Modelling topological phases of matter 2018–9

Prof. Dr. Dr. h.c. Ralf Meyer

June 7, 2022

Contents

1 Introduction

Physicists have recently discovered materials with exotic properties, which are explained by topology. This course introduces these exotic materials and some of the mathematics relevant to their study.

I do not require many prerequisites. I only require the analysis courses of the first two years in Göttingen, including Analysis on Manifolds and Functional Analysis (see, in particular, my lecture notes on Functional Analysis from the Summer Term 2018). I do not expect a strong background in physics. A course in quantum mechanics would be helpful to follow this course. I will, however, begin with a brief crash course on the mathematics of quantum mechanics, which suffices for this course. The main topics in this course are mathematical in nature and are motivated by applications in physics. The better your background in physics, the more will you be able to see the physical motivation of the mathematical questions.

The classification of topological insulators will use the language of homotopy theory of topological spaces. The particular homotopy groups that occur are related to K-theory, but I will not highlight this because I do not want to assume the audience to know about K-theory of C*-algebras.

The materials that this course will focus on are called "topological insulators". "Topological insulators are electronic materials that have a bulk band gap like an ordinary insulator but have protected conducting states on their edge or surface." [5] These surface states conduct electricity, although the interior of the material is insulating. The current on the surface may be quantised and have other exotic properties. It is believed that these edge currents are quite robust. In particular, they cannot dissipate into the interior of the material because it does not conduct. "The surface or edge states of a topological insulator lead to a conducting state with properties unlike any other known one-dimensional 1D or 2D electronic systems. In addition to their fundamental interest, these states are predicted to have special properties that could be useful for applications ranging from spintronics to quantum computation." [5]

We will only do a little quantum mechanics, not enough to derive the electric properties of materials from anything close to first principles. Our treatment is phenomenological in nature and gives only a very rough argument why Hamiltonians of a certain type may model the electronic properties of materials. These Hamiltonians are described through functions from a d-torus with values in self-adjoint $N \times N$-matrices, where d is the dimension of the material and N is the number of internal degrees of freedom per unit cell of the crystal in our model. Another important point will be the description of symmetries in quantum mechanics. Some of the most interesting topological materials are based on symmetries which are realised by anti-unitary operators on the Hilbert space, such as time-reversal symmetry. "A topological insulator, like an ordinary insulator, has a bulk energy gap separating the highest occupied electronic band from the lowest empty band. The surface or edge in two dimensions of a topological insulator, however, necessarily has gapless states that are protected by time-reversal symmetry." [5]

Besides the Hamiltonian, our material comes with a distinguished energy, the Fermi level. The material is a conductor if the Fermi energy belongs to the spectrum of the

Hamiltonian. If not, then it is an insulator. Assume that we have an insulator. Then the energy bands below the Fermi energy define a vector bundle over the d-torus. If there are no further symmetries to take into account, then the (stable) isomorphism class of this vector bundle is the topological phase of the material. Interesting effects occur if this vector bundle is non-trivial. So the first topic after the physical background will be a study of vector bundles over tori. The stable isomorphism classes of vector bundles over a compact topological space X form a group, which is the reduced topological K-theory of X. We shall briefly review some basic properties of K-theory. We need not go far because we mainly need vector bundles over the d-tori for dimensions $d = 1, 2, 3$ (and maybe $d = 4$ for systems that are also periodic in time).

Then we consider the variations on K-theory that occur in the presence of symmetries. Here we focus on those symmetries that are either anti-unitary or anti-commute with the Hamiltonian. There are 10 different combinations of these symmetries. Correspondingly, all 10 K-theory groups of a space appear. We use Clifford algebras to describe these different K-theory groups systematically. We compute some of these K-groups for low-dimensional tori. These computations involve explicit formulas for certain characteristic classes.

The most interesting features of topological materials appear only when we break periodicity by adding an edge, that is, when we truncate the Hamiltonians to a half-space. This replaces one circle by a Toeplitz algebra. This is an example of a C*-algebra, and K-theory for C*-algebras is essential to study the bulk–boundary correspondence, which determines the conductivity properties of the boundary of a material from the properties of the bulk. We shall use van Daele's description of K-theory, which is not the standard one, and we shall use the K-theory long exact sequence without proof in order to describe the bulk–boundary correspondence.

2 A crash course in quantum mechanics

In classical mechanics, the state of a physical system is described by the positions and velocities of all objects in the system. The possible positions form a submanifold X of $\mathbb{R}^{3N}$, where N is the number of particles and X describes constraints on the particles. Usually, if there are no constraints, $X = \mathbb{R}^{3N}$. Taking velocities into account gives points in the tangent space TX of X (or a subspace of TX if there are more constraints on the possible velocities). We assume for simplicity that the dynamics of the system is generated by a potential $V\colon X \to \mathbb{R}$, called *potential energy*. The total energy is the sum of the potential energy and the *kinetic energy*. We assume that the latter is

$$E_{\text{kin}} = \sum_{j=1}^{N} m_j \frac{\|\dot{x}_j\|^2}{2},$$

where $m_j \in (0, \infty)$ and $\dot{x}_j \in \mathbb{R}^3$ are the mass and the velocity of the jth particle, respectively. Then the *momentum* of the jth particle is $m_j \cdot \dot{x}_j \in \mathbb{R}^3$. So in our simple model momenta and velocities differ only by the constant factors m_j. (In general, momenta belong to the cotangent space T^*X instead of the tangent space, and the relationship between momenta and velocities may become non-linear if the formula for the kinetic energy is more complicated than quadratic.)

2.1 Hilbert space and states in quantum mechanics

The situation in quantum mechanics is quite different. Quantum mechanics only makes probabilistic statements. An infinite amount of energy would be needed to localise an object at a single point. Instead, a particle is described by a *wave function*, which is already an extended objects. For a single particle moving freely in $\mathbb{R}^3$ – without spin, we will learn about spin a bit later – the wave function is a unit vector ψ in the Hilbert space $L^2(\mathbb{R}^3, \mathrm{d}x)$. And k particles of different type moving in $\mathbb{R}^3$ (still without spin) are described by a unit vector in the Hilbert space $L^2(\mathbb{R}^{3k}, \mathrm{d}x)$. There is a catch if these particles are all of the same kind. If they are all electrons or all protons or all neutrons, then we must replace $L^2(\mathbb{R}^{3k}, \mathrm{d}x)$ by the subspace of *antisymmetric functions* with respect to exchanging the coordinates. Particles like this are called *Fermions*. Some particles, most notably photons, are *Bosons*, and for them we should use the subspace of *symmetric functions* instead. Summing up, there is a mildly complicated recipe for choosing the Hilbert space that describes a given physical system. The state of the system is described by a unit vector in that Hilbert space.

Lemma 2.1. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be infinite-dimensional, separable Hilbert spaces, that is, neither is finite-dimensional and both contain countable dense subsets. Let $\psi_i \in \mathcal{H}_i$ for $i = 1, 2$ be unit vectors. There is a unitary operator $U\colon \mathcal{H}_1 \to \mathcal{H}_2$ with $U(\psi_1) = \psi_2$.*

Proof. Fix $i \in \{1, 2\}$. The orthogonal complement of ψ_i in $\mathcal{H}_i$ has an orthonormal basis. It is countable because the Hilbert space in question is separable and infinite-dimensional. Add ψ_i to extend this to an orthonormal basis $(\varphi_0, \varphi_1, \varphi_2, \dots)$ of $\mathcal{H}_i$ with

$\varphi_0 = \psi_i$. Writing vectors in $\mathcal{H}_i$ in this orthonormal basis gives a unitary operator $U_i \colon \mathcal{H}_i \to \ell^2(\mathbb{N})$, $U_i(\xi)(n) \coloneqq \langle \varphi_n \,|\, \xi \rangle$. It maps ψ_i to the characteristic function $|0\rangle$ of 0. Now $U = U_2^{-1} U_1 \colon \mathcal{H}_1 \to \mathcal{H}_2$ will do the job. □

Roughly speaking, all choices of a Hilbert space and a unit vector in it are equivalent. So this choice does not yet contain information about the physical system. This information is in the observables.

