] is reconstructed in Fig. 5 from the published data. Clear indications of a modular symmetry are already evident in this diagram (compare Fig. 2), even with the large uncertainty in the data.

Fig. 5 shows our reconstruction of temperature-driven scaling data (discrete icons) exploring the plateaux transitions $1 = \oplus \overset{\otimes}{\longleftrightarrow} \oplus = 2 \overset{\otimes}{\longleftrightarrow} 3 = \oplus \overset{\otimes}{\longleftrightarrow} \oplus = 4$ in a semiconducting InGaAs/InP heterojunction with 2D electron density $n = 3.4{\times}10^{11}\,\mathrm{cm}^2$, mobility $\mu = 35\,000\,\mathrm{cm}^2/\mathrm{Vs}$ and effective mass $m^* = 0.047\,m_e$, in the temperature range 4.2 K (top) to 0.5 K (bottom).[14]

Comparison with a modular scaling flow (solid lines) with quantum critical points at $\otimes = 1/2, 3/2$ and $5/2$ reveals a Γ_{T}-symmetry in the transport data (compare Fig. 2).

In the three decades following this pioneering experiment technology has improved and error bars have shrunk. In the following we shall see that not only have experiments failed to contradict the symmetry, the agreement with the coldest experiments, where the symmetry is expected to be most accurate, is now in some cases at the per mille level.

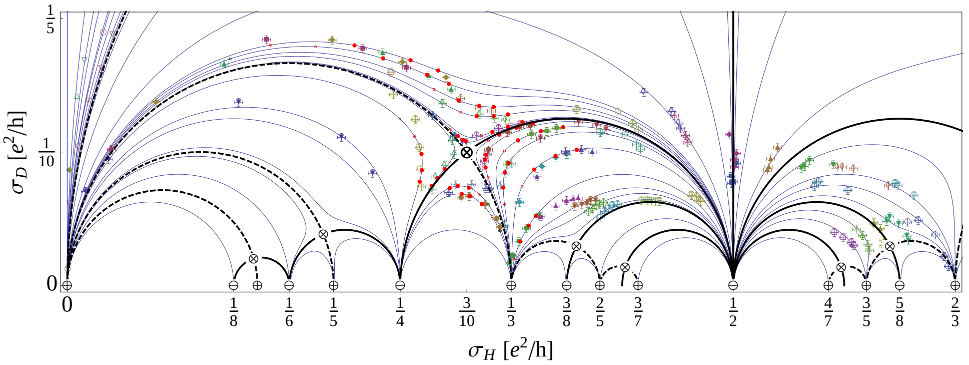


Figure 6. (Color online) (placeholder) Fractional scaling-flow in a GaAs/GaAlAs heterojunction.[16]

B. Plateaux transitions in GaAs/GaAlAs

Figs. 6 and 7 provide further evidence for the existence of a universality class with Γ_T-symmetry that unifies the IQHE (Fig. 7) with the FQHE (Fig. 6).

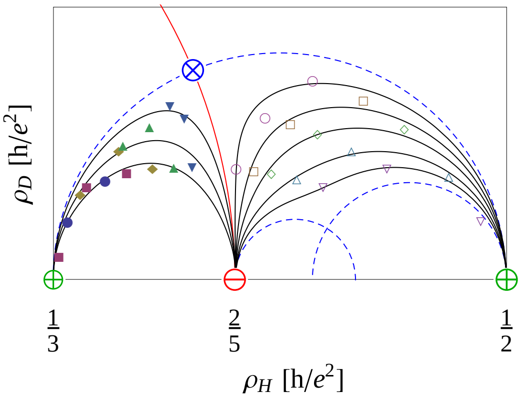


Figure 7. (Color online) (placeholder) Scaling flow between integer plateaux in a GaAs/GaAlAs heterojunction.[17]

C. Plateau-insulator transition in Cr(BiSb)Te

The QHE can take place on the top of 3-dimensional topological insulators,[18] like bismuth antimonide $Bi_{1-x}Sb_x$ which was the first 3-dimensional topological insulator to be discovered.[19] The surface Brillouin zone of this material has a single Dirac cone, seemingly contradicting the Nielsen-Ninomiya theorem. This problem is remedied by the existence of a partner Dirac fermion at the bottom of the topological insulator.[18] In this way, the surface of the 3-dimensional topological insulator can be seen as a sort of effective 2-dimensional Dirac material. The effective edges of these 2-dimensional surface systems are magnetic domain walls along which the charge carriers move. In total the conductivity reads $\sigma_H^{\mathrm{top}} + \sigma_H^{\mathrm{bottom}} = n_t + n_b + 1$. If the two Dirac fermions contribute equally to the conductivity, the Hall spectrum consists of odd integers. Also in this case the trivial insulator phase is evaded.

Fig. 8 and 9 show our reconstruction of temperature-driven scaling data (discrete icons) exploring the plateau-insulator transition $0 = \circledast \overset{\otimes}{\longleftrightarrow} \oplus = 1$ in a 2D ferromagnetic topological insulator (thin film of $Cr_x(Bi_{1-y}Sb_y)_{2-x}Te_3$ grown on a semi-insulating InP (111) substrate).[20] After applying an external magnetic field $B = 14\,\mathrm{T}$ to saturate the magnetization, the magnetic field strength was set to zero and experiments were performed at different temperatures with tunable gate voltage. In order to compensate for what is presumably a systematic error of unknown origin, the data in Fig. 8 has been shifted slightly to the left so that the plateaux are integer-valued.

In both cases, comparison with a modular scaling flow (solid lines) with a quantum critical point at $\otimes = (1+i)/2$ (compare Fig. 2) reveals that these transport data are in excellent agreement with Γ_T-symmetry.

D. Plateaux transitions in mercury telluride

Bulk mercury telluride is a semi-conductor of the II-VI type,[21] but when used to create a quantum well (HgCdTe/HgTe/HgCdTe) the electronic properties depend crucially on the thickness d of the sample. This thickness introduces a parameter which can be tuned to find quantum phase transitions. For thin wells with thickness below the critical thickness $d_c \approx 6.3\,\mathrm{nm}$ the material has a normal band structure, whereas for wide wells ($d > d_c$) the band structure is inverted.[21,22] At critical thickness the band gap closes and a single Dirac cone appears in the Brillouin zone.

In addition to having a highly specific energy spectrum with an inverted band structure, the 2DEG in a wide HgTe quantum well is characterized by a low effective mass, $m^* = 0.02\,m_e$ (m_e is the electron mass).[23] The low effective mass causes a large Landau level separation

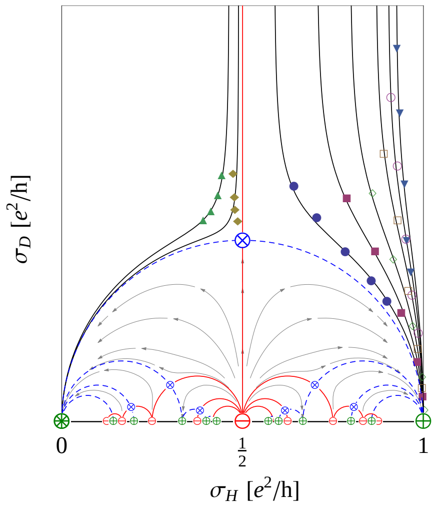


Figure 8. (Color online) Reconstruction of temperature-driven scaling data (discrete icons) exploring the plateau-insulator transition $0 = \circledast \overset{\otimes}{\longleftrightarrow} \oplus = 1$ in a 2D ferromagnetic topological insulator (a thin film of $\mathrm{Cr}_x(\mathrm{Bi}_{1-y}\mathrm{Sb}_y)_{2-x}\mathrm{Te}_3$ grown on a semi-insulating InP(111) substrate).[20]

$\Delta E = \hbar qB/m^*c$, and the QHE survives to relatively high temperatures. In [24 and 25] a strong integer effect was observed up to $T \sim 10 - 15\,\mathrm{K}$.

Fig. 10 shows our reconstruction of temperature-driven scaling data (discrete icons) exploring the plateaux transitions $1 \overset{\otimes}{\longleftrightarrow} 2 \overset{\otimes}{\longleftrightarrow} 3$ in a heterostructure $\mathrm{Hg}_x\mathrm{Cd}_{1-x}\mathrm{Te}/\mathrm{HgTe}/\mathrm{Hg}_x\mathrm{Cd}_{1-x}\mathrm{Te}$ ($x \approx 0.7$) with a 20.3 nm wide HgTe quantum well.[25] Since this thickness is well above d_c there should be no Dirac cones in the bulk Brillouin zone. The sample was grown by molecular beam epitaxy on a GaAs substrate, symmetrically modulation doped with In at both sides of the quantum well, yielding a mobility of 22 m^2/Vs and an electron gas density of about 1.5×10^{15} m^2.[25,26]

The longitudinal and Hall resistivities were measured with a constant 1 A current in the temperature range $2.9 - 50\,\mathrm{K}$, and a magnetic field strength in the $0 - 9\,\mathrm{T}$ range. There is clear evidence for plateaux at $\nu = 1, 2, 3$ and 4, obtained for magnetic fields in the range $1.8 - 8\,\mathrm{T}$. For most mangetic field values the system exhibited scaling behaviour for the five lowest temperatures $T = 2.9, 4.1, 6.1, 8.1$ and 10 K, and in one instance also for 15 and 20 K. In some cases, close to the fix points only the three lowest temperatures were usable.

Comparison with a modular scaling flow (solid lines) with quantum critical points at $\otimes = (2n + 1 + i)/(2n^2 + 2n + 1) = 1 + i$, $(3 + i)/5$, $(5 + 1)/13$, $(7 + i)/25$, $\ldots$ reveals a Γ_{T}-symmetry in the transport data (compare Fig. 2).

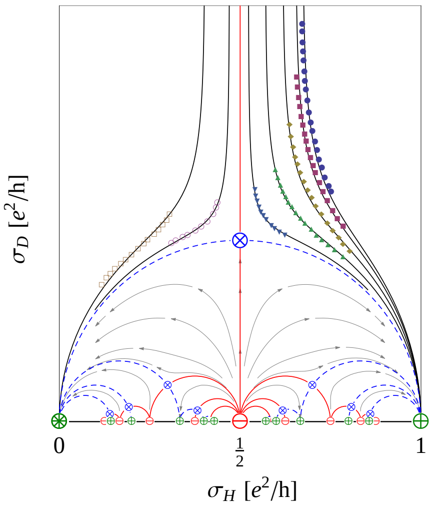


Figure 9. (Color online) Reconstruction of temperature-driven scaling data (discrete icons) exploring the plateau-insulator transition $0 = \circledast \overset{\otimes}{\longleftrightarrow} \oplus = 1$ in a 2D ferromagnetic topological insulator ((a thin film of $\mathrm{Cr}_x(\mathrm{Bi}_{1-y}\mathrm{Sb}_y)_{2-x}\mathrm{Te}_3$ grown on a semi-insulating InP(111) substrate).[20]

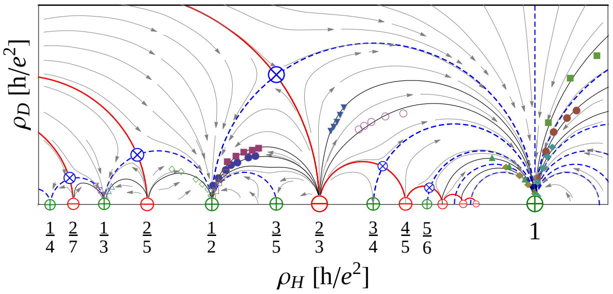


Figure 10. (Color online) Reconstruction of temperature-driven scaling data (discrete icons) exploring the plateaux transitions $1 \overset{\otimes}{\longleftrightarrow} 2 \overset{\otimes}{\longleftrightarrow} 3$ in a HgTe/HgCdTe heterostructure with a wide HgTe quantum well.[25]

E. Plateau-insulator transitions in bismuth antimonide

In [27] the QHE was studied by measuring surface conductivities on the top and bottom of a 3-dimensional topological insulator, bismuth antimonide. Two 8 nm thick TI films of $(\mathrm{Bi}_{1-x}\mathrm{Sb}_x)_2\mathrm{Te}_3$ ($x = 0.84, 0.88$) were grown on insulating InP (111) substrates using molecular beam epitaxy. Quantum Hall signatures were found at magnetic field strengths above 14 T, for temperatures ranging from 700 mK down to 40 mK, at various gate voltages V_G.

Fig. 11 shows our reconstruction of their temperature-driven scaling data (discrete icons) exploring the two plateau-insulator transitions $(-1 = \oplus \overset{\otimes}{\longleftrightarrow} \circledast \overset{\otimes}{\longleftrightarrow} \oplus = 1)$.

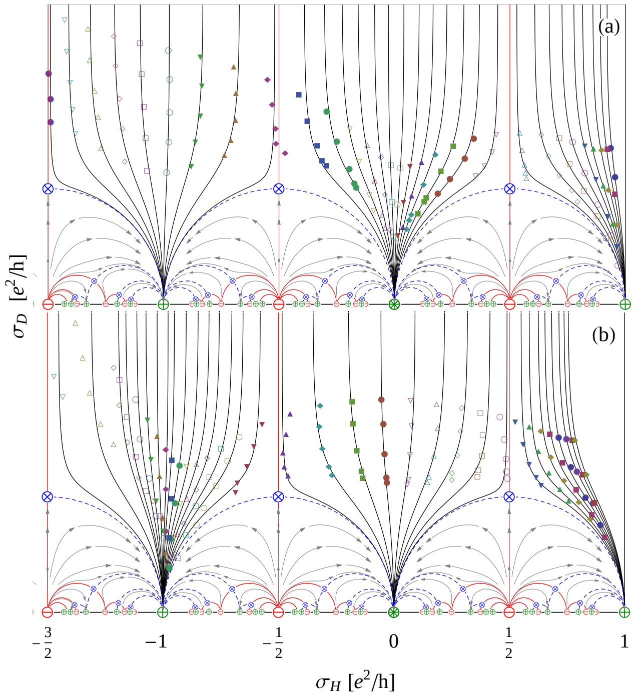


Figure 11. (Color online) Reconstruction of temperature-driven scaling data (discrete icons) exploring the plateaux-insulator transitions $-1 = \oplus \overset{\otimes}{\longleftrightarrow} \circledast \overset{\otimes}{\longleftrightarrow} \oplus = 1$ in a bismuth antimonide topological insulator $(\mathrm{Bi}_{1-x}\mathrm{Sb}_x)_2\mathrm{Te}_3$, with (a) $x = 0.88$, and (b) $x = 0.84$.[27]

Inaccessible data points and clear statistical outliers were not considered when sampling the data.

Comparison with a modular scaling flow (solid curves) with quantum critical points at $\otimes = (\pm 1 + i)/2$, $(\pm 3 + i)/2, \ldots$ reveals a Γ_{T}-symmetry in the transport data (compare Fig. 2).

F. Plateaux transitions in black phosphorus

In addition to graphene, black phosphorus is only other 2D atomic crystal discovered with a QHE.

Fig. 12 shows a scaling flow in two dimensional black phosphorus. In the experiment by Li et. al. few-layer black phosphorus was sandwiched between two layers of hexagonal boron nitride (hBN) and placed on a graphite back-gate to create a van der Waals heterostructure. The thin bottom hBN layer of ~25 nm allows for the electrons in the graphite to screen the impurity potential in at the black phosphorus-hBN interface, bringing the Hall mobility up to 6000 cm^{2}V^{-1}s^{-1}. This large mobility allows for the detection of a QHE in this material.[28]

The data is extracted from Fig. 7 in the supplementary material of [28]. Experiments measuring longitudinal and Hall resistances were made at fixed magnetic fields of 27, 29, 31 and 33 T and temperatures 1.7, 4.1, 4.6, 6, 8 and 10 K by varying the back gate voltage from -2 to -0.7 V. Plateaus were discovered for filling factors $\nu = 1$, 2 and 3. Due to overlap of the Hall resistance curves, an area of ~ ±013 V about the apparent infliction point of the 1-2 transition had to be excluded. The curves for 8 and 10 K were also excluded on condition that direct resistance should drop to zero at a plateau whereas they in every case exceeded 1 kΩ for the $\nu = 1$ plateau.

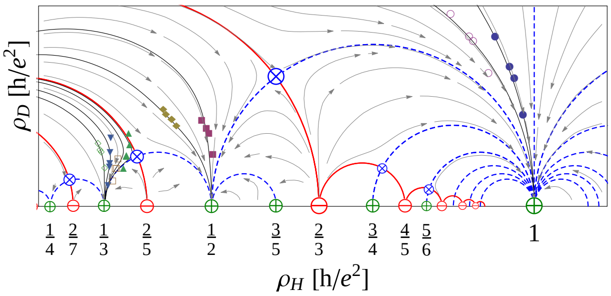


Figure 12. (Color online) Reconstruction of temperature-driven scaling data (discrete icons) exploring the plateaux transitions $1 \overset{\otimes}{\longleftrightarrow} 2 \overset{\otimes}{\longleftrightarrow} 3$ in black phosphorus.[28]

As the direct resistivity is determined by $\rho_D = (L_y/L_x)R_D$ and the aspect ratio (L_y/L_x) of the Hall bar was not given, it has been chosen equal to 3 for a best fit of the data. This value for the ratio is consistent with the optical image of the black phosphorus/hBN/graphite heterostructure given in Fig. 1 (a) of [28].