2.2 Observables

Definition 2.2. An *observable* in a quantum mechanical system described by a Hilbert space $\mathcal{H}$ is a possibly unbounded, self-adjoint operator on $\mathcal{H}$.

We are particularly interested in the observables that correspond to the positions and momenta of the particles in the system because we expect these to determine the state of the system, as in classical mechanics. Another important observable is the total energy.

Example 2.3. Suppose we are dealing with a single particle in $\mathbb{R}^3$ (without spin), described by a unit vector ψ in the Hilbert space $L^2(\mathbb{R}^3, \mathrm{d}x)$. The *position observables* are the operators that multiply a wave function with the coordinate functions:

$$(X_j \psi)(x_1, x_2, x_3) \coloneqq x_j \cdot \psi(x_1, x_2, x_3)$$

for $j = 1, 2, 3$, $x_1, x_2, x_3 \in \mathbb{R}$, and $\psi \in \mathcal{H}$, such that $x_j \cdot \psi \in \mathcal{H}$. We have $\psi \in \mathcal{H}$ and $x_j \cdot \psi \in \mathcal{H}$ for $j = 1, 2, 3$ if and only if

$$\int_{\mathbb{R}^3} (1 + x_1^2 + x_2^2 + x_3^2) \cdot |\psi(x_1, x_2, x_3)|^2 \, \mathrm{d}x < \infty.$$

So the operator X_j is only defined on a dense subspace of $\mathcal{H}$. This is typical of unbounded operators.

Definition 2.4. A (densely defined) *unbounded operator* on a Hilbert space $\mathcal{H}$ is a linear map $T \colon \mathcal{D}_T \to \mathcal{H}$ for a dense linear subspace $\mathcal{D}_T \subseteq \mathcal{H}$.

The *adjoint* T^* of an unbounded operator is defined on $\psi \in \mathcal{H}$ if and only if there is a vector $T^*\psi \in \mathcal{H}$ such that $\langle T^*\psi \,|\, \varphi \rangle = \langle \psi \,|\, T\varphi \rangle$ for all $\varphi \in \mathcal{D}_T$. An unbounded operator $T \colon \mathcal{H} \supseteq \mathrm{Dom}(T) \to \mathcal{H}$ is *self-adjoint* if $T^* = T$.

The position operator, as defined above, is indeed self-adjoint in this sense. Self-Adjointness is a subtle notion. It is much stronger than the condition $\langle T\psi \,|\, \varphi \rangle = \langle \psi \,|\, T\varphi \rangle$ for all $\varphi, \psi \in \mathcal{D}_T$ (such operators are called *formally self-adjoint* or *symmetric*). The technical nuance is that the domain of T^* for a formally self-adjoint operator T may be much bigger than the domain of T.

Unbounded operators are important for quantum mechanics in general, and so we must use them in this section. We will later describe topological materials by phenomenological models where the relevant observables are bounded. So we shall not highlight the technical difficulties of unbounded operators. An example of a self-adjoint operator for which the domain is crucial information is discussed in the proof of Theorem 3.3.

Example 2.5. The *momentum operators* for a single particle are given by the formula

$$(P_j\psi)(x_1,x_2,x_3) \coloneqq -\mathrm{i}\hbar\frac{\partial}{\partial x_j}\psi(x_1,x_2,x_3)$$

for $j=1,2,3$. Here the domain is the maximal subspace of $\psi\in L^2(\mathbb{R}^3,\mathrm{d}x)$ for which $P_j\psi$ defined as a distribution belongs to $L^2(\mathbb{R}^3,\mathrm{d}x)$. More explicitly, P_j is defined at $\psi\in L^2(\mathbb{R}^3,\mathrm{d}x)$ and maps it to $\xi\in L^2(\mathbb{R}^3,\mathrm{d}x)$ if

$$\langle\xi\,|\,\varphi\rangle = \langle\psi\,|\,P_j^*\varphi\rangle$$

for all smooth functions $\varphi\colon\mathbb{R}^3\to\mathbb{C}$ of compact support, where

$$(P_j^*\varphi)(x_1,x_2,x_3) \coloneqq \mathrm{i}\hbar\frac{\partial}{\partial x_j}\varphi(x_1,x_2,x_3).$$

Roughly speaking, we have shifted the unboundedness of differentiation to the auxiliary function φ. This is how distributions and weak solutions of differential equations work. The momentum operators are self-adjoint. The quickest proof uses that self-adjointness is invariant under conjugation by unitaries and that the Fourier transform on $\mathbb{R}^3$ intertwines P_j and the multiplication operator X_j (up to a constant factor depending on the mass and the conventions used in the Fourier transform).

2.3 Functional calculus and spectral measure

How are "observables" seen in experiments? This depends on the *Borel functional calculus*. Let T be a possibly unbounded self-adjoint operator on a Hilbert space $\mathcal{H}$. The Borel functional calculus for T maps a Borel function $f\colon\mathbb{R}\to\mathbb{C}$ to a normal operator $f(T)$ on $\mathcal{H}$. It satisfies $\overline{f}(T)=\big(f(T)\big)^*$. So $f(T)$ is self-adjoint if f is real-valued. The operator $f(T)$ is bounded of norm at most $\|f\|_\infty$ if f is bounded. The map $f\mapsto f(T)$ is an algebra homomorphism (this claim must be interpreted suitably for unbounded functions, but we do not discuss this here). If $f\geq 0$, then $f(T)\geq 0$. If $f(x)\coloneqq x$, then $f(T)=T$ as expected. And the function $R_\lambda(x)\coloneqq(\lambda-x)^{-1}$ for $\lambda\in\mathbb{C}\setminus\mathbb{R}$ is mapped to the resolvent $R_\lambda(T)\coloneqq(\lambda-T)^{-1}$. Actually, $f(T)$ depends only on the restriction of f to the spectrum $\sigma(T)\subseteq\mathbb{R}$; that is, $f(T)=g(T)$ if $f|_{\sigma(T)}=g|_{\sigma(T)}$.

Proposition 2.6. *Let $\psi\in\mathcal{H}$ be a state, that is, a unit vector, and let T be a possibly unbounded self-adjoint operator on $\mathcal{H}$. There is a unique regular Borel probability measure μ_ψ on $\mathbb{R}$ such that*

$$\langle\psi\,|\,f(T)\psi\rangle = \int_{\mathbb{R}} f(x)\,\mathrm{d}\mu_\psi(x)$$

for all bounded Borel functions $f\colon\mathbb{R}\to\mathbb{C}$. The support of μ_ψ is $\sigma(T)$.

Proof. Let $\mathrm{C}_0(\mathbb{R})$ be the space of continuous functions on $\mathbb{R}$ that vanish at ∞. We define a linear functional

$$\xi\colon\mathrm{C}_0(\mathbb{R})\to\mathbb{R},\qquad f\mapsto\langle\psi\,|\,f(T)\psi\rangle.$$

The properties of the functional calculus imply that this linear functional is contractive and positive, that is, $\|f\|_\infty \leq 1$ implies $|\langle \psi \,|\, f(T)\psi\rangle| \leq 1$ and $f \geq 0$ implies $\langle \psi \,|\, f(T)\psi\rangle \geq 0$. The Riesz–Markov–Kakutani Representation Theorem gives a unique regular Borel measure μ_ψ on $\mathbb{R}$ such that

$$\langle \psi \,|\, f(T)\psi\rangle = \int_{\mathbb{R}} f(x)\,\mathrm{d}\mu_\psi(x)$$

for all $f \in \mathrm{C}_0(\mathbb{R})$. We have not defined the Borel functional calculus. So we cannot really say why this formula continues to hold for bounded Borel functions. In fact, the Borel functional calculus is defined to make this formula true. If f is the constant function 1, then $f(T)$ is the identity operator on $\mathcal{H}$. Therefore, $\int_{\mathbb{R}} 1\,\mathrm{d}\mu_\psi(x) = \mu_\psi(\mathbb{R}) = 1$, that is, the measure μ_ψ is a probability measure; here we use that ψ is a unit vector. □

The measure defined in Proposition 2.6 is called the *spectral measure* of T and ψ. These spectral measures link quantum mechanical theory and experiment:

> *Assume that an experiment prepares the quantum mechanical system in the state ψ and then observes T. If this experiment is repeated many times, then the distribution of the measured values converges to the probability distribution μ_ψ. In particular, the expectation value of the measurement is $\langle \psi \,|\, T\psi\rangle$ because the expectation value of μ_ψ is $\langle \psi \,|\, T\psi\rangle$. And the measured values belong to the support of μ_ψ almost surely. Since μ_ψ is supported in $\sigma(T)$, it follows that the measured values belong to $\sigma(T)$ almost surely, no matter which state ψ is used.*

Now let $X_1, \ldots, X_n$ be observables that "strongly" commute, that is, their resolvents commute as bounded operators. Strong commutation for unbounded operators is a subtle property, but we do not highlight this here. Bounded operators strongly commute if and only if they commute in the usual sense. Assuming strong commutation, there is a joint Borel functional calculus, which defines a densely defined normal operator $f(X_1, \ldots, X_n)$ for each Borel function $f\colon \mathbb{R}^n \to \mathbb{C}$, which is bounded if f is bounded. This map from Borel functions to operators has similar properties as for a single self-adjoint operator. The definition of the spectral measure also extends to several strongly commuting observables. And so does the link to experiments:

> *Several strongly commuting observables may be measured simultaneously, and the probability distribution of the measured values is given by the spectral measure associated to these observables and the state of the system.*

Strongly commuting operators $X_1, \ldots, X_n$ also have a "joint" spectrum in $\mathbb{R}^n$. To define it, let

$$\gamma\colon \mathrm{C}_0(\mathbb{R}^n) \to \mathbb{B}(\mathcal{H})$$

be the joint continuous functional calculus for the operators $X_1, \ldots, X_n$. The kernel of γ is an ideal in $\mathrm{C}_0(\mathbb{R}^n)$. There is a closed subset $\Sigma \subseteq \mathbb{R}^n$ such that $\gamma(f) = 0$ if and

only if $f|_\Sigma = 0$. There is a unique injective *-homomorphism $\dot\gamma\colon \mathrm{C}_0(\Sigma) \to \mathbb{B}(\mathcal{H})$ with $\dot\gamma(f|_\Sigma) = \gamma(f)$ for all $f \in \mathrm{C}_0(\mathbb{R}^n)$. The closed subset $\Sigma \subseteq \mathbb{R}^n$ is the joint spectrum of $X_1, \dots, X_n$. It has the following physical interpretation: the simultaneous measurement of $X_1, \dots, X_n$ almost surely gives values in Σ, no matter which state ψ is used.

Example 2.7. We continue Example 2.3, considering the position observables X_1, X_2, X_3 on the Hilbert space $L^2(\mathbb{R}^3, \mathrm{d}x)$ describing a single particle. These three self-adjoint operators strongly commute. Therefore, they have a joint Borel functional calculus and a joint spectral measure in any state ψ. If $f\colon \mathbb{R}^3 \to \mathbb{C}$ is a bounded Borel function, then $f(X_1, X_2, X_3)$ is simply the operator of pointwise multiplication by f:

$$(f(X_1, X_2, X_3)\psi)(x_1, x_2, x_3) \coloneqq f(x_1, x_2, x_3) \cdot \psi(x_1, x_2, x_3)$$

for all $(x_1, x_2, x_3) \in \mathbb{R}^3$, $\psi \in L^2(\mathbb{R}^3, \mathrm{d}x)$. So the joint spectral measure μ_ψ is given by

$$\begin{aligned}\int_{\mathbb{R}^3} f(x)\,\mathrm{d}\mu_\psi(x) &= \langle \psi \mid f(X_1, X_2, X_3)\psi\rangle = \int_{\mathbb{R}^3} \overline{\psi(x)} f(x)\psi(x)\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3 \\ &= \int_{\mathbb{R}^3} f(x)|\psi(x)|^2\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3.\end{aligned}$$

So μ_ψ is the measure with the density function $|\psi(x)|^2 \in L^1(\mathbb{R}, \mathrm{d}x)$. Therefore, experiments that measure the position of the particle determine exactly the function $|\psi|^2$, as the density of the probability measure that describes the results of all position measurements.

The above results also show that the joint spectrum of the position observables is all of $\mathbb{R}^3$. Indeed, the position of a particle can be anywhere in $\mathbb{R}^3$.

The image of the Borel functional calculus for the position observables is the algebra $L^\infty(\mathbb{R}^3, \mathrm{d}x)$, represented on $\mathcal{H} = L^2(\mathbb{R}^3, \mathrm{d}x)$ by pointwise multiplication operators. This subalgebra of $\mathbb{B}(\mathcal{H})$ is *maximal Abelian*, that is, an operator $T \in \mathbb{B}(\mathcal{H})$ that commutes with all multiplication operators must itself be a multiplication operator. Therefore, any observable that can simultaneously be measured together with all three position observables must itself be of the form $f(X_1, X_2, X_3)$ for some Borel function f. When we simultaneously measure X_1, X_2, X_3 and $f(X_1, X_2, X_3)$ and get the values x_1, x_2, x_3, y, then almost surely $y = f(x_1, x_2, x_3)$. So measuring $f(X_1, X_2, X_3)$ in addition to X_1, X_2, X_3 gives no new information. Roughly speaking, we cannot measure more than the three position observables at the same time. These simultaneous measurements tell us the absolute value of ψ, but not its phase $\psi/|\psi|$. So they do not determine the state ψ uniquely. Our intuition from classical mechanics says that we should measure both position and momentum to describe the state of the system completely. The argument above already shows, however, that they cannot commute with the position observables. So they cannot be measured simultaneously with position.

2.4 Self-adjoint operators and representations of the real numbers

The joint Borel functional calculus applies to the three momentum operators. This allows us to define the operators $f(P_1, P_2, P_3)$ for any Borel function $f\colon \mathbb{R}^3 \to \mathbb{C}$. This is particularly interesting for the function $f_k(p) \coloneqq \exp(-\mathrm{i}k \cdot p)$ for $k, p \in \mathbb{R}^3$. The

corresponding operator $f_k(P) = \exp(-\mathrm{i}k \cdot P)$ is unitary. Since the functional calculus is a unital *-homomorphism and $\exp(-\mathrm{i}k \cdot p) \cdot \exp(-\mathrm{i}l \cdot p) = \exp(-\mathrm{i}(k+l) \cdot p)$ and $\exp(-\mathrm{i}0 \cdot p) = 1$ for all $k, l, p \in \mathbb{R}^3$, these unitaries satisfy $f_k(P) \cdot f_l(P) = f_{k+l}(P)$ and $f_0(P) = \mathrm{id}_{\mathcal{H}}$. The properties of the Borel functional calculus also imply that the map $\mathbb{R}^3 \ni k \mapsto f_k(P)\psi \in \mathcal{H}$ is continuous for each $\psi \in \mathcal{H}$. So $k \mapsto f_k(P)$ is a *continuous representation* of the additive group $\mathbb{R}^3$. Formally,

$$\frac{\partial}{\partial k_j} \exp(-\mathrm{i}k \cdot P)\psi \bigg|_{k=0} l = -\mathrm{i}P_j\psi = -\hbar\frac{\partial}{\partial x_j}\psi.$$

Stone's Theorem turns this into a precise statement. First, the vector ψ belongs to the domain of P_j and $P_j\psi = \mathrm{i}\varphi$ if and only if the function $\mathbb{R} \to \mathcal{H}$, $k_j \mapsto \exp(-\mathrm{i}k_j \cdot P_j)\psi$ is differentiable at 0 with derivative φ. So a self-adjoint operator is determined uniquely by the continuous group representation of $\mathbb{R}$ that it generates by functional calculus. Conversely, any continuous representation of $\mathbb{R}$ is of this form for a unique self-adjoint operator. And this generalises to a canonical bijection between continuous representations of the additive group $\mathbb{R}^n$ and n-tuples of strongly commuting self-adjoint operators.