Comparison with a modular scaling flow (solid lines) with quantum critical points at $\otimes = (2n+1+i)/(2n^2+2n+1) = 1+i$, $(3+i)/5$, $(5+1)/13$, $(7+i)/25, \ldots$ reveals a Γ_{T}-symmetry in the transport data (compare Fig. 2)

V. UNIVERSALITY CLASS Γ_{R}

A. Plateau-insulator transition in GaAs/GaAlAs

Fig. 13 shows our reconstruction of temperature-driven scaling data (discrete icons) exploring the plateau-insulator transition $0 = \circledast \overset{\otimes}{\longleftrightarrow} \oplus = 2$ in a GaAs/GaAlAs heterojunction.[29] Comparison with a modular scaling flow (solid lines) with a quantum critical point at $\otimes = 1+i$ reveals a Γ_{R}-symmetry in the transport data (compare Fig. 2).

B. Plateau-insulator transition in graphene

Fig. 14 shows our reconstruction of temperature-driven scaling data (discrete icons) exploring the plateau-insulator transition $0 = \circledast \overset{\otimes}{\longleftrightarrow} \oplus = 2$ in graphene.[30] In

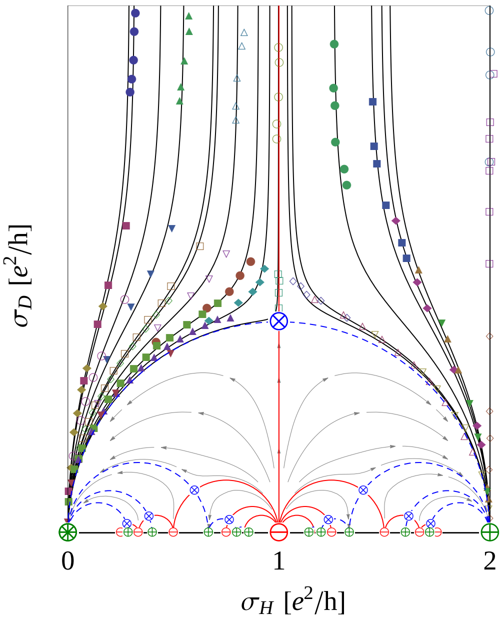


Figure 13. (Color online) Reconstruction of temperature-driven scaling data (discrete icons) exploring the plateau-insulator transition $0 = \circledast \overset{\otimes}{\longleftrightarrow} \oplus = 2$ in a GaAs/GaAlAs heterojunction.[29]

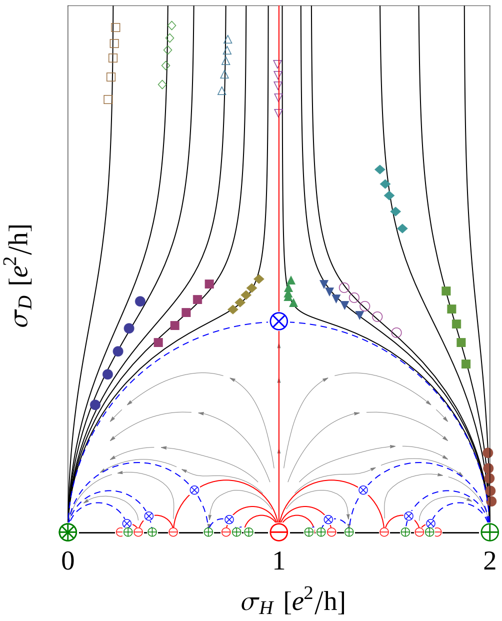


Figure 14. (Color online) Reconstruction of temperature-driven scaling data (discrete icons) exploring the plateau-insulator transition $0 = \circledast \overset{\otimes}{\longleftrightarrow} \oplus = 2$ in graphene.[30]

order to compensate for what is presumably a systematic error of unknown origin, the dataset close to the dashed blue semi-circle has been shifted up by 2%, so that the flow does not violate the semi-circle law (i.e., so that the flow does not cross the separatrix connecting the plateau $\oplus$ to the insulator $\circledast$ via the critical point $\otimes$). Comparison with a modular scaling flow (solid lines) with a quantum critical point at $\otimes = 1 + i$ reveals a Γ_{R}-symmetry in the transport data (compare Fig. 2).

In this experiment large-area $(0.6 \times 0.1\,\mathrm{mm}^2)$ monolayer graphene devices were made by epitaxial growth on SiC-substrate. In the devices, a buffer layer of graphene made partial covalent bonds with the exposed Si atoms and only the top graphene layer was conducting. Experiments were made in the temperaure range $2.6 - 25$ K with magnetic fields in the range $0.1 - 9$ T.

According to the authors, Si-C covalent bonds and defects such as interfacial dangling bonds affect the electrical environment of the graphene sheet and graphene-substrate coupling might break its sublattice symmetry. In addition, the carrier density was engineered as low as $n \approx 10^{15}\mathrm{m}^{-2}$, and low carrier density reduces the screening of Coulomb potential fluctuations, thus enhancing the SiC substrate effect on the conducting graphene sheet.[30] This may be the reason for the Γ_{R} symmetry and the resulting insulator phase distinguishing this system from the usual case of monolayer graphene with no 0-plateau and Γ_{S} symmetry.

The data that best fit the flowlines are taken from one of the least disordered samples which also had the highest surface roughness (called EG2 in [30]). Data taken from the other sample (EG3) has been modified by a constant shift of $0.03e^2/h$ in the positive σ_D-direction, to make it fit the flowlines perfectly. The need for this shift may be the result of a small systematic error in the experiment, however the shift in itself may be well within the random error of the experiment.

VI. UNIVERSALITY CLASS Γ_{S}

Plateau-plateau transition $(-1 = \oplus \overset{\otimes}{\longleftrightarrow} \oplus = 1)$, obtained by measuring surface conductivities of the bismuth antimonide topological insulator $(\mathrm{Bi}_{1-x}\mathrm{Sb}_x)_2\mathrm{Te}_3$.[27]

VII. UNIVERSALITY CLASS Γ_{Q}

We have already mentioned the spectrum of plateaux observed in some experiments on graphene. The competition between several scales is not easy to disentangle, especially in crossover regions where the lowest Landau level may be more resistant to symmetry breaking than the higher levels. However, so far it seems that the symmetries we have discussed (compare Fig. 1) suffice to account for the plateau-data.

A much more stringent test is, as we have seen in the non-relativistic case, to compare the unstable fixed points with experimental quantum critical points. Scaling experiments on graphene are still in their infancy, and the paucity of data means that this analysis is far from conclusive. Unfortunately, so far a meaningful comparison is only possible for the doubly degenerate IQHE, which should be compared with the phase- and flow diagram in Fig. 2.

Because of the zero-mode there is no QHI ($\oplus_\sigma = \circledast = 0$) in this case, so Γ_T and Γ_R are immediately eliminated as potential symmetries. A glance at the defining characteristics of the groups in eq.(1) shows that, up to a factor of two, Γ_S is the only viable candidate. Because of degeneracy the observed conductivity should be doubled,[31] $\sigma \to 2\sigma = G(\sigma)$, giving the Γ_{Q}-symmetric phase- and flow-diagram shown in the bottom panel of Fig. 2.

An immediate consequence is that fractional plateaux in the doubly degenerate QHE should appear only at $\sigma_H = 2(2n+1)/(2m+1) \not\ni \pm 1/3$. In fact, $\sigma_H = 1/3$ has also been observed, but only when the magnetic field is so strong that one expects the spin-valley degeneracy to be lifted.

A. IQHE in graphene

Fig. 15 is a reconstruction of some experimental quantum Hall data for graphene,[32–34] compared with modular critical points (blue $\otimes$). As explained in Section II, ideally we would like to have a family of scaling data deep inside the scaling domain, in which case we could obtain the experimental critical point from the temperature independent crossing point of the curves. Unfortunately such data are still not available for graphene. The family of data published recently are consistent with our estimate, but not good enough to resolve any discrepancy in detail.[34] This is why only the data obtained at the lowest temperature (4.1 K) has been used in Fig. 15(c).

In lieu of such "family portraits" the following symmetry argument has been used to extract the experimental critical points shown in Fig. 15. Since the transitions probed in these experiments are between integer plateaux, we expect the critical points in the conductivity plane to be precisely half way between two plateaux, shifted horizontally by $4e^2/h$ relative to its neighbours. If we are close enough to the scaling domain the critical point should therefore lie on the vertical lines with $\sigma_H = 4n$ (compare Fig. 2). Mapping these lines to the experimental plots in the resisitivity plane gives the continuous (red) semi-circles shown in Fig. 15(d). The points where an experimental graph crosses these arcs is therefore our best estimate for the location of critical points. These are the values of $\rho(\otimes)$ used in Fig. 15(a-c) to identify the critical values of the magnetic field.

Fig. 15 (a) is our reconstruction of the first data on the $2 \overset{\otimes}{\longleftrightarrow} 6$ transition, discovered by Zhang et al. in 2005.[32,35]

Fig. 15 (b) shows the $2 \overset{\otimes}{\longleftrightarrow} 6 \overset{\otimes}{\longleftrightarrow} 10 \overset{\otimes}{\longleftrightarrow} 14$ transitions discovered in 2009.[33] The latter two are magnified in the inset, but the distinction between experimental and Γ_{Q} critical points is still not resolved in this plot.

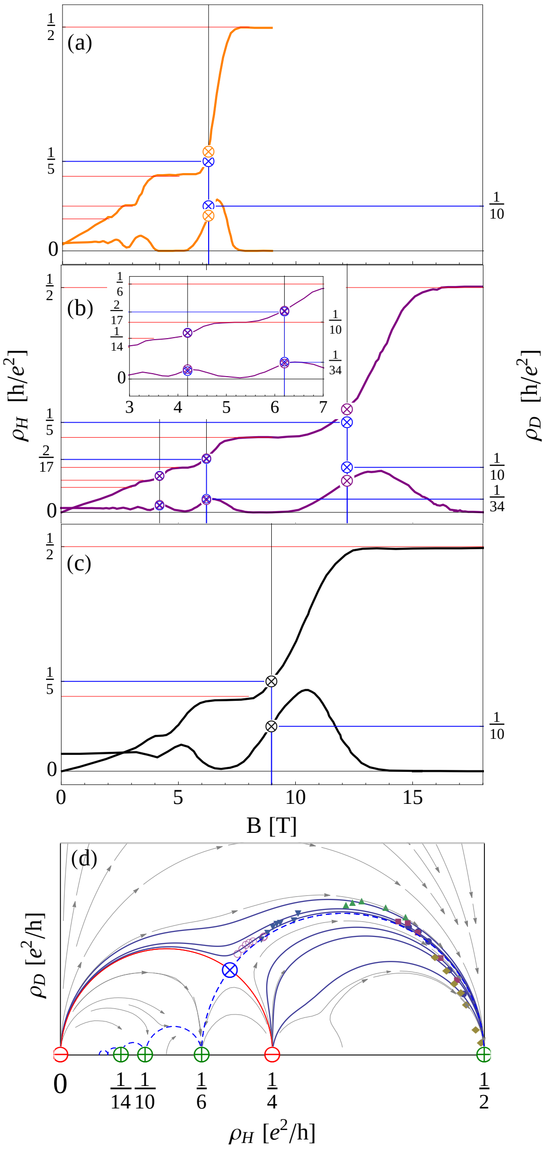


Figure 15. (Color online) (a-c) Experimental quantum Hall data for graphene reconstructed from [32–34], compared with modular critical points (blue $\otimes$). (d) Scaling flow derived from reconstructed graphene data published in [34], superimposed on the phase diagram with Γ_{Q}-symmetry (compare Fig. 2).

Fig. 15 (c) shows more recent data on the $2 \overset{\otimes}{\longleftrightarrow} 6 \overset{\otimes}{\longleftrightarrow} 10$ transitions.[34] In this case the fixed point of Γ_{Q} is completely eclipsed by the experimental critical point. In all cases the overlap of experiment and theory is reasonable, and possibly within experimental error, although no error analysis has been provided by these authors.

Fig. 15 (d) shows a scaling flow derived from reconstructed graphene data published in [34], superimposed on the phase diagram with Γ_{Q}-symmetry (compare Fig. 2).

We see that, at least for the first transition, it is possible that the earlier experiments had not reached the scaling limit, which is where an approximate low-energy symmetry would appear. The good agreement with the most recent data in Fig. 15 (c) notwithstanding, it is premature to claim that these experiments unambiguously demonstrate the emergence of a modular symmetry in graphene. This question can only be settled by more accurate experiments involving transitions to fractional plateaux.

B. Spin-resolved spectrum

The spectrum of observed plateaux in graphene changes when the magnetic field is increased above about 10 T. New integer plateaux at $\sigma_H = 0, \pm 1, 3, \pm 4$ have been reported.[36,37] As already mentioned fractional values at $\sigma_H = 1/3, \ldots$, which do not fit the symmetry of doubly degenerate graphene, have also been reported.[37–39]

When the field is this strong it is no longer a good approximation to ignore the Zeeman splitting, and this is presumably the reason the spectrum changes. The crossover region is difficult to disentangle, but no obvious contradiction with modular symmetry is apparent.

C. FQHE in graphene

Since they were discovered in 2009 many fractional plateaux have been found in graphene.[37,38] A recent study found some intriguing new fractional plateaux in graphene:[39]

$$\sigma_H / G_K = \begin{cases} \frac{1}{3}, \frac{2}{3}, \frac{2}{5}, \frac{3}{5}, \frac{3}{7}, \frac{4}{7}, \frac{4}{9} & \text{for } 0 < \nu < 1 \\[2ex] \frac{4}{3}, \frac{8}{5}, \frac{10}{7}, \frac{14}{9} & \text{for } 1 < \nu \end{cases}$$

The first sequence is consistent with Γ_{T}, in which case both the spin and pseudo-spin has been resolved. Barring coincidences, the second sequence appears to be constrained to have only even numerators. Since $4/3, 8/5 \notin \oplus_{\mathrm{Q}}$, the only possibility appears to be Γ_{R}, which has plateaux

$$\oplus_{\mathrm{R}} = \frac{2n}{2m+1} \ni \frac{4}{3}, \frac{8}{5}, \frac{10}{7}, \frac{14}{9}, \ldots .$$

A possible interpretation is that for higher levels the pseudo-spin symmetry has been completely broken (with no surviving Dirac cones), while the spin remains degenerate. This is consistent with the first sequence appearing for strong fields and the second sequence appearing at lower fields, and also with the theoretical expectation that the lowest level will be most susceptible to symmetry breaking.

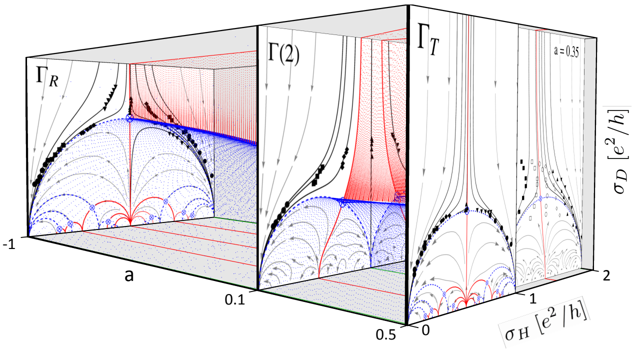


Figure 16. (Color online) Reconstruction of temperature-driven scaling data (discrete icons) exploring the transitions $0 = \circledast \overset{\otimes}{\longleftrightarrow} \oplus = 1 \overset{\otimes}{\longleftrightarrow} \oplus = 2$ in GaAs with self-assembled InAs dots, for various values of the spin-splitting (parametrized by a), which was tuned using a backgate voltage.[40] All flow lines (solid curves) are theoretical. They were derived by numerical integration from the unique one-parameter family φ_a of $\Gamma(2)$-invariant RG-potentials, as discussed in Sects. II C and VIII.

VIII. REDUCED MODULAR SYMMETRY

We turn now to an experiment that explored the transition from degenerate (unpolarized) to non-degenerate (fully polarized/spin-split) bands, by tuning the spin-splitting using a backgate voltage. By the arguments discussed in the first section, we expect these data to interpolate between the two maximal submodular symmetries Γ_{R} (unpolarized) and Γ_{T} (polarized). When the Zeeman splitting is between these extremes the modular symmetry must be at least partially broken, but possibly only to the main congruence group $\Gamma(2)$.

Fig. 16 shows a reconstruction of temperature-driven scaling data (discrete icons) exploring the transitions $0 = \circledast \overset{\otimes}{\longleftrightarrow} \oplus = 1 \overset{\otimes}{\longleftrightarrow} \oplus = 2$ in GaAs with self-assembled InAs dots.[40] The transition from degenerate (unpolarized) to non-degenerate (fully polarized/spin-split) bands is explored by tuning the spin-orbit interaction using a backgate voltage, and compared to the family of physically viable $\Gamma(2)$-invariant RG-potentials (compare Sect. II) $\varphi_a \propto \ln\lambda + (a-1)\ln(\lambda - 1)$,[11] with values of the real parameter a ranging from $a_{\mathrm{R}} = -1$ to $a_{\mathrm{T}} = 1/2$ in this experiment. All solid lines are flow trajectories derived by numerical integration from the gradient flow generated by this potential. For clarity we display only those parts of the modular phase boundaries (red sheets) that are above all separatrices (blue canopies).

By comparing the data for the $0 \overset{\otimes}{\longleftrightarrow} 1$ transition with the flow derived from $\varphi_{1/2}$ (left hand side of front panel), we see that the scaling flow in this case appears to respect Γ_{T}-symmetry. This is not so for the $1 \overset{\otimes}{\longleftrightarrow} 2$ transition. A partial fit is obtained at $a \approx 0.35$ (in order to make the flow visible from this viewing angle it is shown on

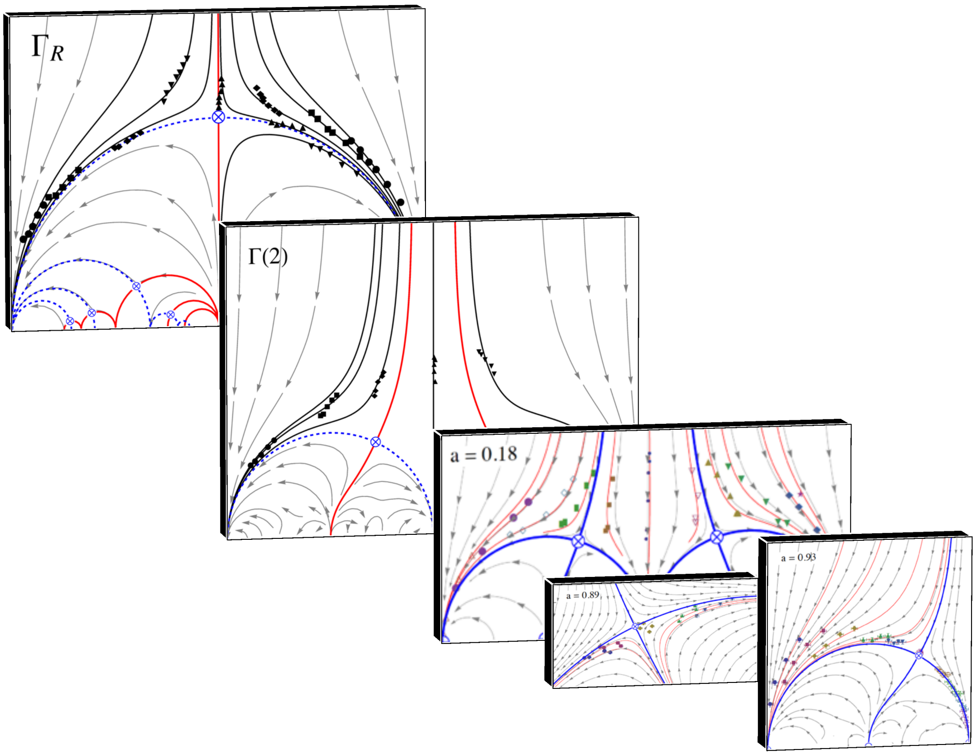


Figure 17. (Color online) (placeholder) Reconstruction of temperature-driven scaling data (discrete icons) exploring various parts of the landscape of $\Gamma(2)$-symmetric scaling flows, derived from a wide range of different 2D materials. (Not to scale, compare Figs. 4 and 16)

a panel only slightly recessed behind the panel at $a = 1/2$ showing the $0 \overset{\otimes}{\longleftrightarrow} 1$ transition), but several of the experimental flow-lines still cross the separatrix (dashed blue semi-circle).

This is a rare example where the $\Gamma(2)$-symmetry appears to be broken, presumably due to the intervention of new physics that is not relevant for the other experiments. It is conceivable that some (maximal?) subgroup of $\Gamma(2)$ has survived, but we have insufficient data to investigate this. It is also conceivable that a systematic error of unknown of origin is responsible, but we have no way of investigating this either. New physics would be more interesting, and it could aid in the construction of a phenomenological function $a = a(B, T, \dots)$ (the dots include material properties that are capable of breaking modular symmetry), which could be used to predict which type of modular symmetry (if any) that is to be expected in the transport coefficients of new materials.

Finally, Fig. 17 summarizes other work not discussed in detail here.[11] Our reconstruction of temperature-driven scaling data (discrete icons), derived from a wide range of different 2D materials, explores various parts of the landscape of $\Gamma(2)$-symmetric scaling flows. As in all our diagrams, solid lines are flow trajectories derived by numerical integration from the gradient flow generated by the RG-potential φ_a.

IX. DISCUSSION

In addition to the fact that some modular predictions have been verified at the *per mille* level, it is perhaps the overall agreement of the unique modular family of level two flow diagrams with a wide range of different materials and experimental circumstances that is the most convincing evidence for "modular universality" in the QHE.

There are some data that appear to disagree with the modular symmetries discussed here. Most, if not all, are "transition materials" that appear to be a superposition of two symmetries. We offer a plausible explanation for one of these anomalies.

Fig. 18 presents a conjectured modular explanation of a peculiar crossover observed in graphene[41]. They appear to find that an insulator phase can inject itself into the standard graphene sequence $-6 \overset{\otimes}{\longleftrightarrow} -2 \overset{\otimes}{\longleftrightarrow} 2 \overset{\otimes}{\longleftrightarrow} 6$, without being accompanied by other new plateaux: $-6 \overset{\otimes}{\longleftrightarrow} -2 \overset{\otimes}{\longleftrightarrow} \circledast \overset{\otimes}{\longleftrightarrow} 2 \overset{\otimes}{\longleftrightarrow} 6$. The QHI at $\sigma_{\circledast} = 0$ means that the original Γ_{Q} has been broken to Γ_{R}:

$$\sigma(\otimes)_Q = 2i \xrightarrow{\text{splits}} \sigma(\otimes)_R = \pm 1 + i,$$

or equivalently, to $\rho(\otimes)_R = (\pm 1 + i)/2$. If so, there should be more structure, signalling new plateaux, emerging in the $\pm 2 \overset{\otimes}{\longleftrightarrow} \pm 6$ transitions, etc. Even if both the new

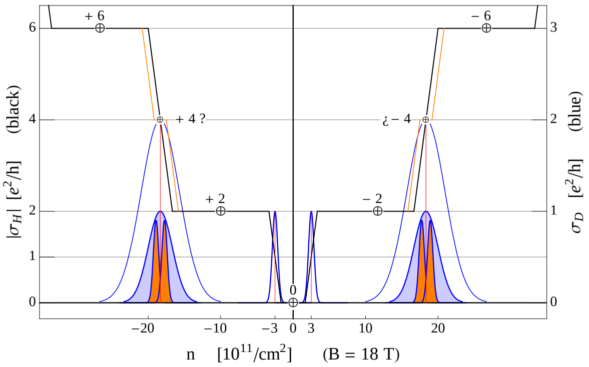


Figure 18. (Color online) (placeholder) Cartoon of data obtained in [41], illustrating how peaks in σ_D can be suppressed by poorly articulated plateaux that still are not visible in σ_H, as seen in this experiment probing the cross-over $\Gamma_\mathrm{Q} \to \Gamma_\mathrm{R}$ (compare Fig. 3).

plateau and the new zero in σ_D are insufficiently developed to be visible, the new zero in σ_D that eventually develops at $\sigma_\oplus = \pm 4$ will force the maximum value of σ_D to shrink. In other words, when a critical point "splits" in order to make room for a new phase, the presence of this new pair of crtical points could at first appear as a suppression of the old peak, as is seen in this experiement. When the plateau is fully developed there will be two peaks instead of one, both smaller than the original peak, compare location of critical points in Fig. 2.

Arguably, the biggest outstanding problem in the QHE is to determine the value(s) of the critical (delocalization) exponent(s), which would completely nail down the quantum Hall universality class(es). This exponent is determined by curvature of the RG-potential at a critical point, and therefore depends on the normalization of the β-function, which does not follow from symmetry alone. Finding it requires information about the dynamics of the collective (emergent) modes relevant at low energy.

A useful analogy is the Ising model. Kramers and Wannier managed to calculate the exact value the critical temperature (location of the critical point) by exploiting a $\mathbb{Z}_2$-duality that is similar to S-duality acting on $\Im\sigma = \sigma_D$, but the value of the critical exponent remained beyond reach until Onsager solved the model completely.

However, while modular symmetry by itself is not sufficient to find the low-energy effective field theory, it does severly limit the supply of candidate models, and may therefore provide valuable assistance in the search for this theory.

[1] C. A. Lütken and G. G. Ross, "Duality in the quantum Hall system," Phys. Rev. B **45**, 11837 (1992).

[2] C. A. Lütken and G. G. Ross, "Delocalization, duality, and scaling in the quantum Hall system," Phys. Rev. B **48**, 2500 (1993).

[3] C. A. Lütken, "Geometry of renormalization group flows constrained by discrete global symmetries," Nucl. Phys. B **396**, 670 (1993).

[4] C. A. Lütken, "Global phase diagrams for charge transport in two dimensions," J. Phys. A: Math. Gen. **26**, L811 (1993).

[5] F. Diamond and J. Schurman, *A First Course in Modular Forms*, Graduate Texts in Mathematics; 228 (Springer Verlag, New York, NY, USA, 2005).

[6] In mathematics the conventional names for these groups are $\Gamma_0(2) = \Gamma_\mathrm{T}$, $\Gamma^0(2) = \Gamma_\mathrm{R}$, and $\Gamma_\theta(2) = \Gamma_\mathrm{S}$, but our notation is simpler and more informative, as it uses one of the generators to label the group.

[7] H. Petersson, "Über die kongruenzgruppen der Stufe 4," JRAM **212**, 63 (1963).

[8] R. Rankin, *Modular Forms and Functions* (Cambridge University Press, Cambridge, 1977).

[9] D. Zagier, *The 1-2-3 of Modular Forms*, Annals of Mathematics Studies No. 97 (Springer Verlag, 2008) Lectures at a Summer School in Nordfjordeid, Norway.

[10] C.P. Burgess and B.P. Dolan, Phys. Rev. B **76**, 113406 (2007).

[11] J. Nissinen and C. A. Lütken, "Renormalization-group potential for quantum Hall effects," Phys. Rev. B **85**, 155123 (2012).

[12] J. Nissinen and C. A. Lütken, "The quantum Hall curve," (2012), arXiv:1207.4693v1 [cond-mat.str-el].

[13] E. Fradkin, *Field Theories of Condenced Matter Physics* (Cambridge university press, 2013).

[14] H.P. Wei, D.C. Tsui, and A.M.M. Pruisken, "Localization and scaling in the quantum Hall regime," Phys. Rev. B **33** (1985).

[15] M. Z. Hasan and C. L. Kane, "Topological insulators," (2010), arXiv:1002.3895v2 [cond-mat.mes-hall].

[16] S.S. Murzin, S.I. Dorozhkin, D.K. Maude, and A.G.M. Jansen, Phys. Rev. B **72**, 195317 (2005).

[17] W. Li, C. L. Vicente, J. S. Xia, W. Pan, D. C. Tsui, L. N. Pfeiffer, and K. W. West, Phys. Rev. Lett. **102**, 216811 (2009).

[18] J. Wang, B. Lian, and S.-C. Zhang, "Quantum anomalous Hall effect in magnetic topological insulators," (2015), arXiv:1409.6715v4 [cond-mat.mes-hall].

[19] Y. Ando, "Topological insulator materials," J. Phys. Soc. Jpn. **82**, 102001 (2013), http://dx.doi.org/10.7566/JPSJ.82.102001.

[20] J.G. Checkelsky, R. Yoshimi, A. Tsukazaki, K.S. Takahashi, Y. Kozuka, J. Falson, M. Kawasaki, and Y. Tokura, "Trajectory of the anomalous Hall effect towards the quantized state in a ferromagnetic topological insulator," Nature Physics **10**, 731 (2014).

[21] M. Knig, S. Wiedmann, C. Brne, A. Roth, H. Buhmann, L. W. Molenkamp, X.-L. Qi, and S.-C. Zhang, "Quantum Spin Hall Insulator State in HgTe Quantum Wells," Science **318**, 766–770 (2007).

[22] B. A. Bernevig, T. L. Hughes, and S.-C. Zhang, "Quantum Spin Hall Effect and Topological Phase Transition in HgTe Quantum Wells," Science **314**, 1757–1761 (2006).

[23] E. B. Olahanetsky, S. Sassine, Z. D. Kvon, N. N. Mikhailov, S. A. Dvoretsky, J. C. Portal, and A. L. Aseev, "Quantum Hall liquid-insulator and plateau-to-plateau transitions in a high mobility 2DEG in a HgTe quantum well," Pis'ma v Zh. ksper. Teoret. Fiz. **84**, 661–665 (2006).

[24] Y. G. Arapov, S. V. Gudina, V. N. Neverov, S. M. Podgornykh, M. R. Popov, G. I. Harus, N. G. Shelushinina, M. V. Yakunin, N. N. Mikhailov, and S. A. Dvoretsky, "Temperature Scaling in the Quantum-Hall-Effect Regime in a HgTe Quantum Well with an Inverted Energy Spectrum," Semiconductors **49**, 1545–1549 (2015).

[25] S. V. Gudina, Y. G. Arapov, V. N. Neverov, S. M. Podgornykh, M. R. Popov, N. G. Shelushinina, M. V. Yakunin, S. A. Dvoretsky, and N. N. Mikhailov, "2D-localization and delocalization effects in quantum Hall regime in HgTe wide quantum wells," Phys. Status Solidi C **13**, 473–476 (2016).

[26] Y.G. Arapov, S.V. Gudina, V.N. Neverov, S.M. Podgornykh, M.R. Popov, G.I. Harus, N.G. Shelushinina, M.V. Yakunin, N.N. Mikhailov, and S.A. Dvoretsky, "Temperature scaling in the quantum-Hall-effect regime in a HgTe wide quantum well with an inverted energy spectrum," Semiconductors **49**, 1545 (2015).

[27] R. Yoshimi, A. Tsukazaki, Y. Kozuka, J. Falson, K.S. Takahashi, J.G. Checkelsky, N. Nagaosa, M. Kawasaki, and Y. Tokura, "Quantum Hall effect on top and bottom surface states of topological insulator $(\mathrm{Bi}_{1-x}\mathrm{Sb}_x)_2\mathrm{Te}_3$ films," Nat. Commun. 6:6627 doi: 10.1038/ncomms7627 (2015).

[28] L. Li, F. Yang, G. J. Ye, Z. Zhang, Z. Zhu, W. Lou, X. Zhou, L. Li, K. Watanabe, T. Taniguchi, K. Chang, Y. Wang, X. H. Chen, and Y. Zhang, "Quantum Hall effect in black phosphorus two-dimensional electron system," Nature Nanotechnology **11**, 593–597 (2016).

[29] S.S. Murzin, M. Weiss, A.G.M. Jansen, and K. Eberl, Phys. Rev. B **66**, 233314 (2002).

[30] L. I. Huang, Y. Yang, R. E. Elmquist, S. T. Lo, F. H. Liu, and C. T. Liang, "Insulator-quantum Hall transition in monolayer epitaxial graphene," RSC Adv. **6**.

[31] V.P. Gusynin and S.G. Sharapov, Phys. Rev. Lett. **95**, 146801 (2004).

[32] Y. Zhang, H.L. Stormer, and P. Kim, Nature **438**, 201 (2005).

[33] X. Wu, M. Ruan Y. Hu, N.K. Madiomanana, J. Hankinson, M. Sprinkle, C. Berger, and W.A. de Heer, Appl. Phys. Lett. **95**, 223108 (2009).

[34] M. Amado, E. Diez, F. Rossella, V. Bellani, D. Lopez-Romero, and D.M. Maude, J. Phys.: Condens. Matter **24**, 305302 (2012).

[35] K.S. Novoselov, A. Geim, S. Morozov, D. Jiang, M.I. Katsnelson, V. Grigorieva, and S.V. Dubonos, Nature **438**, 197 (2005).

[36] Y. Zhang, Z. Jiang, J.P. Small, M.S. Purewal, Y.-W. Tan, M. Fazlollahi, J.D. Cudlow, J.A. Jaszczak, H.L. Stormer, and P. Kim, Phys. Rev. Lett. **96**, 136806 (2006).

[37] X. Du, I. Skachko, F. Duerr, A. Lucian, and E.Y. Andrei, Nature **462**, 192 (2009).