For the momentum observables, we may guess a continuous representation of $\mathbb{R}^3$ with generators P_j, namely, the translation representation. Define

$$(\tau_k\psi)(x) \coloneqq \psi(x - \hbar k)$$

for all $k, x \in \mathbb{R}^3$, $\psi \in L^2(\mathbb{R}^3, \mathrm{d}x)$. The operators τ_k are clearly unitary and define a continuous unitary representation of $\mathbb{R}^3$ on $L^2(\mathbb{R}^3, \mathrm{d}x)$. We compute $\frac{\partial}{\partial k_j}\tau_k\psi(x)\Big|_{k=0} = -\hbar\frac{\partial}{\partial x_j}\psi(x)$. So the generators of τ_k are indeed the differential operators $P_j \coloneqq -\mathrm{i}\hbar\frac{\partial}{\partial x_j}$ for $j = 1, 2, 3$.

2.5 The Hamiltonian

What is the energy observable for a single particle? Clasically, the kinetic energy of a particle is given by the momentum: $E_{\mathrm{kin}} = P^2/2m$. This suggests to define it in quantum mechanics by

$$E_{\mathrm{kin}} \coloneqq \sum_{j=1}^{3} \frac{1}{2m} P_j^2 = -\sum_{j=1}^{3} \frac{\hbar^2}{2m}\frac{\partial^2}{\partial x_j^2}.$$

The potential energy is a (Borel) function $V\colon \mathbb{R}^3 \to \mathbb{R}$ of position. So it should be described by the Borel functional calculus as $V(X_1, X_2, X_3)$. We have mentioned in Example 2.7 that this is the operator of pointwise multiplication by V: $(V\psi)(x) \coloneqq V(x) \cdot \psi(x)$. The total energy is the sum of both pieces:

$$H \coloneqq -\sum_{j=1}^{3} \frac{\hbar^2}{2m}\frac{\partial^2}{\partial x_j^2} + V(x). \tag{2.1}$$

This very important operator is called the *Schrödinger operator*. It generates the dynamics of the quantum mechanical system. Namely, let ψ_t be the state of the system at time t.

This function $\mathbb{R} \to \mathcal{H}$ solves the linear ordinary differential equation

$$\mathrm{i}\hbar \frac{\mathrm{d}\psi_t}{\mathrm{d}t} = H\psi_t.$$

If ψ is the state at time $t = 0$, then this is solved by

$$\psi_t = U_t(\psi), \qquad U_t := \exp(-\mathrm{i}Ht/\hbar).$$

For this to make sense, we need the operator H above to be self-adjoint. This is a subtle issue. We have not even specified the domain of H carefully, and self-adjointness depends on the domain. There are, in fact, potentials for which the operator H above is not self-adjoint. This is related to the problem in classical mechanics that an object may escape to ∞ (or arrive from ∞) in finite time; then the time-development does not exist as a one-parameter group of homeomorphisms of phase space. We shall ignore this issue here and refer to [12, 13].

The description of dynamics above uses a picture where the states change and the observables are constant in time. So measuring the observable S at time t when the system is in the state ψ at time 0 amounts to measuring S on the state $U_t(\psi)$. This is equivalent to measuring the observable $U_{-t}SU_t$ on ψ. So there is an equivalent picture where the state is treated as constant and the observables change by $S \mapsto U_{-t}SU_t = \exp(\mathrm{i}Ht/\hbar)S\exp(-\mathrm{i}Ht/\hbar)$. This picture is less intuitive, but often useful.

2.6 Bound states

One of the most important questions about a quantum mechanical system concerns its long-term dynamics. We are particularly interested in the "transport" properties of H: is a state ψ "moved" by the dynamics generated by H? There are some clear cases for non-movement and movement, and there is a lot of room in between where more sophisticated notions are needed to describe the transport behaviour. We begin with the obvious case of non-movement.

Definition 2.8. A unit vector ψ with $H(\psi) = \lambda \cdot \psi$ for some $\lambda \in \mathbb{R}$ is called a *bound state* of the Hamiltonian H.

So a bound state is just an eigenvector of H. Let ψ be a bound state of H with eigenvalue λ. Then $\exp(-\mathrm{i}Ht/\hbar)\psi = \exp(-\mathrm{i}\lambda t/\hbar)\cdot\psi$. So the dynamics generated by H only multiplies ψ with a time-dependent phase factor. Such a phase factor does not change the spectral measures for any observable T because

$$\langle \exp(-\mathrm{i}\lambda t/\hbar)\cdot\psi \mid f(T)\exp(-\mathrm{i}\lambda t/\hbar)\cdot\psi\rangle = \langle \psi \mid f(T)\psi\rangle$$

for all operators $f(T)$ and all $t \in \mathbb{R}$. Hence when we prepare the system in the state ψ, then measuring an observable T at some time $t \in \mathbb{R}$ gives the same probability distribution for all times t. In particular, the position of the particle is described by the same probability distribution for all times t. So the particle described by the wave function ψ does not move. This justifies calling states as above "bound" states.

In general, the eigenvectors of a self-adjoint operator on a Hilbert space with different eigenvalue are orthogonal. There is a maximal orthonormal set of eigenvectors. But this need not be a basis. To understand what happens, we turn the Hamiltonian into a multiplication operator. One version of the Spectral Theorem for self-adjoint operators says that this is possible:

Theorem 2.9. *Let $\mathcal{H}$ be a Hilbert space and let T be a self-adjoint operator on $\mathcal{H}$. There are a σ-finite measure space (X, μ), a measurable function $\hat{T}\colon X \to \mathbb{R}$, and a unitary $U\colon \mathcal{H} \to L^2(X, \mu)$, such that UTU^* is the operator of pointwise multiplication with $\hat{T}$ on $L^2(X, \mu)$, with its canonical domain $\{\psi \in L^2(X, \mu) : \hat{T} \cdot \psi \in L^2(X, \mu)\}$.*

This theorem makes the long-term dynamics of a quantum mechanical system transparent. The functional calculus is compatible with conjugation by unitaries:

$$U \exp(-\mathrm{i}Tt/\hbar)U^* = \exp(-\mathrm{i}UTU^*t/\hbar).$$

And for the multiplication operator UTU^*, exponentiation simply gives the operator of multiplication by the function $\exp(-\mathrm{i}\hat{T}t/\hbar)$.

Now we consider the most obvious case of movement. This is the "free" dynamics, where the potential energy vanishes, so the Hamiltonian H is just the kinetic energy:

$$H = -\sum_{j=1}^{3} \frac{\hbar^2}{2m} \frac{\partial^2}{\partial x_j^2}.$$

In a corresponding classical system without potential energy, a particle just moves with constant speed. In particular, unless its constant speed is exactly 0, the particle will eventually leave any finite region. In quantum mechanics, the free Hamiltonian H above commutes with the momentum operators. Therefore, the momentum is constant in time: $\exp(-\mathrm{i}Ht/\hbar)P_j \exp(\mathrm{i}Ht/\hbar) = P_j$ for $j = 1, 2, 3$. So the particle moves with constant momentum. And the probability that this momentum vanishes is 0. So we expect that, almost surely, the particle will eventually leave any compact region in position space. To make this precise, we should turn H into a multipliccation operator. This is accomplished here by the Fourier transform on $\mathbb{R}^3$, which turns H into the operator of pointwise multiplication with the function $c \cdot (x_1^2 + x_2^2 + x_3^2)$ for an unimportant positive constant c that depends on conventions. In particular, the operator H has no eigenvectors that are in $L^2(\mathbb{R}^3, \mathrm{d}x)$, that is, there are no bound states. The functions $x \mapsto \exp(-\mathrm{i}k \cdot x)$ for $k \in \mathbb{R}^3$ are eigenfunctions of the differential operator H with eigenvalue $\hbar\|k\|^2/2m$, but these functions are not square-integrable. In a sense, this function outside the Hilbert space describes a wave that travels with constant speed. The Fourier transform $\hat{f}(x) \coloneqq \int_{\mathbb{R}^3} f(k) \exp(-\mathrm{i}k \cdot x)\, \mathrm{d}k$ gives us mixtures of such waves. If f has small support around k, the function $\hat{f}$ describes a "wave packet" that consists of waves that have approximately the same speed. The dynamics will multiply the function f pointwise with $\exp(-\mathrm{i}ct\|k\|^2)$. This factor oscillates more and more wildly on the support of f, even if the latter is very small. Nevertheless, the support does not change. So the wave packet will still keep the same distribution of momenta, only the shape of the wave packet may change as time flows.