[38] K.I. Bolotin, F. Ghahari, M.D. Schulman, H.L.Stormer, and P. Kim, Nature **462**, 196 (2009).

[39] B.E. Feldman, B. Krauss, J.H. Smet, and A.Yacoby, Science **337**, 1196 (2012).

[40] Y.-T. Wang, G-H. Kim, C. F. Huang, S.-T. Lo, W.-J. Chen, J. T. Nicholls, L.-H. Lin, D. A. Ritchie, Y. H. Chang, C.-T. Liang, and B. P. Dolan, "Probing temperature-driven flow lines in a gated two-dimensional electron gas with tunable spin-splitting," J. Phys.: Condens. Matter **24**, 405801 (2012).

[41] L. Zhang et al., Phys. Rev. Lett. **105**, 046804 (2010).

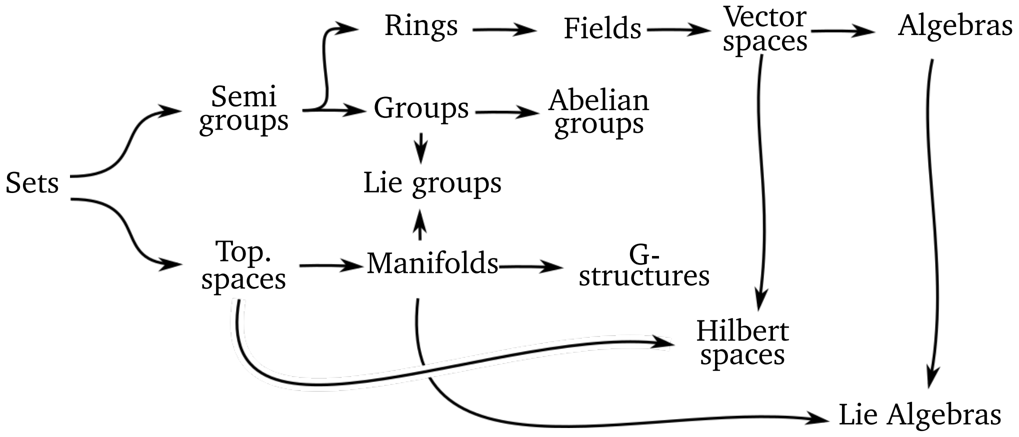
Appendices

A

Mathematical structures

A.1 A hierarchy of structure

We will encounter two large classes of structures in this work - the algebraic and the geometric. These are shown in the below figure, where roughly speaking the top half represents the algebraic and the bottom the geometric. Of course, these structures like to play, and often we will meet objects with several types of structure.



By algebraic structure we mean the following. Given a set S with elements s_i we can add structure to it by defining one or more binary operations that satisfy certain axioms. Binary operations are maps S × S → S that produce a third element of the set by a first and a second. One of the most important structures stemming from one binary operation is that of a group structure. If a set G has a binary operation $\circ : G \times G \to G$ that satisfy closure and associativity the set is called a semi-group. By including an identity element one gets a monoid. These naturally appear in the discussion of effective field theories. By adding a fourth axiom, namely invertibility, the set is promoted to a group. In this case the binary operation is called a group multiplication. When the multiplication is commutative the group is called commutative or Abelian.

With more than one binary operation the two most important structures are those of rings and vector spaces. A ring is a set with two operations normally called addition and multiplication, both of which are binary maps on the same set. The trivial example of a ring is the ring of integers. A field is a ring where the two operations are commutative. A vector space is a set with operations $+ : \mathrm{V} \times \mathrm{V} \to \mathrm{V}$ and $\cdot : \mathrm{V} \times \mathrm{F} \to \mathrm{V}$ where F is a field. In this sense, a vector space is an Abelian group with respect to $+$, with the additional operation of "scalar multiplication" by elements of the field F. In principle one can add more and more structure to the sets. For example, a ring that is equipped also with a differentiation operation is called a differential ring.

These algebraic types of mathematical structure are well known to most physicists, and we will not spend much (if any) time discussing them except for the case by case appearances. The Non-algebraic structures, however, will play a much more central role in our discussion. Foundational for our geometric discussions are topological spaces. A topological space is a set X together with a collection of subsets $\mathrm{S} = \{\mathrm{S}_i\}$ such that unions and intersections of these are in S and X itself is in S. The subsets are called open sets. If S_i is a open set that contains $x \in \mathrm{X}$ and S_i is contained in a subset $\mathrm{V} \subset \mathrm{X}$ we call V a neighborhood of the point x. We will mostly be dealing with connected topological spaces. These are the topological spaces that cannot be seen as a union of disjoints sets. These notions will allow us to define manifolds, and from them the notion of a G-structure.

A.2 Categories and functors

Category theory is a framework for reasoning in mathematics, where many general ideas are captured. Roughly speaking, a category is some collection of objects and a collection of morphisms (also called "arrows") between them [7]. Maybe most important is the change of perspective category theory presents regarding mathematical structures. In a category the relation between objects is put on equal footing as the objects themselves, and discussing one without the other is in some sense an incomplete story. In fact, one can generalize this idea and study a category where the objects are the morphisms, and morphisms are morphisms of morphisms.

A category $\mathcal{C}$ is a collection of objects $a, b, c, \ldots$ in $\mathrm{Ob}(\mathcal{C})$ and a collection

$\mathrm{Hom}_{\mathcal{C}}(a,b)$ of morphisms

$$a \xrightarrow{f} b$$

between any two objects [8][7]. These morphisms satisfy a composition rule so that given two morphisms we can create a third

$$a \xrightarrow{f} b \ , \ b \xrightarrow{g} c$$

$$a \xrightarrow{g \circ f} c$$

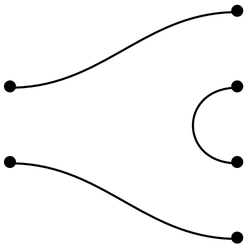
There is also a special morphism $1_a \in \mathrm{Hom}_{\mathcal{C}}(a,a)$ called the identity, which can be composed with any other morphism from the left or right without any effect. Often the objects in a category will be sets equipped with additional structure, like groups or vector spaces. The morphisms, which then for example may be maps between groups, are then often chosen to be maps that preserve the relevant structure. For example the category Set of sets have as objects sets and as morphisms functions between sets. In the category Grp of groups the objects are groups and the morphisms maps between groups compatible with the axioms of the group composition. These maps are called group homomorphisms. This is closely related to both group actions and group representations, which we discuss shortly.

One can also define maps between categories. Given categories $\mathcal{C}$ and $\mathcal{D}$ the map sends $\mathrm{Ob}(\mathcal{C})$ to $\mathrm{Ob}(\mathcal{D})$, and the morphisms in the one category to morphisms in the second in such a way that the composition rule is satisfied. These maps are called functors [7]. For example, one could consider a category where the objects are categories and morphisms are functors. This category is often denoted Cat. We discuss how such functors give rise to the idea of an action of a set with algebraic structure on the objects of some category in the appendix. In particular group actions and group representations naturally emerge from this discussion.

One can also generalize to the higher categories, where one also have morphisms between morphisms. If we call the morphisms in the above discussions 1-morphisms, the morphisms between 1-morphisms are called 2-morphisms. The general case where we have n-morphisms is called a n-category. For example, a 2-category can be drawn as

$$\begin{tikzcd} a \arrow[r, bend left=50, "f"{name=U}] \arrow[r, bend right=50, "g"'{name=D}] & b \arrow[Rightarrow, from=U, to=D, "\alpha"] \end{tikzcd}$$

Here the composition rules are somewhat more complicated, see for example [7]. From this perspective a set is simply a 0-category, where there is only objects. The category $\mathcal{C}at$ of categories is in fact a 2-category where so-called natural transformations are 2-morphisms. These are maps between functors $\alpha : \mathrm{F} \to \mathrm{G}$. If $f : c_1 \to c_2$ is a morphisms in $\mathcal{C}$, then a natural transformations associated to c_i a morphism $\alpha_{c_i} : \mathrm{F}(c_1) \to \mathrm{G}(c_i)$. The map α is a natural transformation if $\mathrm{G}(f) \circ \alpha_{c_1} = \alpha_{c_2} \circ \mathrm{F}(f)$. These 2-morphisms in $\mathcal{C}at$ in fact play a role in the representation theory of groups. Another interesting example is the category Mfd_n constricted as follows. The objects are 0-manifolds, e.g. an union of points. The 1-morphisms are 1-manifolds connecting the points, called cobordisms. The 2-morphisms are 2-manifolds connecting the 1-manifolds and so on.



The 2-morphisms would be surfaces interpolating between two of these one dimensional morphisms, i.e. a cobordism between cobordisms. This is similar to the (1-) category nCob where the objects are (n-1) dimensional manifolds, and the morphisms are n-manifolds.

A.3 Monoids, groups and actions

Groups are often imagined as an abstraction of the intuitive notion of a symmetry. Note however that the definition of a group does not include any reference to the way a group acts on certain objects. This intuitive picture of a group as some set of transformations often results from studying the morphisms that respect the group structure.

Recall from our discussion of category theory that a functor is a map $\mathrm{F} : \mathcal{C} \to \mathcal{D}$ between categories that respect the associativity of the category morphisms and the identity. Imagine a category $\mathcal{C}_\bullet$ with only one object $\bullet$. The morphisms are arrows that start and end on $\bullet$. To be a category these arrows need to be both associative and have an identity. If we recall the definition of a monoid, these structures exactly coincide. If the arrows of the category $\mathcal{C}_\bullet$ are invertible, we get a group. Is this way monoids and groups can be seen as one-object categories [36].

The notion of an action of a monoid or a group on some objects can be formally represented by a functor $\mathcal{A} : \mathcal{C}_{\bullet} \to \mathcal{D}$ where $\mathcal{D}$ contains the desired objects. Since we only have one object in our original category, the functor associates to $\bullet$ a single object $\mathcal{A}(\bullet)$ in the $\mathcal{D}$ category. The morphisms $f_i : \bullet \to \bullet$ get realized as maps $\mathcal{A}(f_i) : \mathcal{A}(\bullet) \to \mathcal{A}(\bullet)$ that are compatible with the properties of the original morphisms. This is abstractly the notion of a monoid action or group action.

Sometimes the name group (monoid) action is reserved to the case where the category $\mathcal{D}$ is the category Set. In this case the functor associates to the morphisms in the abstract one-object category elements of $\mathrm{A}ut(\mathrm{X})$, where $\mathrm{X} = \mathcal{A}(\bullet)$. More conventional notation then denotes a group (monoid) action as

$$\mathrm{G} \times \mathrm{X} \to \mathrm{X}$$

that takes one element of the group or monoid and lets it act on a point $x \in \mathrm{X}$, denoted $g(x)$ [35]. We will discuss what one can do with such group actions further when we discuss topological spaces.

When the category $\mathcal{D}$ is the category of vector spaces, the group action is called a group representation. First, note that many sets well known to physicists have a natural group structure. Most importantly, the set of $n \times n$ matrices $\mathrm{GL}_n(\mathbb{R})$ is a group under matrix multiplication. Almost all groups we will encounter will be subgroups of this group, i.e. a subset of $\mathrm{GL}_n(\mathbb{R})$ that by themselves satisfy the four group axioms. We can also consider the linear map on a finite dimensional vector space V, denoted GL(V). However, due to the remarkable classification of finite dimensional vector spaces[1] these coincide.

We want to discuss group actions on vector spaces. To do this we need the notion of a morphism of groups that can cary the group structure from the group over to a realization as linear maps on a vector space. A group homomorphisms is a map that preserves the group structure [74]. Given groups G and H with group multiplications $\circ_{\mathrm{G}}$ and $\circ_{\mathrm{H}}$ respectively, a group homomorphism is a map $f : \mathrm{G} \to \mathrm{H}$ such that

$$f(g_1 \circ_{\mathrm{G}} g_2) = f(g_1) \circ_{\mathrm{H}} f(g_2)$$

A representation of the group is then defined as follows. If GL(V) is the group of linear transformations on a vector space V, a group representation is a group homomorphism

$$\phi : \mathrm{G} \to \mathrm{GL}(\mathrm{V}).$$

The space V is then called the representation space [35]. The notion of equivalence of representations is defined as follows. If $f : \mathrm{V}_1 \to \mathrm{V}_2$ is a invertible map of

[1] Two vector spaces are isomorphic if and only if they have the same dimensions. Hence they are isomorphic to $\mathbb{R}^n$. A equally remarkable statement can be made for so-called separable Hilbert spaces - every one of them is isomorphic to the Hilbert space $\ell^2(\mathbb{N})$ of square summable sequences.

vector spaces, and ϕ_1 and ϕ_2 are their respective representations of a group G. Then if

$$f \circ \phi_1(g) = \phi_2(g) \circ f$$

the two representations are called isomorphic [35]. Representation theory aims to find all non-isomorphic representations. In the representation space V of a representation has no invariant subspaces W ⊂ V, the representation is called irreducible [35].

A.4 Classification and moduli spaces

The main lesson from category theory is that the objects of a category and its morphisms are equally important, and both are needed to define a category. In this sense we should not discuss one without the other. This naturally motivates may interesting classification questions, where equivalence classes are searched for in terms of some map compatible with a given structure. Classifications naturally introduce the concept of a moduli space. The name "moduli" comes from B. Riemanns study of surfaces with complex structure, where the complex structures were classified with a parameter, or modulus. We will use the words moduli and parameter more or less synonymously in this thesis.

A classification problem often comes in two steps. First, one tries to find discrete invariants, often of the topological type. This divides the objects one is studying into disjoint cases one can study in more detail on their own. The second step is to see if the objects in one of these discrete classes are parametrized by some remaining continuous parameter, or *moduli*. It may also be that this moduli parametrizes not just a single object, but rather a whole isomorphism class of these objects. The space of these moduli is what we will refer to as a moduli space. In mathematical literature, a moduli space sometimes has stricter definitions, but these will not be important for our purposes. See for example [42].

So a moduli space is a space $\mathcal{M}$ where each point corresponds to a class of the objects we would like to study. We can imagine a sort of family $\mathcal{U}$ of objects as a distribution over the moduli space, such that every class appears once. Formally we have a continuous map

$$\pi : \mathcal{U} \to \mathcal{M}$$

such that $\pi^{-1}(p)$ is the equivalence class associated with $p \in \mathcal{M}$. The fact that this map is continuous intuitively means that two nearby points in the moduli space represents "very similar" objects. As a trivial example consider the classification of circles. A circle is specified by a single number, namely the radius. In this case the moduli space can be taken to be $\mathbb{R}^+$ where $\pi^{-1}(r) = \mathbb{S}^1_r$ is a circle of radius r.

Circles with approximately the same radius are "roughly identical".

B

Topology

Topology is one of many non-algebraic structures we will meet in this thesis. In particular we will discuss the morphisms that preserve topological properties of a space. Intuitively these are continuous deformations of the space.

Let X_1 and X_2 be topological spaces, and $f : X_1 \to X_2$ a map between them. The spaces are said to be homeomorphic if there exists such a map f which is continuous and has a continuous inverse f^{-1} [45]. This puts an equivalence relation on the spaces, where $X_1 \sim X_2$ if they can be deformed into each other by a homeomorphism. We say that the spaces fall into the same equivalence class, or more precisely homeomorphic equivalence class. We will discuss equivalence relations in more detail shortly.

As an example, a torus can be continuously deformed into a sphere with a handle, with no cutting or ripping. However, the homeomorphisms are not always intuitively clear and we need a somewhat more sophisticated tool for classifying spaces topologically. In essence we will do something very familiar to physicists - we use "conserved quantities" to simplify the problem. We would like to find some quantities that are preserved under homeomorphisms, i.e. topological invariants. One important note is the following: If all topological invariants we are aware of coincide for two spaces, they are not necessarily homeomorphic. This is because we can not claim to know of all possible topological invariants, and we will surely not cover more than a few in our discussions. What we *can* say is that if two spaces have different topological invariants they are for sure *not* homeomorphic.

B.1 From equivalence relations to homotopy groups

Homotopy groups will be one of our main tools for studying topological properties of a space. The first homotopy group is a group associated with a space X such that homeomorphic spaces have homotopy groups that are isomorphic. In this way, we can use the group to say something about the topological structure of X.

Before moving on to homotopy groups, let us consider quotient spaces. If X is a topological space and $\pi : X \rightarrow Y$ is a map from X to a space Y, one can define subsets $U \subset Y$ to be open if $\pi^{-1}(U)$ is open. If π is surjective and continuous it is called a quotient map [45] .

An equivalence relation on X is a sort of identification $\sim$ of elements in X. An equivalence relation has to satisfy the following relations [45]

$$x \sim x \forall x \in X$$

$$x \sim y \rightarrow y \sim x$$

$$x \sim y \,,\; y \sim z \,,\, \rightarrow x \sim z$$

The equivalence class $[x]$ is defined to be

$$[x] = \{\forall y \in X | y \sim x\}$$

The space of all such classes are denoted $X/\sim$, as to indicate that one "divides out" the equivalence. The projection map $\pi : X \rightarrow X/\sim$ that sends

$$\pi : x \rightarrow [x]$$

is a quotient map [45] and the resulting space is called a quotient space. This is a very neat way of constructing new spaces from old. Intuitively, the equivalence relation identifies point in X and the projection map cuts and glues the space accordingly, resulting in the new space $X/\sim$. For example, if we take the real line and identify all integer spaced points, the quotient space would be a circle, at least topologically.

The idea of homotopy groups is to probe the topology of a space by the deformation of loops in that space. These loops feel topological obstructions as we deform them, and can therefore be used to say something about the topology of a space. We here make these notions more precise. The following definitions and results are taken from [45].

We let X be a topological space. We assume the space to be path-connected in the sense that any two points may be connected by a continuous curve. By a loop ℓ we mean a map

$$\ell : [0,1] \to \mathrm{X}$$

where $\ell(0) = \ell(1) = x_0 \in \mathrm{X}$. Since the space is assumed path connected the point x_0 plays no special role, and may be moved around arbitrarily on X. Given two loops we may glue them together by what is called concatenation defined by

$$\ell_1 * \ell_2 = \begin{cases} \ell_1(2t) & \text{if } t \in [0,1/2] \\ \ell_2(2t-1) & \text{if } t \in [1/2,1] \end{cases}$$

This simply means we traverse the second loop after the other, moving at double parameter speed. We have here assumed that the two loops both start and end at the same (but arbitrary) point x_0. Two loops are said to be homotopic if there exist a continuous map $\mathrm{L} : [0,1] \times [0,1] \to \mathrm{X}$, which we denote $\mathrm{L}(s,t) \equiv \mathrm{L}_s(t)$, where $\mathrm{L}_0(t) = \ell_1(t)$ and $\mathrm{L}_1(t) = \ell_2(t)$. In other words $\mathrm{L}_s(t)$ describes a continuous family of loops in X parametrized by $s \in [0,1]$ that start at one loop ℓ_1 and end at the other ℓ_2.

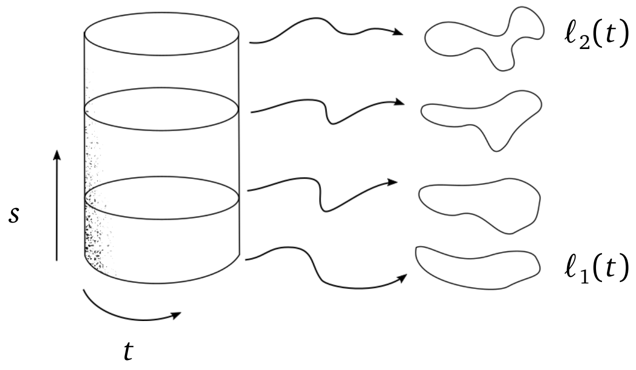


Figure B.1: The continuous family of loops can be seen as a deformation of the first loop into the second. For any fixed s there is a loop homotopic to any other.

This introduces a equivalence relation on loops, i.e. $\ell_1 \sim \ell_2$ if the two are homotopic[1]. More generally, two maps are said to be homotopic if they can be deformed into each other. A loop that is homotopic to the trivial loop, i.e the point x_0, is called null homotopic, meaning that it may be continuously deformed to a point. With this notion of equivalence, we can construct equivalence classes

$$[\gamma] = \{\ell : [0,1] \to \mathrm{X} | \ell \sim \gamma\}$$

[1] One can easily see that this relation satisfies the three requirements to be a equivalence relation.

of all loops that are homotopic to some particular loop γ. One can further extend the concept of concatenation to equivalence classes by the following observation. Assume we have two equivalence classes of loops $[\ell]$ and $[\gamma]$, and two representatives of each of them, i.e.

$$\ell_1, \ell_2 \sim \ell$$

$$\gamma_1, \gamma_2 \sim \gamma$$

Then $\ell_1 * \gamma_1 \sim \ell_2 * \gamma_2$, since the two loops may be deformed individually. This means we have a well defined operation $*$ on the whole equivalence classes $[\ell]$ and $[\gamma]$. By noting that the concatenation with the trivial loop is an identity operation, and that an inverse may be constructed by

$$[\gamma(t)]^{-1} = [\gamma(1-t)]$$

the set of equivalence classes of loops gains a group structure. This is called the first homotopy group of X and it denoted

$$(\{[\gamma]\}, *) \equiv \pi_1(\mathrm{X})$$

To identify this group one has to find the topologically distinct loops in X, i.e. study the deformation of circles on X. For example, consider the case of $\mathrm{X} = \mathbb{R}^3 \backslash \{(0,0,z)\}$ i.e. Euclidian 3-space with a line removed. In this space there is an infinite number of topologically non-equivalent classes of loops, namely those that wind around the removed line an positive/negative integer number of times. Concatenating the class of loops that wind one time with the class that wind two times clearly yields the class of loops that wind three times, and we see that the first homotopy group behaves as the integers under addition:

$$\pi_1(\mathbb{R}^3 \backslash \{(0,0,z)\}) = (\mathbb{Z}, +)$$

By equality we here really mean group isomorphism.

These groups of loops are not only interesting by themselves but are useful when classifying and constructing spaces. For example, the statement that the first homotopy group is the trivial group is simply the statement that the space X has no holes or other topological obstructions. In this case we call X simply connected.

Now, consider two topological spaces X_1 and X_2 and an onto map $\pi : \mathrm{X}_2 \to \mathrm{X}_1$ between them. Let also V be a neighborhood of x in X_1. Then, if $\pi^{-1}(\mathrm{V})$ is a disjoint union of subsets in X_2, i.e.

$$\pi^{-1}(\mathrm{V}) = \bigsqcup_i u_i \,, \;\; u_i \subset \mathrm{X}_2$$

and the restricted map $\pi|_{u_i}$ is a homeomorphism of u_i and V then π is called a covering map and (X_2, π) is called a cover of X_1. In the case where $\pi_1(\mathrm{X}_2) = \{e\}$ we call X_2 a universal cover and denote it $\mathrm{X}_2 \equiv \tilde{\mathrm{X}}$.

We will often find ourselves studying curves on a space X_1 that is not necessarily simply connected. In this case it is a good idea to lift the curves to a covering. If $\pi : \mathrm{X}_2 \to \mathrm{X}_1$ is a covering map, and $f : \mathrm{Y} \to \mathrm{X}_1$ is a continuous map, a lift is defined as any map $\tilde{f} : \mathrm{Y} \to \mathrm{X}_2$ such that $\pi \circ \tilde{f} = f$ [45]. For lifting curves the map is a parametrization of a curve $f : \mathbb{R} \to \mathrm{X}_1$. The following results will be useful [45].

Path Lifting Property: *Let $f : (0,1) \to \mathrm{X}_1$ be a path starting at $f(0)$. For $\pi : \mathrm{X}_2 \to \mathrm{X}_1$ a covering map and p a point in X_2 such that $\pi(p) = f(0)$ there exists an unique lift $\tilde{f}$ such that $\tilde{f}(0) = p$.*

Monodromy Theorem: *Let $f, g : (0,1) \to \mathrm{X}_1$ be paths starting and ending at the same points. Let also $\pi : \mathrm{X}_2 \to \mathrm{X}_1$ be a covering map, and $\tilde{f}$ and $\tilde{g}$ be the lifted curves both starting at some $p \in \mathrm{X}_2$. Then $\tilde{f} \sim \tilde{g}$ if and only if $f \sim g$ and if $f \sim g$ then $\tilde{f}(1) = \tilde{g}(1)$.*

Monodromy Acion: *Let $\pi : \mathrm{X}_2 \to \mathrm{X}_1$ be a covering map, and $x \in \mathrm{X}_1$ some point in the base space. Then there is a group action of $\pi_1(\mathrm{X}_1)$ on the points $\pi^{-1}(x) = \{p_i | \pi(p_j) = x\}$ called the monodromy action, acting by $[\gamma] : p \to \tilde{\gamma}(1)$.*

For a much more detailed discussion of lifts of coves and the monodromy action see [45]. In summary these results means the following. If we are given a collection of curves at a point $x \in \mathrm{X}_1$ (not necessarily homotopic) we can lift these curves to a covering space, where all the curves will start at the same point p, but end somewhere else in $\pi^{-1}(x)$. These endpoints are obtained by the action of the first homotopy group. These results is in fact what allows us to define quantum mechanics on topologically non-trivial spaces - the inequivalent ways to quantize a system is in bijective correspondence with the one dimensional representations of the first homotopy group.

B.2 Classification of surfaces

In the case of two dimensions, the most important topological invariant is the Euler Characteristic. To understand this invariant we need the notion of a triangulation. A polyhedra is a surface that has flat polygonal faces, straight lines as edges and corners called vertices. The process of creating this polyhedra out of a smooth surface is called triangulation. A typical examples is the tetrahedron, which may be seen as triangulations of the sphere. The Euler characteristics for a surface X homeomorphic to a polyhedra P is defined by

$$\chi(\mathrm{X}) \equiv \mathrm{V}(\mathrm{P}) - \mathrm{E}(\mathrm{P}) + \mathrm{F}(\mathrm{P}) \tag{B.1}$$

where $\mathrm{V}(\mathrm{P}), \mathrm{E}(\mathrm{P}), \mathrm{F}(\mathrm{P})$ are the number of vertices, edges and faces of P respectively. A famous theorem due to Poincaré and Alexander [60] tells us that all polyhedra homeomorphic to the surface X yields the same Euler characteristics $\chi(\mathrm{X})$. For example, the 2-sphere is homeomorphic to a cube in 3-space, which has 8 vertices, 12 edges and 6 faces. Hence the sphere has $\chi(\mathrm{S}^2) = 8 - 12 + 6 = 2$. The sphere is also homeomorphic to a tetrahedron (triangular pyramid) which has 4 vertices, 6 edges and 4 faces. Similarly then $\chi(\mathrm{S}^2) = 4 - 6 + 4 = 2$. A torus is somewhat more tricky. The torus is obtained by identifying points in the plane related by translation in the x and y directions, i.e. "gluing" together the sides of a rectangle. We can thus draw edges and vertices on this rectangle to make the counting relatively simple.

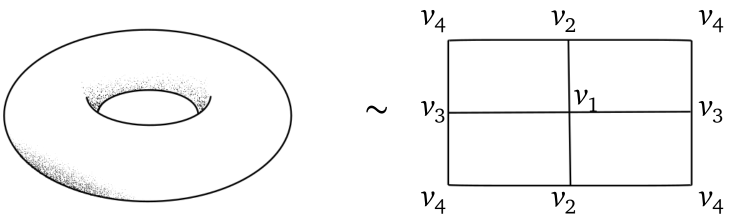


Figure B.2

We have to remember that edges are identified when we do the counting and not count vertices and edges twice. After some counting we see that $\chi(\mathbb{T}^2) = 4 - 8 + 4 = 0$. Thus the sphere and the torus are not topologically equivalent. This is of course not a surprise as the torus has a hole in it while the sphere does not, and thus can not be continuously deformed into each other without ripping and tearing. The intuitive idea of "how many holes" a surface has may be somewhat formalized. Intuitively we may think of a surface with g holes as a topologically trivial surface (the sphere) with g tori on it, or rather g handles. The number g is called the genus of the surface. This introduces the idea of a connected sum, writen $\mathrm{X}_1\sharp\mathrm{X}_2$, which is defined by cutting a hole in two surfaces and connecting them by a cylinder [60]. For example the surface $\mathbb{T}^2\sharp\mathbb{T}^2$ is given geometrically by

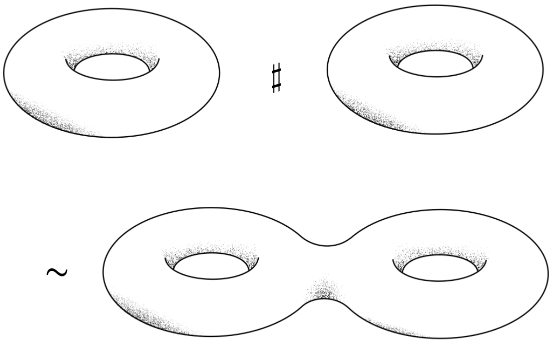


Figure B.3

A surface that can not be expressed as a connected sum of two surfaces is called prime.

Note that by triangulating the surface resulting from a connected sum, the Euler characteristics changes not only to a sum of the characteristics for the factor surfaces [60]. Imagine removing from each triangulated surface a triangle and connecting the two with a triangular cylinder. Then the number of faces decrease by -1 on each surface, while the total surface has gained three edges and three faces from the triangular cylinder. Inserting this into the formula for Euler characteristics we find that

$$\chi(X_1 \sharp X_2) = \chi(X_1) + \chi(X_2) - 2$$

Consider now the surface $\Sigma_g \equiv \mathbb{T}^2 \sharp ... \sharp \mathbb{T}^2$ of g factors of the torus, i.e. (topologically) an arbitrary genus g surface. Since the torus has vanishing Euler characteristics we have

$$\begin{aligned}
\chi(\Sigma_2) &= -2 = 2 - 2 \cdot 2 \\
\chi(\Sigma_3) &= -4 = 2 - 2 \cdot 3 \\
\chi(\Sigma_4) &= -6 = 2 - 2 \cdot 4 \\
&\vdots \\
\chi(\Sigma_g) &= 2 - 2g
\end{aligned}$$

What should be intuitively clear is that any nicely behaved[2] surface is homeomorphic to either the 2-sphere or Σ_g. In other words, For such surfaces there are two relevant prime surfaces: the sphere and the torus. This result implies that the classification of surfaces by homeomorphisms is given uniquely from the genus g of the surface. More abstractly stated, the topological isomorphisms classes of compact surfaces corresponds to the set of positive integers.

[2] Compact, connected, orientable, Hausdorf and so on.

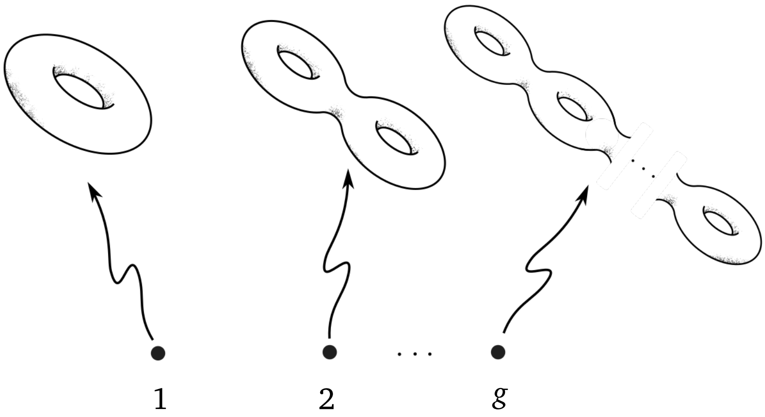


Figure B.4: Classification of real orientable surfaces by their genus. Each real integer g corresponds to a class of homeomorphic surfaces.

C

Lie groups and their algebras

Several places in this thesis we make use of Lie groups and algebras. We will not present a detailed discussion of the subject, and refer the reader to for example [35] for applications in quantum theory or [34] for a general discussion of Lie theory and representations.

A Lie group is a smooth n-manifold G with a group structure [74]. The group multiplication

$$\circ : \mathrm{G} \rightarrow \mathrm{G}$$

has to respect the smooth structure of the manifold, and hence are smooth maps from the manifold to itself. The left action of a group on itself is defined by $h \rightarrow \mathrm{L}_g h = gh$. Similarly, one could define a right action R_g, but we will not need this here. We know that a map between manifolds induce a map on the tangent spaces. In this case, the left action induces a map $\mathrm{L}_g^* : \mathrm{T}_h\mathrm{G} \rightarrow \mathrm{T}_{gh}\mathrm{G}$ that relates tangent spaces over G. Imagine a vector field V over G. If the vector field at gh can be obtained from the vector field at g by L_g^*, the vector field is called left-invariant [74]. In some sense, the vector field is generated by a tangent space and the group structure.

In particular, let X be a vector at $\mathrm{T}_e\mathrm{G}$, the tangent space at the identity element. We can then construct a vector field V_X that at a point g obeys

$$(\mathrm{V}_\mathrm{X})_g = \mathrm{L}_g^* \mathrm{X}$$

Recall that the Lie bracket $[\cdot, \cdot]$ mapped vector fields to vector fields, so we also have $[\mathrm{X}, \mathrm{Y}]_{gh} = \mathrm{L}_g^*[\mathrm{X}, \mathrm{Y}]_h$. By collecting all left-invariant fields and equipping this vector space with the binary map that is the Lie bracket, we get an algebra of vector fields. This is called the Lie algebra of G, and denoted $\mathfrak{g}$. Since these fields can be obtained by the pushforward L_g^* from the identity, we associate the Lie algebra with the tangent space $\mathrm{T}_e\mathrm{G}$ at the identity equipped with the Lie bracket [74]. If we consider $\{\mathrm{X}_a\}$ as a basis for this tangent space, we can express $[\mathrm{X}_a, \mathrm{X}_b]$

as a linear combination

$$[X_a, X_b] = f_{ab}^{c} X_c$$

The coefficients f_{ab}^{c} are called structure constants. They specify the Lie algebra, and hence the Lie group for infinitesimal transformations.