We could now try to make precise in which sense the states of the form $\hat{f}(x) \coloneqq \int_{\mathbb{R}^3} f(k) \exp(-\mathrm{i}k \cdot x)\, \mathrm{d}k$ describe "transport" of a particle with approximately constant speed, or that the dynamics will eventually move the particle out of any compact region in $\mathbb{R}^3$. But the details get subtle and confusing. There are different ways to define what it means for a state to remain "localised" (not moving) or to be "delocalised" (moving), and the different notions do not agree. The existing mathematical theory has its merits, but it cannot yet explain all the phenomena expected by physicists.

2.7 Rotations and angular momentum

Another important symmetry of space besides translations are rotations. Since the Lebesgue measure on $\mathbb{R}^3$ is invariant under rotations, they act on $L^2(\mathbb{R}^3, \mathrm{d}x)$ in an obvious way by unitaries: if $A \in \mathrm{SO}(3)$, then the corresponding unitary operator on $L^2(\mathbb{R}^3, \mathrm{d}x)$ is the unitary defined by $(\sigma_A \psi)(x) \coloneqq \psi(A^{-1} \cdot x)$ for all $\psi \in L^2(\mathbb{R}^3, \mathrm{d}x)$, $x \in \mathbb{R}^3$. This defines a unitary representation of $\mathrm{SO}(3)$ on $L^2(\mathbb{R}^3, \mathrm{d}x)$.

Any rotation may be written as a product of rotations around the coordinate axes (compare "Euler angles"). Rotations around an axis form a circle, which is a quotient of $\mathbb{R}$. So a continuous representation of the group $\mathrm{SO}(3)$ is determined uniquely by the generators of rotations around the three coordinate axes. Up to the constant $\hbar$, these generators are the *angular momentum operators* L_1, L_2, L_3. Rotations around the x_1-axis are of the form

$$R_1(\varphi) \coloneqq \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\varphi) & -\sin(\varphi) \\ 0 & \sin(\varphi) & \cos(\varphi) \end{pmatrix}.$$

They act on $L^2(\mathbb{R}^3, \mathrm{d}x)$ by

$$\sigma(R_1(\varphi))\psi(x) = \psi(R_1(-\varphi)x) = \psi(x_1, \cos(\varphi)x_2 + \sin(\varphi)x_3, -\sin(\varphi)x_2 + \cos(\varphi)x_3).$$

The generator of this continuous representation of $\mathbb{R}$ is the angular momentum operator L_1:

$$(L_1\psi)(x) = \mathrm{i}\hbar \frac{\partial}{\partial \varphi} \sigma(R_1(\varphi))\psi(x)\Big|_{\varphi=0} = \mathrm{i}\hbar x_3 \frac{\partial \psi}{\partial x_2} - \mathrm{i}\hbar x_2 \frac{\partial \psi}{\partial x_3}.$$

The formulas for the other axes are obtained by cyclic permutation of the coordinates:

$$\begin{aligned} L_1 &= \mathrm{i}\hbar x_3 \frac{\partial}{\partial x_2} - \mathrm{i}\hbar x_2 \frac{\partial}{\partial x_3}, \\ L_2 &= \mathrm{i}\hbar x_1 \frac{\partial}{\partial x_3} - \mathrm{i}\hbar x_3 \frac{\partial}{\partial x_1}, \\ L_3 &= \mathrm{i}\hbar x_2 \frac{\partial}{\partial x_1} - \mathrm{i}\hbar x_1 \frac{\partial}{\partial x_2}. \end{aligned}$$

2.8 Spin

Our discussion so far only applies to particles of spin 0. Electrons or protons have spin 1/2. The correct Hilbert space to describe such particles is $L^2(\mathbb{R}^3, \mathrm{d}x) \otimes \mathbb{C}^2$. Equivalently, we

take square-integrable functions with values in $\mathbb{C}^2$ or pairs of square-integrable scalar-valued functions. These two functions correspond to the spin-up and spin-down internal degrees of freedom. More generally, a particle with spin $s/2$ for $s \in \mathbb{N}$ is described by the Hilbert space

$$\mathcal{H} \coloneqq L^2(\mathbb{R}^3, \mathrm{d}x) \otimes \mathbb{C}^{s+1}.$$

The position and momentum operators are defined on this Hilbert space by exactly the same formulas as on $L^2(\mathbb{R}^3, \mathrm{d}x)$, now applied to functions $\psi\colon \mathbb{R}^3 \to \mathbb{C}^{s+1}$. The angular momentum operators on $\mathcal{H}$ are of the form

$$L_i^{\mathrm{tot}} \coloneqq L_i \otimes 1 + 1 \otimes S_i, \qquad i = 1, 2, 3,$$

for certain self-adjoint matrices S_i acting on $\mathbb{C}^{s+1}$. Each L_i^{tot} is self-adjoint and so generates a continuous representation σ_i of $\mathbb{R}$ on $\mathcal{H}$. When we interpret $\sigma_i(\varphi)$ as the action of rotation by the angle φ, we meet a problem, however: the rotation by $2\pi = 360^\circ$ is the operator $\psi \mapsto (-1)^s\psi$, which is not the identity map for half-integer spin. Only $\sigma_i(4\pi) = \sigma_i(2\pi)^2$ is always the identity map.

Consider the unitary $\psi \mapsto \lambda \cdot \psi$ for some $\lambda \in \mathbb{C}$ with $|\lambda| = 1$, briefly, $\lambda \in \mathrm{U}(1)$. If T is a self-adjoint operator on $\mathcal{H}$, then $\langle \psi \,|\, T\psi \rangle = \langle \lambda \cdot \psi \,|\, T(\lambda \cdot \psi) \rangle$. The same holds for $f(T)$ for any Borel function. So spectral measures are not changed by the unitary above. Therefore, applying this unitary has no physical consequences. Multiplication by a scalar in $\mathrm{U}(1)$ is, in a sense, the "identity" automorphism of the physical system. This is best understood in the picture where a unitary U moves observables instead of states. Instead of mapping states by $\psi \mapsto U\psi$ and leaving observables fixed, we may leave states fixed and move observables by $T \mapsto U^*TU$. This has the same effect on the inner products $\langle \psi \,|\, T\psi \rangle$ and is therefore physically equivalent. Now $U^*TU = T$ if U is scalar multiplication with some $\lambda \in \mathrm{U}(1)$. So it describes the identity automorphism of the space of observables. In fact, the representations of $\mathbb{R}$ generated by the total angular momentum operators L_i^{tot} generate a representation of $\mathrm{SO}(3)$ on observables. That is, when we write $g \in \mathrm{SO}(3)$ as a product of rotations around the coordinate axes, say, $g = R_1(\varphi_1)R_2(\varphi_2)R_3(\varphi_3)$, then the resulting operator $\sigma(g) \coloneqq \sigma_1(\varphi_1)\sigma_2(\varphi_2)\sigma_3(\varphi_3)$ on $L^2(\mathbb{R}^3, \mathrm{d}x) \otimes \mathbb{C}^{s+1}$ depends on the choice of $\varphi_1, \varphi_2, \varphi_3$, but only by a scalar in $\mathrm{U}(1)$. So the operator $T \mapsto \sigma(g)T\sigma(g)^*$ on observables is well defined. And if $g, h \in \mathrm{SO}(3)$, then $\sigma(g)\sigma(h) = \tau(g, h)\sigma(gh)$ for some $\tau(g, h) \in \mathrm{U}(1)$ (actually, we even have $\tau(g, h) \in \{\pm 1\}$ here). So the action of $\mathrm{SO}(3)$ on observables becomes an ordinary group representation.

Definition 2.10. Let G be a group. A map $\varrho\colon G \to \mathrm{U}(\mathcal{H})$ that satisfies

$$\varrho(gh) = \tau(g, h)\varrho(g)\varrho(h)$$

for all $g, h \in G$ with scalars $\tau(g, h) \in \mathrm{U}(1)$ is called a *projective representation* of G.

The upshot of the discussion above is that the "representation" of rotations that is relevant for a particle of spin $s/2$ is only a projective representation of $\mathrm{SO}(3)$ on $L^2(\mathbb{R}^3, \mathrm{d}x) \otimes \mathbb{C}^{s+1}$ if s is odd; in particular, this is the case for electrons and protons. But a projective representation is good enough for quantum mechanics because it induces an ordinary group representation on the algebra of observables.