To relate the Lie group and the Lie algebra we need the exponential map. First, note that a curve $\gamma : \mathbb{R} \to G$ with the property $\gamma(t) \circ \gamma(t') = \gamma(t + t')$ can be considered a Abelian subgroup of G. Note that $\gamma(0) = e$. These curves are called one-parameter subgroups [60][74]. The tangent vectors to this curve are left-invariant vectors, and we can think of the one-parameter subgroup as being generated by the tangent vector at e [74][35]. Let $X \in T_e G$ and consider a curve γ_X generated by this vector by the pushforward L_g^*. Then the exponential map $\exp : T_e G \to G$ is defined by [35]

$$\exp(tX) = \gamma_X(t)$$

When we are dealing with matrix groups, which is particularly relevant for finite dimensional representations, the exponential can be defined as a Taylor series

$$e^X = \sum_n \frac{1}{n!} X^n$$

This map satisfies [35]

$$e^0 = 1$$
$$\det(e^X) = e^t r(X)$$
$$e^{X+Y} = e^X e^Y \text{for commuting fields.}$$
$$(e^X)^{-1} = e^{-X}$$

By differentiating term by term in the series we see that $\gamma_X{}'(t)|_0 = X$. Often this is how one identified the generators of the group.

Recall that a group homomorphism is a map from preserves the group structure. If $\phi : G_1 \to G_2$ is a homomorphism, there will be a map $\psi : \mathfrak{g}_1 \to \mathfrak{g}_2$ between Lie algebras such that $\phi(e^X) = e^{\psi(X)}$ [35]. This will be particularly important for representations of G. If $\rho : G \to GL(V)$ is a representation on V, there is a representation [35] $\pi : \mathfrak{g} \to gl(V)$ such that

$$\rho(e^X) = e^{\pi(X)}$$

Particularly interesting in quantum mechanics are the unitary representations, where the group homomorphism is a map $\rho : G \to U(\mathcal{H})$, to the set of unitary operators on a (finite dimensional) Hilbert space $\mathcal{H}$. Recall that unitary operators are those that satisfy $\langle Av, Aw \rangle = \langle v, w \rangle$.

D

Some aspects of quantum mechanics

D.1 Operator algebraic quantization

To construct the relevant Hilbert space it is fruitful to study the algebra of observables in the theory. First recall some facts from classical mechanics. The tangent vectors to the curve representing the particle trajectory trough M describe the velocity of the particle. If we imagine the set of all paths from $x(t_1)$ to $x(t_2)$, the physical path is the one satisfying the variational principle. Given a Lagrangian $\mathrm{L} : \mathrm{TM} \rightarrow \mathbb{R}$ as a function of coordinates and velocities, the variational principle states that

$$\delta \mathrm{S} = \delta \int dt \mathrm{L} = 0$$

The Lagrangian is normally obtained by $\mathrm{L} = \mathrm{T} - \mathrm{V}$ where T is the kinetic energy and V a potential on M [60]. The canonical momentum conjugate to the coordinate x^μ is defined to be

$$p_\mu = \frac{\partial \mathrm{L}}{\partial \dot{x}^\mu}$$

Note that under a coordinate transformation, the canonical momentum transformers a 1-form on M. By performing a Legendre transform of L where we replace $\dot{x}^\mu$ with the canonical momentum, we get the Hamiltonian formulation of classical mechanics that lives on the cotangent bundle $\mathrm{T}^*\mathrm{M}$. Here the greek indices run over the spatial dimensions of M. The Hamiltonian is given by

$$\mathrm{H}(x,p) = p_\mu \dot{x}^\mu - \mathrm{L}(x, \dot{x}(p))$$

Note that a trivialization of the cotangent bundle is essentially $\mathbb{R}^{2n}$ consisting of 2n-tuples $(x^1, ..., x^n; p_1, ..., p_n)$. The cotangent space is in the applications of classical mechanics called the phase space of the system denoted $\mathcal{P}$. The state of the system is a point in phase space, and time evolution is a flow in $\mathcal{P}$. The equations of motion, obtained by the variational principle, reads [60]

$$\dot{x}^\mu = \frac{\partial \mathrm{H}}{\partial p_\mu}$$

$$\dot{p}_\mu = -\frac{\partial \mathrm{H}}{\partial x^\mu}$$

These can be rewritten by introducing an algebra of observables. Consider smooth functions C^∞ on $\mathcal{P}$ and define the Poisson brackets by

$$\{f, g\} = \frac{\partial f}{\partial x^\mu}\frac{\partial g}{\partial p_\mu} - \frac{\partial f}{\partial p_\mu}\frac{\partial g}{\partial x^\mu}$$

for $f, g \in \mathrm{C}^\infty$. The tuple $(\mathrm{C}^\infty(\mathcal{P}), \{\cdot,\cdot\})$ makes up the algebra of classical observables $\mathscr{A}_\mathrm{C}$. Note that for the coordinates and momenta themselves we have $\{x^\mu, p_\nu\} = \delta^\mu_\nu$. Now let F be an observable. The time evolution of this quantity is given by the equation

$$\dot{\mathrm{F}} = \frac{\partial \mathrm{F}}{\partial x^\mu}\dot{x^\mu} + \frac{\partial \mathrm{F}}{\partial p_\mu}\dot{p_\mu} = \{\mathrm{F}, \mathrm{H}\}$$

by using the equations of motion. Hence there is a close relation between the algebra of observables and the time evolution of an element of the algebra.

Canonical quantization is the procedure of mapping the classical algebra to a quantum version of it

$$\mathscr{Q} : \mathscr{A}_\mathrm{C} \longrightarrow \mathscr{A}_\mathrm{Q}$$

where, by postulate, the quantum observables are Hermitian operators acting on a Hilbert space $\mathcal{H}$ of states [60]. The standard rules of quantization are

$$\mathscr{Q} : \{\cdot,\cdot\} \to [\cdot,\cdot] = i\hbar\mathscr{Q}(\{\cdot,\cdot\})$$

where $\hbar$ is Plancks constant. The Hilbert space is then constructed as the eigenspace of the observables. In particular, we are interested in the maximal set of commuting observables, as these operators can be simultaneously diagonalized. For the canonical coordinates and momenta, the quantizatized operators satisfies $[x^\mu, p_\nu] = i\hbar\delta^\mu_\nu$. The typical solution is to take x^μ simply as multiplication by x^μ, and $p_\mu = i\hbar\partial_\mu$. Other observables $\mathcal{O} = \mathcal{O}(x, p)$ can then be obtained from the operator form of the canonical variables. The Hilbert space is in this case imagined by be consisting of the delta-like position states $|x\rangle$ with general state $|\psi\rangle = \int d\mathrm{vol}_\mathrm{M}\psi(x)|x\rangle$. The state is determined by the wavefunction $\psi(x) = \langle x|\psi\rangle$ which lives in the Hilbert space $\mathrm{L}^2(\mathrm{M})$.

A popular method of constructing a Hilbert space of states is by the so called ladder operator method. This way of quantizing is sometimes also called the algebraic approach, in the sense that it relies heavily on Lie algebra theory. We try to present a clear and precise description of this method.

Formally, quantum mechanics deals with the construction and diagonalization of a Cartan subalgebra of the algebra of observables. The Cartan subalgebras are the maximal collection of commuting operators, with Hermitian generators $\{\mathrm{H}_i\}$, not to be confused with the Hamiltonian. Here i runs from 1 to m, called the rank of the subalgebra. Imagine that we have diagonalized the H_i. Given a representation on a Hilbert space, we label the states $|\lambda\rangle$ where

$$\mathrm{H}_i\,|\lambda\rangle = \lambda_i\,|\lambda\rangle$$

Now assume that there are operators $\mathrm{A}_{\pm\alpha}$ with $\mathrm{A}_{-\alpha} = \mathrm{A}^{\dagger}_{+\alpha}$ satisfying the commutation relations

$$[\mathrm{H}_i, \mathrm{A}_{\pm\alpha}] = \mp\alpha_i \mathrm{A}_{\pm\alpha}$$

for each i. Then one can trivially show that the state $\mathrm{A}_{\pm\alpha}\,|\lambda\rangle$ has eigenvalues

$$\mathrm{H}_i\mathrm{A}_{\pm\alpha}\,|\lambda\rangle = (\lambda \mp \alpha_i)\mathrm{A}_{\pm\alpha}\,|\lambda\rangle$$

In this way the operators $\mathrm{A}_{\pm\alpha}$ raises and lowers the eigenvalues and are therefore called ladder operators. The standard way to proceed is then the following. After picking the Cartan generators, one should try to construct operators of the type $\mathrm{A}_{\pm\alpha}$ so that the above algebraic relations hold. Then, by finding a lowest eigenstate of one of the Cartan generators one can construct a larger Hilbert space of states by the action of $a_{+\alpha}$.

The standard example of this method is the harmonic oscillator, with Hamiltonian $\mathrm{H} = \frac{1}{2}(p^2 + x^2)$. Classically $\{x, p\} = 1$, so by canonical quantization $[x, p] = i$. We have set all physical parameters to 1 in this discussion for the sake of clarity. Recall that x and p generate shifts is momentum and position respectively. The Hamiltonian does not commute with these transformations, and we can only diagonaliz one operator. We naturally pick H which plays the role of our single Cartan generator. By defining operators

$$a = \frac{x + ip}{\sqrt{2}}\ ;\ a^{\dagger} = \frac{x - ip}{\sqrt{2}}$$

The Hamiltonian can be written $\mathrm{H} = a^{\dagger}a + 1/2$. The operators $a, a^{\dagger}$ satisfies

$$[a^{\dagger}, a] = 1\ ;\ [\mathrm{H}, a^{\dagger}] = [a^{\dagger}a, a^{\dagger}] = a^{\dagger}\ ;\ [\mathrm{H}, a] = [a^{\dagger}a, a] = -a$$

and will be the ladder operators. Hence, if we have a eigenvalue $a^{\dagger}a\,|\lambda\rangle = \lambda\,|\lambda\rangle$ the state $a\,|\lambda\rangle$ has eigenvalue $\lambda - 1$. In fact, the spectrum of $a^{\dagger}a$ is $\mathbb{N}$, and $a^{\dagger}a$ is often called the number operator N. To see that the spectrum consists of integers, assume that λ is non-integer and that n is the closest integer above λ. Then the state $a^n\,|\lambda\rangle$ would have eigenvalue $\lambda - n$ which is negative. However, since

$$\lambda = \langle\lambda|\mathrm{N}|\lambda\rangle = (a\,|\lambda\rangle)^{\dagger}a\,|\lambda\rangle \geq 0$$

this is a contradiction and λ must be a positive integer. Hence the Hamiltonian has the spectrum $n + 1/2$ for $n \in \mathbb{N}$. From an algebraic perspective this situation is almost identical to that of a particle in a magnetic field.

D.2 Magnetic fields and translation symmetry

When we discussed topological insulators and the Hall effect there is a small problem that we overlooked. The Hamiltonian of a particle moving in a magnetic field is not translationally invariant, meaning Blochs theorem does not apply. Physically however it should be clear that a homogenous magnetic field is translation invariant, but the electromagnetic gauge field is not. However, by introducing so called magnetic translations the problem can be overcome.

We consider the quantum dynamics of a particle in a magnetic field. The particle motion can classically be seen as a curve $x : [t_0, t_1] \to \mathrm{M}$ trough a space M. We will consider simply 2+1 dimensional Minkowski space in this section. The classical action of the particle moving freely is

$$\mathrm{S} = \int dt \frac{1}{2} m \dot{x}_\mu \dot{x}^\mu$$

Given a U(1) gauge field $\mathrm{A}_\mu dx^\mu$ over M we add the current-gauge field term

$$\mathrm{S} = \int dt \frac{1}{2} m \dot{x}_\mu \dot{x}^\mu - q \int dt \dot{x}^\mu \mathrm{A}_\mu$$

We write $\mathrm{A}_0 = \phi$. For an electron with $q = -e$ the action takes the form

$$\mathrm{S} = \int dt \frac{1}{2} m \dot{x}_\mu \dot{x}^\mu + e \int dt (\phi - x^i \mathrm{A}_i)$$

where latin indices run over spatial dimensions. The magnetic and electric fields are as usual defined by

$$\mathrm{B}^i = \epsilon^{ijk} \partial_j \mathrm{A}_k$$
$$\mathrm{E}^i = \partial_t \mathrm{A}_i - \partial_i \phi$$

For a magnetic field only the Lagrangian reduces to the form

$$\mathrm{L} = \frac{1}{2} m \dot{x}_\mu \dot{x}^\mu u - e \dot{x}^i \mathrm{A}_i$$

In the Hamiltonian framework with canonical momentum $p_i = m\dot{x}_i - e\mathrm{A}_i$ the Hamiltonian reads

$$\mathrm{H} = \frac{1}{2m}(p_i + e\mathrm{A}_i)(p^i + e\mathrm{A}^i)$$

The mechanical momentum is $\pi_i = p_i + e\mathrm{A}_i = m\dot{x}_i$. To quantize the system we need the classical Poisson brackets. We will quantize the mechanical momentum as this is gauge independent. The Poisson brackets read

$$\{\pi_i, \pi_j\} = e(\partial_j \mathrm{A}_i - \partial_i \mathrm{A}_j)$$

Using $\epsilon_{\mu\sigma\lambda}\epsilon^{\mu\nu\rho} = \delta^{\nu}_{\sigma}\delta^{\rho}_{\lambda} - \delta^{\rho}_{\lambda}\delta^{\nu}_{\lambda}$ we can write this as

$$\{\pi_i, \pi_j\} = -e\epsilon_{ijk}\mathrm{B}^k$$

For a electron moving in two spatial dimensions with a perpendicular magnetic field these relations reduce to

$$\{\pi_x, \pi_y\} = -e\mathrm{B}$$

where B is the magnetic field strength. Canonical quantization then promotes these phase space functions to Hermitian operators on a Hilbert space by

$$\mathscr{Q} : \{\pi_x, \pi_y\} \to [\pi_x, \pi_y] = -ie\hbar\mathrm{B}\mathbb{I}$$

By introducing the complexified operators

$$a = (2\pi e\hbar\mathrm{B})^{-1/2}(\pi_x - i\pi_y)$$

$$a^{\dagger} = (2\pi e\hbar\mathrm{B})^{-1/2}(\pi_x + i\pi_y)$$

which satisfies $[a, a^{\dagger}] = 1$ we can write the single particle Hamiltonian

$$\mathrm{H} = \hbar\omega_{\mathrm{B}}(a^{\dagger}a + 1/2)$$

Hence this problem has the same algebraic structure as that of the Harmonic oscillator, with a Hilbert space constructed similarly. The energy levels $\mathrm{E}_n = \hbar\omega_{\mathrm{B}}(n + 1/2)$ are called Landau levels. For more on particles in magnetic fields in connection with the Hall effect see [80], from which parts of this discussion is borrowed.

More relevant for us is the observation that the magnetic Hamiltonian is not translational invariant. This is an obstacle if we want to discuss the integer Hall effect as a topological insulator, since these depended on Blochs theorem. This obstacle can be overcome by introducing magnetic translations. Recall that the magnetic field is $\mathrm{B}^{\mu} = \epsilon^{\mu\nu\rho}\partial_{\nu}\mathrm{A}_{\rho}(x)$. If we translate by R this becomes

$$\epsilon^{\mu\nu\rho}\frac{\partial}{\partial(x + \mathrm{R})^{\mu}}\mathrm{A}_{\rho}(x + \mathrm{R}) = \epsilon^{\mu\nu\rho}\partial_{\nu}\mathrm{A}_{\rho}(x + \mathrm{R})$$

Since the magnetic field is homogenous this must equal $\epsilon^{\mu\nu\rho}\partial_{\nu}\mathrm{A}_{\rho}(x)$. Hence the freedom we have is to pick $\mathrm{A}_{\rho}(x+\mathrm{R}) = \mathrm{A}_{\rho}(x)+\partial_{\rho}\omega(x)$ for some function ω since the contraction of a symmetric tensor with a antisymmetric one vanishes. This is of course nothing but a U(1) gauge transformation. The combined transformation

$$\mathscr{T}_{\mathrm{R}} = e^{ie\omega(x)}\mathrm{T}_{\mathrm{R}} = e^{ip_{\mu}\mathrm{R}^{\mu}+ie\omega(x)}$$

is called a magnetic translation and is the combination of a ordinary translation and a gauge transformation. These are indeed symmetries of the system and commutes with the Hamiltonian if we have periodic boundary conditions.