For the group SO(3), we may get rid of the adjective "projective" when we replace it by its universal covering, which is isomorphic to the group SU(2). In general, a Lie group with torsion-free fundamental group has no non-trivial projective representations. That is, any projective representation becomes an ordinary representation when we multiply the unitaries with suitable scalars in U(1). Therefore, passing to a universal covering removes the difficulty of projective representations. The price is to replace the rotation group SO(3) by the less intuitive group SU(2). It also follows that groups such as $\mathbb{R}^n$ or $\mathbb{T}^n := \mathbb{R}^n/\mathbb{Z}^n$ have no non-trivial projective representations because $\mathbb{R}^n$ has trivial fundamental group and $\mathbb{T}^n$ has the torsion-free fundamental group $\mathbb{Z}^n$.

Both SO(3) and SU(2) are compact groups. Therefore, any representation of them splits as a direct sum of irreducible representations in a unique way, and irreducible representations are finite-dimensional. The group SU(2) has a unique irreducible representation on $\mathbb{C}^d$ for each $d \geq 1$. This is exactly the representation that is generated by the spin-angular momentum operators S_1, S_2, S_3 on $\mathbb{C}^d$, that is, with spin $s = (d-1)/2$. So the possible values of the spin are in bijection with the irreducible representations of the group SU(2) or, equivalently, the irreducible projective representations of the group SO(3).

I do not plan to go much further into the story of the spin. Honesty requires that it be mentioned because the most important particles for our course have spin $1/2$. So we should replace the Hilbert space $L^2(\mathbb{R}^3, \mathrm{d}x)$ by $L^2(\mathbb{R}^3, \mathrm{d}x) \otimes \mathbb{C}^2$. If I knew more quantum chemistry, I should say much more about the spin because features of the spin ("spin–orbit coupling") have been the main idea to find materials that are non-trivial topological insulators. It must be pointed out also that spin–orbit coupling is a *relativistic* effect. The Schrödinger equation treated above does not yet contain it.

2.9 Symmetries

We have already discussed some special symmetries – translations and rotations – and we have found that these are represented on the Hilbert space of the quantum mechanical system by a projective representation of the underlying symmetry group. But there are also symmetries that are represented by *anti*-unitary operators.

Definition 2.11. An *anti-unitary* operator on a Hilbert space $\mathcal{H}$ is an $\mathbb{R}$-linear bijection $\Theta\colon \mathcal{H} \to \mathcal{H}$ with $\langle \Theta(\psi) \,|\, \Theta(\varphi)\rangle = \langle \varphi \,|\, \psi\rangle$ for all $\varphi, \psi \in \mathcal{H}$. An *involution* is a map $\mathcal{H} \to \mathcal{H}$ whose square is the identity map.

By polarisation, the condition $\langle \Theta(\psi) \,|\, \Theta(\varphi)\rangle = \langle \varphi \,|\, \psi\rangle$ for all $\varphi, \psi \in \mathcal{H}$ holds if and only if $\Theta(\psi)$ is a unit vector for all unit vectors ψ.

Wigner's Theorem says that any symmetry of a quantum-mechanical system is realised by a unitary or anti-unitary operator on the Hilbert space (compare [3, Section 1.1]). So there are no other possibilities. The product of two anti-unitary operators is unitary.

The symmetry of time-reversal cannot act by a unitary operator. This is because it leaves positions unchanged, but it maps momenta (like velocities) to their negatives. A unitary operator U doing this cannot leave the canonical commutation relations

$[X_j, P_j] = \mathrm{i}\hbar$ intact:

$$[U^* X_j U, U^* P_j U] = [X_j, -P_j] = -\mathrm{i}\hbar = -U^* \mathrm{i}\hbar U.$$

But if U is anti-unitary, then $-\mathrm{i}\hbar = U^* \mathrm{i}\hbar U$. So time-reversal symmetry must act by an anti-unitary operator Θ. When we do time-reversal twice, it should be a trivial symmetry, that is, $\Theta^2 = \lambda$ for some $\lambda \in \mathrm{U}(1)$. Actually, since $\Theta^2 \circ \Theta = \Theta \circ \Theta^2$ and $\overline{\lambda} \circ \Theta = \Theta \circ \lambda$, we must have $\lambda = \pm 1$. Both signs occur. Roughly speaking, $\Theta^2 = +1$ for a single boson (a particle with integer spin) and $\Theta^2 = -1$ for a single fermion (a particle with half-integer spin).

Example 2.12. Let Θ be the operator of complex conjugation on $L^2(\mathbb{R}^3, \mathrm{d}x)$. That is, $(\Theta\psi)(x) \coloneqq \overline{\psi(x)}$ for all $x \in \mathbb{R}^3$. This operator is an anti-unitary involution. Let H be a Schrödinger operator as in (2.1). Then $\Theta^{-1} H \Theta = \Theta H \Theta = H$. This implies $\Theta^{-1} \exp(\mathrm{i}tH)\Theta = \exp(-\mathrm{i}tH)$. So the action of Θ has the effect of replacing a time t by $-t$. In fact, Θ is the standard time-reversal symmetry for a particle of spin 0.

If Θ_0 and Θ are anti-unitary operators on $\mathcal{H}$, then $U \coloneqq \Theta\Theta_0^{-1}$ is a unitary operator on $\mathcal{H}$ with $\Theta = U\Theta_0$. In this way, we may rewrite all anti-unitaries Θ on $\mathcal{H}$ through unitaries and a "reference anti-unitary" Θ_0. We are particularly interested in anti-unitary operators with square ± 1. If, say, $\Theta_0^2 = 1$, then $\Theta^2 = (U\Theta_0)^2 = U\Theta_0^{-1} U \Theta_0$, and we want this to be ± 1. Since $UU^* = 1$, this is equivalent to $U^* = \pm\Theta_0^{-1} U \Theta_0$ or $U = \pm(\Theta_0^{-1} U \Theta_0)^*$. If Θ_0 is a complex conjugation operator as in Example 2.12, then $\Theta_0^{-1} U \Theta_0$ behaves like the complex conjugate of U and so $(\Theta_0^{-1} U \Theta_0)^*$ behaves like a transpose of U. So the condition $U = \pm(\Theta_0^{-1} U \Theta_0)^*$ means that U is symmetric or skew-symmetric. We now consider a simple case of this.

Example 2.13. For a single electron, the time-reversal symmetry on $L^2(\mathbb{R}^3, \mathrm{d}x) \otimes \mathbb{C}^2$ acts by an operator of the form $\Theta = (1 \otimes U)\Theta_0$, where Θ_0 is complex conjugation as in Example 2.12 and $U \in \mathbb{M}_2(\mathbb{C})$ is a unitary matrix that satisfies $U\overline{U} = -1$, so that $\Theta^2 = -1$; here $\overline{U}$ means the matrix with the complex conjugate entries. For a unitary matrix, $U\overline{U} = -1$ is equivalent to $\overline{U} = -U^*$, and to $U = -U^t$, that is, U is skew-symmetric. Up to multiplication by scalars, which does not matter, this forces

$$U = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

So

$$\Theta(\psi_1, \psi_2) = (-\overline{\psi_2}, \overline{\psi_1})$$

for $\psi_1, \psi_2 \in L^2(\mathbb{R}^3, \mathrm{d}x)$.

Let Θ be a unitary or anti-unitary operator on the Hilbert space $\mathcal{H}$ of a quantum mechanical system. For it to be a symmetry, it is usually required that Θ commutes with the Hamiltonian, $\Theta \circ H = H \circ \Theta$.

Definition 2.14. A *time-reversal symmetry* of the system is an anti-unitary operator with $\Theta \circ H = H \circ \Theta$ and $\Theta^2 = \pm 1$.

Time-reversal symmetry is a crucial feature for the most interesting topological insulators. Since we are mostly dealing with Fermions, the sign in most cases is $\Theta^2 = -1$.

There is no special name for unitary operators with $\Theta \circ H = H \circ \Theta$, such as rotations or translations or reflections: these are just called symmetries. In the context of topological materials, most authors also allow operators that satisfy $\Theta \circ H = -H \circ \Theta$ instead of $\Theta \circ H = H \circ \Theta$.

Definition 2.15. An operator with $\Theta \circ H = -H \circ \Theta$ and $\Theta^2 \in \mathrm{U}(1)$ is called a *chiral symmetry* when it is unitary, and a *particle-hole symmetry* when it is anti-unitary.

The name particle-hole symmetry will be explained at the end of Section 2.10. If Π is a particle-hole symmetry, then Π is anti-unitary and hence $\Pi^2 = \pm 1$ as for time-reversal symmetries. If Ξ is a chiral symmetry, then so is $\mu \cdot \Xi$ for any $\mu \in \mathrm{U}(1)$, and this unitary even describes the same symmetry. Since $(\mu \cdot \Xi)^2 = \mu^2 \cdot \Xi^2 = \mu^2 \cdot \lambda$, the choice $\mu = \lambda^{-1/2}$ makes $(\mu \cdot \Xi)^2 = 1$. So it is no loss of generality to assume $\Xi^2 = 1$ for chiral symmetries.

The condition $\Xi \circ H = -H \circ \Xi$ is equivalent to $\Xi^{-1} \circ H \circ \Xi = -H$. Then H and $-H$ have the same spectrum because the spectrum is invariant under conjugating by unitaries and anti-unitaries. So a particle-hole or chiral symmetry is impossible if the Hamiltonian is bounded below and not bounded above. Since this is quite common, some authors consider symmetries with $\Xi \circ H = -H \circ \Xi$ to be unphysical and try to avoid them by describing the system in other ways (see, for instance, [9]). We shall argue below that Hamiltonians that are unbounded both above and below are physically reasonable in certain circumstances. And we shall describe materials using tight-binding models. Since these have bounded Hamiltonians anyway, they admit chiral and particle-hole symmetries.

A Schrödinger operator with potential does not have any obvious chiral or particle-hole symmetries. The topological insulator materials studied so far are based on spin–orbit coupling, which is a relativistic effect. So the Schrödinger operator is not the right starting point for such systems. In fact, the physics articles usually work in tight-binding models, work in a very abstract setting without writing down concrete models, or study a Schrödinger operator with periodic potential.

Let G be the group of generalised symmetries of the physical system, allowing time-reversal, particle-hole and chiral symmetries. As in [3], we define

$$\varphi(g) = \begin{cases} +1 & \text{if } g \text{ is unitary,} \\ -1 & \text{if } g \text{ is anti-unitary,} \end{cases} \qquad c(g) = \begin{cases} +1 & \text{if } Hg = gH, \\ -1 & \text{if } Hg = -gH. \end{cases}$$

for $g \in G$. A simple calculation shows:

Lemma 2.16. *The two maps $\varphi, c\colon G \rightrightarrows \{\pm 1\}$ are group homomorphisms, that is, $\varphi(gh) = \varphi(g)\varphi(h)$ and $c(gh) = c(g)c(h)$ for all $g, h \in G$.*

We combine the two homomorphisms into a single homomorphism

$$(\varphi, c)\colon G \to \{\pm 1\} \times \{\pm 1\} = \{1, \theta, \pi, \xi\},$$

where $1 = (+1, +1)$, $\theta = (-1, +1)$, $\pi = (-1, -1)$, $\xi = (+1, -1)$. The range of (φ, c) is a subgroup $H \subseteq G$. There are five possibilities for H: the trivial group, the whole group, and the groups $\{1, x\}$ for $x \in \{\theta, \pi, \xi\}$. If $H = \{1\}$, then the system has only ordinary symmetries. If $H = \{\theta\}$, then there is a time-reversal symmetry, but neither a particle-hole nor a chiral symmetry. If $H = \{\pi\}$, there is a particle-hole symmetry, but neither a time-reversal nor a chiral symmetry. If $H = \{\xi\}$, there is a chiral symmetry, but neither a time-reversal nor a particle-hole symmetry. If $H = \{1, \theta, \pi, \xi\}$ then all kinds of symmetries are present. The argument above shows that if the system has two of the symmetries $\{\Theta, \Pi, \Xi\}$, then it automatically has the third type of symmetry. For instance, a product $\Theta\Pi$ of a time-reversal symmetry and a particle-hole symmetry is a chiral symmetry because $\theta\pi = \xi$.

Most of the work on topological insulators so far focusses on the case when $G = G_0 \times H$, where G_0 only contains translations, and (φ, c) is the projection to the second factor. In particular, there are canonical lifts Θ, Π or Ξ in G for the non-trivial elements in H. And these form an isomorphic copy of $H \subseteq \{\pm 1\}^2$. In particular, these symmetries commute with one another (and with G_0 as well), and they square to 1. When we treat symmetries as Hilbert space operators, however, then a relation such as $\Theta^2 = 1$ in G only means that $\Theta^2 \in \mathrm{U}(1)$ because all scalars in $\mathrm{U}(1)$ give the trivial symmetry of the quantum mechanical system. For time-reversal and particle-hole symmetries, we know that $\Theta^2 = \pm 1$ and $\Pi^2 = \pm 1$. For a chiral symmetry Ξ it is no loss of generality to assume $\Xi^2 = 1$. Listing all the possible symmetry types and signs gives the ten cases listed in Table 1. The rows in this table are ordered in a way that becomes natural when one describes these symmetries using Clifford algebras.

H	Θ^2	Π^2
$\{1\}$		
$\{1, \Xi\}$		
$\{1, \Theta\}$	$+1$	
$\{1, \Theta, \Pi, \Xi\}$	$+1$	-1
$\{1, \Pi\}$		-1
$\{1, \Theta, \Pi, \Xi\}$	-1	-1
$\{1, \Theta\}$	-1	
$\{1, \Theta, \Pi, \Xi\}$	-1	$+1$
$\{1, \Pi\}$		$+1$
$\{1, \Theta, \Pi, \Xi\}$	$+1$	$+1$

Table 1: Table of symmetry types.

2.10 Single-Particle approximation

Recall that a single electron is described by a wave function in $\mathcal{H} \coloneqq L^2(\mathbb{R}^3, \mathrm{d}x) \otimes \mathbb{C}^2$. Then a system of k electrons is described by the exterior power $\Lambda^k \mathcal{H}$, completed to a

Hilbert space; this is the same as the subspace of antisymmetric vectors in the Hilbert space tensor product of k copies of $\mathcal{H}$. However, this Hilbert space is so high-dimensional that it is impossible to work with. Already a sample that weighs only 1 g contains about 10^{24} electrons. So we need approximate models for all actual computations.

The models that we shall consider are all "single-particle models". Roughly speaking, we pretend that we study one of the many electrons, moving in the potential created by the other electrons and the atomic nuclei. So we work in the same Hilbert space $\mathcal{H}$ as for a single electron. The Hamiltonian H is an "effective" Hamiltonian for the electron that we singled out. It is reasonable to expect H to be a Schrödinger operator as in (2.1) once again. But now we must use an "effective" mass and an "effective" potential, which take into account the effects of all the other electrons. When we move our electron, then by the Coulomb repulsion this will also move the other nearby electrons. So we are effectively moving several electrons at once. This increases the inertial mass of the electron and forces us to work with an "effective" mass in the kinetic energy. Physicists also speak of a quasi-particle because the object that we are describing is not a true particle, but a particle-like approximation for the behaviour of one electron in the presence of many others.

A crucial feature is still missing, namely, the Pauli exclusion principle: two electrons cannot occupy the same state. This is made mathematically precise by the construction of the Hilbert space for a system of several electrons: we use the *anti*symmetric subspace in the tensor product. In the single-particle approximation, states occupied by other electrons are forbidden for the electron that we are looking at. A flexible Ansatz to model this is to add an operator R on $\mathcal{H}$ with $0 \leq R \leq 1$, such that the probability for a state ψ to be occupied is $\langle \psi \,|\, R\psi \rangle$; this also allows the occupation of states to be uncertain. Under simplifying assumptions that we shall not discuss, statistical physics predicts that this operator R is the following function of the Hamiltonian (that is, the energy) and the temperature:

$$W_T(E) = \left(\exp\left(\frac{E - \mu}{k_B T} \right) + 1 \right)^{-1},$$

where k_B is the Boltzmann constant, T is the temperature, E is the energy, and μ is the *chemical potential* or *Fermi energy*. The operator R is defined by functional calculus for the Hamiltonian: $R = W_T(H)$. The function W_T is also called the *Fermi–Dirac distribution*.