We pick the symmetric gauge where $A_\mu = \frac{1}{2}\epsilon^{\mu\nu\rho}B_\nu x_\rho$ [80]. Performing a translation on this gauge field by R we can easily read of

$$\omega = \frac{1}{2}\epsilon_{\mu\nu\rho}B^\nu R^\rho x^\mu$$

We consider translations by $a = (n_x L_x, n_y L_y)$, $b = (m_x L_x, m_y L_y)$ where we imagine the L_i's to be the lengths of the Hall sample and the n, m are integers. After some algebra one can show that

$$\mathcal{T}_a\mathcal{T}_b = \exp[ie(n_y m_x - n_x m_y)BL_x L_y/\hbar]\mathcal{T}_b\mathcal{T}_a$$

In this sense, we can view the magnetic translations as a projective version of the ordinary translations. We consider $n_y = m_x = 1$, $n_x = m_y = 0$ and write this extra phase as

$$e^{i\phi/\phi_0}$$

where $\phi = BL_x L_y$ and $\phi_0 = \hbar/e$. Hence for discrete values of the magnetic field where $\phi = m\phi_0$ the translations reduce to the usual commutative translations, and Blochs theorem applies. In this case we have sort of cheated by copying the Hall system over and over, making it translationally invariant. Thus the results from the discussion of bundles and Chern numbers can again be applied, resulting in a quantized Hall conductance σ_H [18].

D.3 Inequivalent quantizations and homotopy theory

In classical mechanics the state of a particle is given by a fixed point in phase space, corresponding to its position in configuration space as well as its canonical momentum. The fact that a state corresponds to a single point in configuration space is no longer true in quantum theory. Here the quantum fluctuations of a particle allow it to probe the global properties of the configuration space. We here want to discuss a beautiful result in quantum theory concerning the topology of the classical configuration space.

THEOREM: *Given a classical system with configuration space* M *the inequivalent ways to quantize the system is in unique correspondence with the one dimensional unitary representations of the first homotopy group of* M.

This result yields a host of interesting quantum phenomena, most importantly the notion of particle statistics. This result stems from the U(1) phase invariance of quantum mechanics. In principle, we are free to use this extra U(1) gauge freedom as we please. However, when the configuration space M has a non-trivial first homotopy group $\pi_1(\mathrm{M}) \neq \{e\}$ there are restrictions on this phase. A similar argument to what we will present below is found in [59].

Let M be a topologically non-trivial space and $\tilde{\mathrm{M}}$ its universal cover. The projection map is $\pi : \tilde{\mathrm{M}} \rightarrow \mathrm{M}$, not to be confused with the homotopy groups. We want to know in what way quantum mechanics on the universal covering space is related to that on M itself. The motivation for this is the fact that unless we are dealing with a simply connected configuration space, the U(1) phase is not globally well defined [59]. We consider wavefunctions to be sections of a complex line bundle.

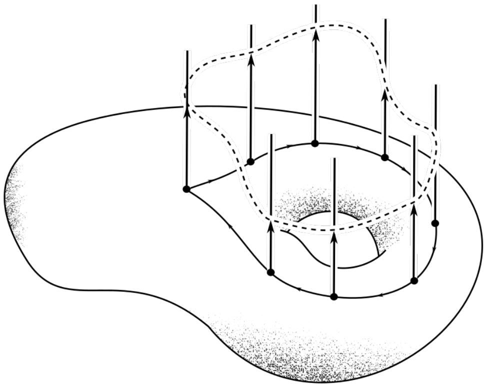


Figure D.1: Quantum mechanics on a topologically non-trivial configuration space, with wavefunctions as sections of a line bundle.

Let us follow a closed loop in M. As we return to the initial point x, the physical state has to be the same, meaning we have to return to the same ray in Hilbert space as before. In other words, there may be a phase ambiguity.

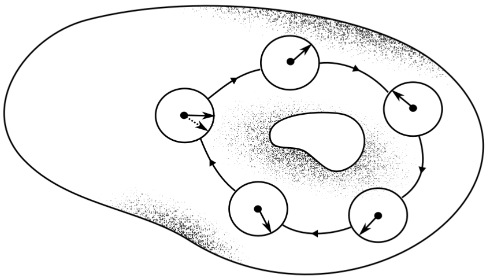


Figure D.2: As we follow a closed loop in the configuration space, we have to arrive at the same physical state, i.e. in the same U(1) equivalence class of wavefunctions.

The discussion on homotopy and covering spaces will be useful here. As we know, there is a natural action of the first homotopy group on the cover that in some sense permutes the points $\pi^{-1}(x)$, for $x \in \mathrm{M}$. From our discussion on line bundles in earlier chapters we know that a bundle over a space with a equivalence relation given by a group action can be considered as a G-invariant bundle on the original space. Hence $\pi_1(\mathrm{M})$-invariant sections over $\tilde{\mathrm{M}}$ should be considered equivalent with sections on M. However, under the action of the homotopy group we can allow a extra phase depending on the homotopy type. Thus, any wave function must satisfy

$$\tilde{\psi}([\gamma]\tilde{x}_1) = \tilde{\psi}(\tilde{x}_2) = \rho([\gamma])\tilde{\psi}(\tilde{x}_1)$$

$$\rho : \pi_1(\mathrm{M}) \to \mathrm{U}(1)$$

when we move around the loop in M. Imagine now moving along a second loop γ_2 after the first γ_1. In this case we have by the same argument that

$$\tilde{\psi}([\gamma_2][\gamma_1]\tilde{x}_1) = \rho([\gamma_1])\rho([\gamma_2])\tilde{\psi}(\tilde{x}_1)$$

But at the same time, we should be able to move along the concatenated curve $\gamma_1 * \gamma_2$ which would mean that

$$\tilde{\psi}([\gamma_2] * [\gamma_1]\tilde{x}_1) = \rho([\gamma_1] * [\gamma_2])\tilde{\psi}(\tilde{x}_1)$$

This leads to the conclusion that the phases must satisfy

$$\rho([\gamma_1] * [\gamma_2]) = \rho([\gamma_1])\rho([\gamma_2])$$

In conclusion, the map $\rho : \pi_1(\mathrm{M}) \to \mathrm{U}(1)$ is a unitary scalar representation of the first homotopy group of the configuration space. Hence for any representation of this group there is a quantum system corresponding to a particular quantization.

The most surprising effect of this result is that of particle statistics. In 1976 J. M. Leinaas and J. Myrheim published a paper [46] in the italian journal Il Nuovo Cimento going into the true nature of identical particles. They argued that the standard argument of exchanging particle coordinates in the wave function has nothing to do with the exchange of two particles, but simply a relabeling of supposedly identical particles. Rather, as we will see, the statistics of identical particles emerges from the topology of the configuration space.

Consider some particle moving on a space M. We assume the space to be topologically trivial in the sense that $\pi_1(\mathrm{M}) = \{e\}$ is the trivial group. In other words, M is simply connected. For a system of many particles moving on M the configuration space is a Cartesian product of n copies of M, i.e.

$$\mathrm{M}^{\times n} = \mathrm{M} \times \ldots \times \mathrm{M}$$

However, as the particles can not pass trough each other, we need to remove the points where the coordinates are equal

$$\mathrm{M}^{\times n} \to \mathrm{M}^{\times n} - \Delta$$

Since the particles are also assumed to be identical, the have to mod out the action of the symmetric group S_n, identifying any two particle positions. In conclusion, we have that the configuration space is

$$\mathrm{C} = (\mathrm{M}^{\times n} - \Delta)/\mathrm{S}_n$$

As we have just seen, when given a configuration space the natural thing to do is to construct quantum mechanics on a universal covering space and in addition do representations of the first homotopy group. For the current configuration space we have

$$\pi_1(\mathrm{C}) = \mathscr{B}_n\,, \;\; \text{d=dim(M)} = 2$$

$$\pi_1(\mathrm{C}) = \mathrm{S}_n\,, \;\; \text{d=dim(M)} \geq 3$$

Here S_n is the symmetric group on n letters and $\mathscr{B}_n$ is the braid group on n strands. To see that this is the case, consider particles moving on a surface Σ. As time passes the worldlines of the particles lies in $\Sigma \times \mathbb{R}$.

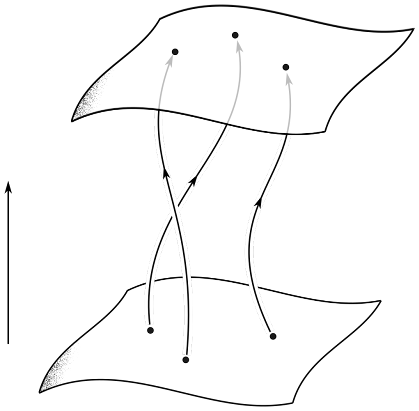


Figure D.3: Worldlines of particles as they move on a surface.

This is nothing but a geometric braid [47]. The set of such braids can be given a group structure simply by placing one braid on top of another. In higher dimension than $2+1$ however, there is enough space to disentangle the braids, reducing the effective action on the worldlines to a permutation. This is why the first homotopy group in $d \geq 3$ is the symmetric group. For the symmetric group, we know that there are two scalar unitary representations, with are simply the symmetric and antisymmetric representations.

$$\psi(x_1,...,x_i,x_j,...,x_n) = \psi(x_1,...,x_j,x_i,...,x_n) = e^{i\cdot 0}\psi(x_1,...,x_j,x_i,...,x_n)$$

$$\psi(x_1,...,x_i,x_j,...,x_n) = -\psi(x_1,...,x_j,x_i,...,x_n) = e^{i\pi}\psi(x_1,...,x_j,x_i,...,x_n)$$

For a general representation we write the phase factor as $e^{i\theta}$ and refer to the angle θ as the statistics type. Bosons correspond to $\theta = 0$, fermions to $\theta = \pi$. The particles with statistics associated with the braid group are called anyons and can take any statistical angle.

It is worth emphasizing that particle statistics is a purely many-body phenomena. It is the topology of the combined many-particle configuration space that gives rise to the inequivalent ways to quantize a many-body system, not single particle configuration spaces. It is quite remarkable that this topological effect can be seen on macroscopic scales, for example in Bose-Einstein condensation.

E

Zeta functions

E.1 The Riemann and Hurwitz zeta functions

Several types of zeta functions has appeared throughout this thesis, both in physical as well as mathematical contexts. The most famous zeta function is the Riemann zeta function, defined as [26][84]

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$$

where $s \in \mathbb{C}$. This series is absolutely convergent for $\Re(s) > 1$, and can be extended to the entire complex plane except the point $s = 1$ [26]. Some values for the zeta functions are

$$\zeta(2) = \frac{\pi^2}{6} \quad ; \quad \zeta(4) = \frac{\pi^4}{90} \quad ; \quad \zeta(6) = \frac{\pi^6}{945}$$

$$\zeta(8) = \frac{\pi^8}{9450} \quad ; \quad \zeta(10) = \frac{\pi^{10}}{93555}$$

$$\zeta(-2n) = 0$$

The points $-2n$ are called the trivial zeros of the zeta function. The Riemann hypothesis states that all on-trivial zeros lie on the line $\Re(s) = 1/2$. At the time of writing this thesis, no solution is in sight.

A generalization of the Riemann zeta function is the Hurwitz zeta function. Similarly to the Riemann zeta, we have the definition [84]

$$\zeta_{\mathrm{H}}(s, a) = \sum_{n=0}^{\infty} (n + a)^{-s}$$

for positive a. For $a = 1$ the Hurwitz zeta and the Riemann zeta coincide. We also have the useful identities

$$\zeta_{\mathrm{H}}(0,a) = \frac{1}{2} - a$$

$$\zeta'(0,a) = \log\Gamma(a) - \frac{1}{2}\log 2\pi$$

These identities are useful when computing functional determinants in quantum field theory.

E.2 Spectral ζ-functions and functional determinants

A further generalization of the above zeta functions is the spectral zeta function, which generalizes the idea of a determinant. In the case of finite dimensional vector spaces, determinants of operators makes sense as a product over eigenvalues, since a finite product of finite numbers must be finite. However, in the case where operators have a countably infinite spectrum, it is not so clear how to define a determinant. The standard trick is to use zeta functions. These infinite dimensional generalizations of determinants naturally appear in quantum field theory, which we will show in a simple case of field theory on the circle.

Consider a finite dimensional vector space V and a operator $\mathrm{Q} : \mathrm{V} \to \mathrm{V}$ with spectrum $\{\mu_n\}$. We define the spectral zeta function by

$$\zeta_{\mathrm{Q}}(s) = \sum_n \mu_n^{-s}$$

Notice that this can be written $\sum \exp(-s\log(\mu_n))$ so that $\exp(-\zeta_q'(0)) = \prod_n \mu_n = \det\mathrm{Q}$. This formula reproduces the determinant as we know it in the finite dimensional case. In stead of naively defining the determinant as a product, we instead use the zeta function definition

$$\det\mathrm{Q} = e^{-\zeta_{\mathrm{Q}}'(0)}$$

which makes sense both in the finite and infinite case. These determinants natural appear in field theory. Consider first a scalar theory $\varphi : \mathbb{R}^2 \to \mathbb{R}$ with partition function

$$\mathcal{Z} = \int \mathcal{D}\varphi e^{-\int d^n x \varphi \mathrm{Q} \varphi}$$

We expand the fields as $\sum_n \alpha_n \varphi_n$ where $\mathrm{Q}\varphi_n = \mu_n \varphi_n$ are eigenfunctions of Q orthogonal in the L^2 sense. The freedom in the fields now lie in the coefficients, and we take as integral measure

$$\mathcal{D}\varphi = \prod_n \frac{d\alpha_n}{\sqrt{\pi}}$$

The partition function then takes the form

$$\mathcal{Z} = \left(\prod_n \int \frac{d\alpha_n}{\sqrt{\pi}}\right) e^{-\sum_n \alpha_n^2 \mu_n} = \prod_n \int \frac{d\alpha_n}{\sqrt{\pi}} e^{-\alpha_n^2 \mu_n} = \det{}^{-1/2}(\mathrm{Q})$$

where we used standard Gaussian integral results. Using our definition of the determinant we have

$$\mathcal{Z} = \exp\left[\frac{1}{2}\zeta'_{\mathrm{Q}}(0)\right]$$

In the operator formalism, the creation and annihilation operators for the fermionic part of Fock space satisfy anticommutation relations to achieve the antisymmetry of fermionic states. In the field integral approach where the fields are classical, this antisymmetry is introduced trough Grassmann algebras. Recall that a vector space together with an operation

$$\cdot : \mathscr{A} \times \mathscr{A} \to \mathscr{A}$$

is called an algebra. A Grassmann algebra is obtained from a vector space with basis $\{\theta_i\}$, $i = 1, ..., n$ with the additional product such that

$$\theta_i \cdot \theta_j + \theta_j \cdot \theta_i = 0$$

We will not keep writing the $\cdot$. Note that $\theta_i \theta_i = 0$, so a generic element of the Grassmann algebra is a polynomial that is at most first order in each of the generators. For example, for $n = 2$ we could have

$$f(\theta_1, \theta_2) = k_0 + k_1\theta_1 + k_2\theta_2 + k_{12}\theta_1\theta_2$$

as any higher order polynomials would include terms like $\theta_1^2 = 0$. Differentiation is defied as one would expect with

$$\frac{\partial \theta_i}{\partial \theta_j} = \delta_{ij}$$

with the additional convention that the theta being differentiated must be moved all the way to the left. This may induce a few minus signs. For example, again with $n = 2$ we have

$$\frac{\partial f}{\partial \theta_1} = k_1 + k_{12}\theta_2$$
$$\frac{\partial f}{\partial \theta_2} = k_2 - k_{12}\theta_1$$

Modulo this sign business however, differentiation is nothing new. Integrals on the other hand is another thing entirely. Note that any integral should satisfy

$$\frac{\partial}{\partial \theta_i} \int d\theta_i f(\theta_i) = \int d\theta_i \frac{\partial f(\theta_i)}{\partial \theta_i} = 0$$

if we assume no boundary terms. For Grassmann variable we define the integral to be the same as the derivative

$$\int d\theta_i f(\theta_i) = \frac{\partial f(\theta_i)}{\partial \theta_i}$$

This may seem strange, but by applying the derivative we see that

$$\frac{\partial}{\partial \theta_i}\int d\theta_i f(\theta_i) = \frac{\partial^2 f(\theta_i)}{\partial \theta_i^2} = 0$$

since all elements are maximal of order one.

Consider now the two sets of Grassmann variable $\{\theta_i\}$ and $\{\overline{\theta}_i\}$ which in some sense are conjugate variable. We want to evaluate the integral

$$\mathrm{I} = \int d\theta_1 d\overline{\theta}_1 ... d\theta_n d\overline{\theta}_n e^{\sum_{i,j=1}^n \overline{\theta}_i \mathrm{M}_{ij}\theta_j}$$

Writing the exponential of the i-sum as the product of exponentials and Taylor expanding the exponential we have that

$$e^{\sum_{i,j=1}^n \overline{\theta}_i \mathrm{M}_{ij}\theta_j} = \prod_i \left(1 + \overline{\theta}_i \sum_j \mathrm{M}_{ij}\theta_j\right)$$

To integrate this notice that integrals over Grassmann algebras essentially picks out the coefficient with most indices, modulo some signs. In the $n = 2$ case, as we then want to integrate over $d\theta_1 d\overline{\theta}_1 d\theta_2 d\overline{\theta}_2$, the only expressions surviving will have to be those that involve all four generators. Writing out the product these terms are

$$\overline{\theta}_1 \mathrm{M}_{11}\theta_1 \overline{\theta}_2 \mathrm{M}_{22}\theta_2 + \overline{\theta}_1 \mathrm{M}_{12}\theta_2 \overline{\theta}_2 \mathrm{M}_{21}\theta_1$$

Writing the generators in appropriate order, the result after taking the integrals (e.g. differentiating) is $\mathrm{M}_{11}\mathrm{M}_{22} - \mathrm{M}_{12}\mathrm{M}_{21} = \det \mathrm{M}$. This result generalizes [28], and we have

$$\mathrm{I} = \int d\theta_1 d\overline{\theta}_1 ... d\theta_n d\overline{\theta}_n e^{\sum_{i,j=1}^n \overline{\theta}_i \mathrm{M}_{ij}\theta_j} = \det \mathrm{M}$$

Notice the difference from the scalar case, where the Gaussian integral of the same type was $\det^{-1/2}\mathrm{M}$. The partition function for fermionic fields ψ is then of the form

$$\mathcal{Z} = \int \mathcal{D}\overline{\psi}\mathcal{D}\psi e^{-\int d^n x \overline{\psi}\mathrm{Q}\psi} = \exp(-\zeta'_{\mathrm{Q}}(0))$$

Of course, the operator Q is usually the Dirac operator $\mathcal{D}$ in d dimensions. Note that both in the scalar and Grassmann case, the field integrals are turned into spectral problems. In this sense, the fields seem almost auxiliary, in the sense that they simply appear in order to calculate determinants.

E.3 Functional determinant on S^1

As promised we look at a simple example. Consider the sigma model $\varphi : \mathrm{S}^1 \to \mathbb{R}$ with action

$$\mathrm{S} = \int_{\mathrm{S}^1} dt(\partial_t\varphi)^2 = \int_{\mathrm{S}^1} dt\varphi(-\partial_t^2)\varphi$$

where we integrated by parts. Hence, we should calculate the zeta-regularized determinant of this Laplacian on the circle. First note that $-\partial_t^2 \exp(iat) = a^2 \exp(iat)$ and periodicity on the circle with circumference c implies that $a_n = \sqrt{\mu_n} = n2\pi/c$. Hence the spectral zeta function is

$$\zeta_{-\partial_t}(s) = 2\left(\frac{c}{2\pi}\right)^{2s} \zeta(2s)$$

where $\zeta(2s)$is the Riemann zeta function, and the factor of 2 in front comes from summing over both the negative and positive integers. Differentiating and changing to the natural variable 2s we have

$$\frac{d}{ds}\zeta_{-\partial_t}(s) = 2\frac{d}{d(2s)}\left[2\left(\frac{c}{2\pi}\right)^{2s}\zeta(2s)\right] = 4\log(c/2\pi)(c/2\pi)^{2s}\zeta(2s)+4(c/2\pi)^{2s}\zeta'(2s)$$

Hence at $s = 0$ we get $\zeta'_{-\partial_t}(0) = -2\log(c)$, by using the zeta function values from above. The determinant is then $\det(-\partial_t^2) = c^2$, and the partition function $\mathcal{Z} = 1/c$.

Bibliography

[1] E. Abdalla, M. B. Abdalla, and D. Rothe. *Non-Perturbative Methods in Two-Dimensional Quantum Field Theory*. World Scientific, 1991.

[2] Alexander Altland and Ben Simons. *Condensed Matter Field Theory*. Cambridge university press, 2010.

[3] Luis Alvarez-Gaumé, Gregory Moore, and Cumrun Vafa. Theta Functions, Modular Invariance and Strings. *Commun. Math. Phys.*, 106(1):1–40, 1989.

[4] Hideo Aoki and Mildred S. Dresselhaus, editors. *Physics of Graphene*. Springer International Publishing Switzerland, 2014.

[5] Michael Atiyah. Topological Quantum Field Theories. *Publ. Math. IHES*, 68:175–186, 1988.

[6] John Baez and Javier P. Muniain. *Gauge Fields, Knots and Gravity*. World Scientific, 1994.

[7] John C. Baez. An Introduction to n-Categories. arXiv:q-alg/9705009, May 1997.

[8] John C. Baez. Quantum Quandaries: A Category-Theoretic Perspective. arXiv:quant-ph/0404040, 2004.

[9] R. A. Bertlmann. *Anomalies in Quantum Field Theory*. Oxford University Press, 1996.

[10] Alexander I. Bobenko and Christian Klein, editors. *Computational Approach to Riemann Surfaces*. Springer Verlag, 2011.

[11] Bernhelm Booß-Bavnbek, Giampiero Esposito, and Matthias Lesch, editors. *New Paths Towards Quantum Gravity*. Springer-Verlag Berlin Heidelberg, 2010.

[12] Jan Hendrik Bruinier, Gerard van der Geer, Hünter Harder, and Don Zagier. *The 1-2-3 of Modular Forms*. Springer-Verlag Berlin Heidelberg, 2008.

[13] Jean-Luc Brylinski. *Loop Spaces, Characteristic Classes and Geometric Quantization*. Birkhäuser, 1993.

[14] C. P. Burgess and C. A. Lütken. One-dimensional flows in the quantum Hall effect. *Nuclear Physics B.*, 500:367–378., 1997.

[15] T. H. Buscher. Path-integral derivation of quantum duality in nonlinear sigma-models. *Physical Review B*, 201(4), 1988.

[16] Sean Carroll. *Spacetime and Geometry*. Pearson Education Limited, 2014.

[17] Xie Chen, Zheng-Cheng Gu, and Xiao-Gang Wen. Local unitary transformation, long-range quantum entanglement, wave function renormalization, and topological order. *Physical Review B*, 82(155138), 2010.

[18] Dariusz Chruscinski and Andrzej Jamiolkowski. *Geometric Phases in Classical and Quantum Mechanics*. Birkhäuser, 2004.

[19] E. de Faria and W. de Melo. *Mathematical Aspects of Quantum Field Theory*. Cambridge university press, 2010.

[20] Pierre Deligne, Pavel Etingof, Daniel S. Freed, Lisa C. Jeffrey, David Kazhdan, John W. Morgan, David R. Morrison, and Edward Witten, editors. *Quantum Fields and Strings: A Course for Mathematicians*, volume 1,2. American Mathematical Society, 1999.

[21] F. Diamond and J. Shurman. *A First Course in Modular Forms*. Springer Schience+Business Media, Inc., 2005.

[22] R. Dijkgraaf. *A Geometrical Approach to Two-Dimensional Conformal Field Theory*. PhD thesis, Utrecht University, 1989.

[23] R. Dijkgraaf. Les Houches Lectures on Fields, Strings and Duality. arXiv:hep-th/9703136v1, 1997.

[24] M R. Douglas. Spaces of Quantum Field Theories. arXiv:hep-th/1005.2779v2, 2010.

[25] Boris Dubrovin. Geometry of 2d topological field theories. arXiv:hep-th/9407018, 1994.

[26] Emilio Elizalde. *Ten Physical Applications pf Spectral Zeta Functions*. Springer-Verlag Berlin Heidelberg, 2 edition, 2012.

[27] Eduardo Fradkin. *Field Theories of Condensed Matter Physics*, volume 2. Cambridge university press, 2013.

[28] Philippe Di Francesco, Pierre Mathieu, and David Sénéchal. *Conformal Field Theory*. Springer Verlag, 1997.

[29] Daniel S. Freed. Short-range Entanglement and Invertible Field Theories. arXiv:1406.7278v2 [cond-mat.str-el], 2014.

[30] Daniel S. Freed and Michael J. Hopkins. Reflection Positivity and Invertible Topological Phases. arXiv:1604.06527v2 [hep-th], 2016.

[31] Juan Mateos Guilarte, José María Muños Porras, and Marina de la Torre Mayado. Elliptic theta functions and the fractional quantum Hall effect. *Journal of geometry and physics*, 27:297–332, 1998.

[32] B. Gustafsson and J. Peetre. Notes on Projective Structures on Complex Manifolds. *Nagoya Math. J.*, 116:63–88., 1989.

[33] Heekyoung Hahn. Eisenstein series associated with Γ0(2). *The Ramanujan Journal*, 15(2):235–257, 2008.

[34] Brian C. Hall. *Lie Groups, Lie Algebras and Representations*. Springer International Publishing Switzerland, 2 edition, 2003.

[35] Brian C. Hall. *Quantum Theory for Mathematicians*. Springer, 2013.

[36] Hans Halvorson, editor. *Deep Beauty*. Cambridge university press, 2011.

[37] Kentaro Hori, Sheldon Katz, Albrecht Klemm, Rahul Pandharipande, Richard Thomas, Cumrun Vafa, Ravi Vakil, and Eric Zaslow. *Mirror Symmetry*, volume 1. American Mathematical Society, 2003.

[38] D. Huybrechts. *Complex Geometry*. Springer Verlag, 2005.

[39] Jürgen Jost. *Riemannian Geometry and Geometric analysis*. Springer Verlag, 1995.

[40] Jürgen Jost. *Geometry and Physics*. Springer-Verlag, 2009.

[41] Lloyd James Peter Kilford. *Modular Forms: A Classical and Computational Introduction.* Imperial College Press, 2008.

[42] Frances Kirwan, Sylvie Paycha, and Tsou Sheung Tsun, editors. *Workshop on Moduli Spaces in Mathematics and Physics*. Hindawi publishing corporation, 1998.

[43] Alexei Kitaev. Periodic table for topological insulators and superconductors. arXiv:0901.2686 [cond-mat.mes-hall], 2009.

[44] Neal Koblitz. *Introduction to Elliptic Curves and Modular Forms*. Springer-Verlag New York Inc., 1984.

[45] John M. Lee. *Introduction to Topological Manifolds*. Springer, 2 edition, 2011.

[46] J. M. Leinaas and J. Myrheim. On the Theory of Identical Particles. *Il Nuovo Cimento*, 37 B(1), 1977.

[47] Joshua Lieber. Introduction to Braid Groups, 2011.

[48] C. A. Lütken. Geometry of renormalization group flows constrained by discrete global symmetries. *Nuclear Physics B*, 396, 1993.

[49] C. A. Lütken. Global phase diagrams for charge transport in two dimensions. *J. Phys. A: Math. Gen*, 26, 1993.

[50] C. A. Lütken. Holomorphic anomaly in the quantum Hall system. *Nuclear Physics B*, 759:343–369, 2006.

[51] C. A. Lütken. private communication, 2017.

[52] C. A. Lütken and G. G. Ross. Duality in the quantum Hall system. *Physical Review B*, 45(4), 1992.

[53] C. A. Lütken and G. G. Ross. Delocalization, duality and scaling in the quantum Hall effect. *Physical Review B*, 48(4), 1993.

[54] C. A. Lütken and G. G. Ross. Experimental pobes of emergent symmetries in the quantum Hall system. *Nuclear Physics B*, 850:321–338, 2011.

[55] C. A. Lütken and G. G. Ross. Quantum critical Hall exponents. *Physics Letters A*, 387:262–265, 2014.

[56] C.A. Lütken. Introduction to the role of modular symmetries in graphene and other two-dimensional materials. *Contemporary Physics*, 2015.

[57] Michele Maggiore. *A Modern Introduction to Quantum Field Theory*. Oxford University Press, 2005.

[58] M.Green, J.Schwarz, and E. Witten. *Superstring Theory Volume 2*. Cambridge university press, 1987.

[59] Giuseppe Morandi. *The Role of Topology in Classical and Quantum Physics*. Springer Verlag, 1992.

[60] M. Nakahara. *Geometry, Topology and Physics*. IOP Publishing, second edition edition, 2004.

[61] Charles Nash. *Differential Topology and Quantum Field Theory*. Academic Press, Inc., 1991.

[62] A. Newlander and L.Nirenberg. Complex Analytic Coordinates in Almost Complex Manifolds. *Annals of Mathematics*, 65(3):391–404, May 1957.

[63] J. Nissinen and C. A. Lütken. Renormalization-group potential for quantum Hall effects. *Physical Review B*, 85(155123), 2012.

[64] J. Nissinen and C. A. Lütken. The quantum Hall curve. arXiv:1207.4693v1 [cond-mat.str-el], 2012.

[65] Qian Niu, D. J. Thouless, and Yong-Shi Wu. Quantized Hall conductance as a topological invariant. *Physical Review B*, 31(6), 1985.

[66] Lars Onsager. Reciprocal Relations in Irreversible Processes. *Physical Review*, 37, 1931.

[67] M.E. Peskin and D.V. Schroeder. *An Introduction to Quantum Field Theory*. Westview Press, 1995.

[68] Robert A. Rankin. *Modular forms and functions*. Cambridge university press., 1977.

[69] Shinsei Ryu, Andreas P Schnyder, Akira Furusaki, and Andreas W W Ludwig. Topological insulators and superconductors: ten-fold way and dimensional hierarchy. *New Journal of Physics*, 12, 2010.

[70] Andreas P Schnyder, Shinsei Ryu, Akira Furusaki, and Andreas W W Ludwig. Classifications of topological insulators and superconductors in three spatial dimensions. arXiv:0803.2786v3 [cond-mat.mes-hall], 2008.

[71] B. Schoeneberg. *Elliptic Modular Functions*. Springer-Verlag, 1974.

[72] J.-P. Serre. *A Course in Arithmetic*. Springer-Verlag New York Inc., 1973.

[73] Tudor D. Stanescu. *Introduction to Topological Quantum Matter and Quantum Computation*. CRC Press, 2017.

[74] Kurt Sundermeyer. *Symmetries in Fundamental Physics*, volume 176. Springer International Publishing Switzerland, 2 edition, 2014.

[75] Richard J. Szabo. *An Introduction to String Theory and D-Brane Dynamics*. Imperial College Press., 2 edition, 2011.

[76] D. J. Thouless, M. Kohmoto, M. P. Nightingale, and M. den Nijs. Quantized Hall Conductance in a Two-dimensional Periodic Lattice. *Physical Review Letters.*, 49(6):405, 1982.

[77] Ulrike Tillmann, editor. *Topology, Geometry and Quantum Field Theory*. 308. Cambridge university press, 2004.

[78] David Tong. Lectures on Kinetic Theory. Lecture notes available at http://www.damtp.cam.ac.uk/user/tong/kinetic.html, 2012.

[79] David Tong. Lectures on String Theory. arXiv:0908.0333v3 [hep-th], 2012.

[80] David Tong. The Quantum Hall Effect. arXiv:1606.06687, 2016.

[81] D. C. Tsui, H. L. Stormer, and A. C. Gossard. Two-Dimensional Magnetotransport in the Extreme Quantum Limit. *Physical Review Letters.*, 48(22), 1982.

[82] K. v. Klitzing, G. Dorda, and M. Pepper. New Method for High-Accuracy Determination of the Fine-Structure Constant Based on Quantized Hall Resistance. *Physical Review Letters.*, 45(6), 1980.

[83] Raimund Varnhagen. Topology and Fractional Quantum Hall Effect. *Nuclear Physics B*, 443(3):501–515, 1995.

[84] Michel Waldschmidt, Pierre Moussa, Jean-Marc Luck, and Claude Itzykson, editors. *From Number Theory to Physics*. Springer-Verlag, 1992.

[85] Jing Wang, Biao Lian, , and Shou-Cheng Zhang. Quantum anomalous Hall effect in magnetic topological insulators. *Physica Scripta*, 2015(T164), 2015.

[86] Lawrence C. Washington. *Elliptic Curves: Number Theory and Cryptography*. Chapman and Hall/CRC, 2003.

[87] Xiao-Gang Wen. *Quantum Field Theory of Many-Body Systems*. Oxford University Press, 2004.

[88] Edward Witten. Three Lectures On Topological Phases Of Matter. arXiv:1510.07698v2 [cond-mat.mes-hall], 2016.

[89] A. B. Zamolodchikov. "irreversibility" of the flux of the renormalization group in a 2d field theory. *Pis'ma Zh. Eksp. Teor. Fiz.*, 43(12):565–567, 1986.

[90] Y. Zhang, Z. Jiang, J. P. Small, M. S. Purewal, Y.-W. Tan, M. Fazolollahi, J. D. Chudow, J. A. Jaszczak, H. L. Stormer, and P. Kim. Landau Level Splitting in Graphene in High Magnetic Fields. *Physical Review Letters*, 96(136806), 2006.

[91] J. Zinn-Justin. *Quantum Field Theory and Critical Phenomena*. Oxford Science Publications, 4 edition, 2002.