For any temperature T, the function W_T is strictly decreasing and satisfies $W_T(\mu) = 1/2$. This characterises the Fermi energy. And

$$\lim_{T \searrow 0} W_T(E) = \begin{cases} 1 & E < \mu, \\ 1/2 & E = \mu, \\ 0 & E > \mu. \end{cases}$$

So in the zero temperature limit, the states below the Fermi energy are completely filled and those states above are empty. Thus our extra electron is forced to occupy a state

above the Fermi level. If E is not an eigenvalue, then we may model this situation more easily by shrinking the Hilbert space to the image of $\chi_{[\mu,\infty)}(H)$, the spectral projection of H for the interval $[\mu, \infty)$. We shall, however, not use this simplification here.

This completes the description of a single-particle model. It is given by the Hilbert space $\mathcal{H}$, the Hamiltonian H, and the Fermi energy μ. This is, of course, a very simple and crude model. It cannot describe truly quantum features of the interaction between the many electrons, and some interesting physical effects – notably, the *fractional* quantum Hall effect – cannot be explained without these. It seems, however, that the explanation of the fractional quantum Hall effect is still not mathematically rigorous and there has not yet been much progress in understanding topological phases in more complete many-particle models. In this course, we will treat many-particle systems only by the one-particle approximation.

Now we can explain the name "particle-hole symmetry" for an anti-unitary operator Π with $\Pi H = -H\Pi$. Here we are treating a system with Fermi energy 0: this is no true restriction because adding a constant to H will not change the physics, and this allows us to shift the Fermi energy. Instead of studying the behaviour of an extra electron, we may also remove an electron and study the behaviour of a "hole". This hole may also be viewed as a positron, the anti-particle of the electron. Replacing electrons by positrons has the effect of replacing H by $-H$ and of taking complex conjugates everywhere (because of charge conjugation). An anti-unitary operator Π with $\Pi H = -H\Pi$ exists if and only if an electron (particle) and a positron (hole) have the same Hamiltonian, that is, if they are governed by the same physics.

For an ordinary quantum mechanical system, it is reasonable to assume that the Hamiltonian is bounded below. A system with a Hamiltonian that is not bounded below has no ground state. The system can drop to lower and lower energy levels, emitting radiation all the time (this is caused, say, by interaction with light). So the system effectively contains an infinite amount of energy. And if the description of the system is complete, then the Hamiltonian should be unbounded above because otherwise there is a maximal amount of energy that may be put into the system. But using sufficiently hard radiation, we can, in principle, put as much energy as we like into any quantum mechanical system. When we are dealing with the effective Hamiltonian for a Fermionic system, however, then it is no problem if the Hamiltonian is unbounded below. All states with an energy below the Fermi energy are filled – or almost filled if the temperature is finite. At zero temperature, the ground state is given by the lowest energy in the spectrum that lies above the Fermi energy. Even at finite temperature, we cannot extract an infinite amount of energy out of the system.

3 The band theory of conductors and insulators

In this section, we explain that the electronic properties of crystalline materials are described by periodic Hamiltonians. We describe the structure of the spectrum of such Hamiltonians: it is a union of intervals, called energy bands. If the Fermi energy belongs to the interior of the spectrum, the material is a conductor. If the Fermi energy lies outside the spectrum, the material is an insulator.

3.1 Hamiltonians for crystalline materials

An actual crystal has a finite size. The mathematical description becomes easier, however, if we pretend its size to be infinite. This is not a bad approximation because even 1 mm is already pretty large from the point of view of quantum mechanics, and the behaviour of electrons in the interior of the crystal, sufficiently far from the boundary, may be described very well by an approximation where we let the crystal extend all the way to infinity. Special things happen at the boundary of the crystal. In fact, these are what make topological insulators so interesting. In this section, we disregard boundary effects and consider a crystal that fills all of space.

The main feature of crystals is periodicity: there is a lattice $\Lambda \subseteq \mathbb{R}^d$, that is, a cocompact discrete subgroup, so that the crystal is invariant under translations by Λ. A crystal is, of course, an object in dimension $d = 3$. We also treat the cases $1 \leq d \leq 2$ because they are easier and there are relevant physical situations like thin foils or wires of materials. And the general theory works for any $d \in \mathbb{N}$. Any lattice is of the form

$$\Lambda = \mathbb{Z}a_1 + \cdots + \mathbb{Z}a_d$$

for some basis $a_1, \ldots, a_d \in \mathbb{R}^d$. To simplify notation, we shall often change the basis so that $\Lambda = \mathbb{Z}^d$. This change of basis is usually not isometric, so it distorts the geometry. Hence there are situations when we have to remember the actual form of Λ.

We are interested in the electronic properties of a crystalline material. *We assume that these are described well in the single-particle approximation* (see Section 2.10). This gives us a certain Hamiltonian H on the single-particle Hilbert space $\mathcal{H}$, which is $L^2(\mathbb{R}^3, \mathrm{d}x) \otimes \mathbb{C}^2$ for an electron in $\mathbb{R}^3$. When we work non-relativistically, it is reasonable to assume that the effective Hamiltonian is a Schrödinger operator as in (2.1) with a potential V. The function V is Λ-periodic because our crystal has this periodicity. I cannot say how to compute this potential or the effective mass in the Schrödinger operator. I shall leave out the multiplicity factor $\mathbb{C}^2$ in the following. The argument below still works if we tensor $L^2(\mathbb{R}^d)$ with $\mathbb{C}^N$ for any $N \geq 1$ and allow the potential to be a matrix-valued periodic function instead of a scalar-valued function.

Our goal is to understand the structure of the spectrum of the Hamiltonian H. We first discuss the translation operators because their interaction with H causes the band structure of the spectrum that we are interested in.

If $x \in \Lambda$, then there is a unitary operator on $\mathcal{H}$ that describes *translation* by x. On $L^2(\mathbb{R}^d, \mathrm{d}x) \otimes \mathbb{C}^N$, it acts by $(S_x f)(y) \coloneqq f(y - x)$. These operators clearly satisfy $S_x S_y = S_{x+y}$ and $S_{-x} = S_x^*$ for $x, y \in \Lambda$, and S_0 is the identity map. Therefore, any

translations S_x for $x \in \Lambda$ may be writen as a product of the translations S_j for $1 \leq j \leq d$ and their adjoints. The operators S_j are unitary and hence have spectrum in U(1).

We assume that the Hamiltonian H strongly commutes with the translations S_x for all $x \in \Lambda$ or, equivalently, with S_j for $1 \leq j \leq d$. This assumption is reasonable because of the periodicity of the crystal. We should be aware, however, that this is an idealisation. A real-life crystal always has impurities. And its temperature will be non-zero. This causes the atomic nuclei to move in a random way. This destroys exact periodicity.

Definition 3.1. Let $f \in L^2(\mathbb{R}^d, \mathrm{d}x)$. Let $\mathbb{T}^d \coloneqq \mathbb{R}^d/\mathbb{Z}^d$. The *Bloch transform* of f is the function $\mathbb{T}^d \times [0,1]^d \to \mathbb{C}$ defined by

$$(\mathcal{B}f)(k,x) \coloneqq \sum_{n \in \mathbb{Z}^d} \exp(2\pi \mathrm{i} k \cdot n) \cdot f(x-n).$$

The Bloch transform is a relative of the Fourier transform. As for the Fourier transform, its definition on square-integrable functions is slightly trickly. If $f \in L^2(\mathbb{R}^d, \mathrm{d}x)$, then the sum $\sum_{n \in \mathbb{Z}^d} \exp(2\pi \mathrm{i} k \cdot n) \cdot f(x+n)$ may diverge for some k, x. The following proposition shows, however, that the sum exists for almost all k, x and defines a function in $L^2(\mathbb{T}^d \times [0,1]^d, \mathrm{d}x)$.

Proposition 3.2. *The Bloch transform defines a unitary operator from $L^2(\mathbb{R}^d, \mathrm{d}x)$ to $L^2(\mathbb{T}^d \times [0,1]^d, \mathrm{d}x)$.*

Proof. We may decompose $\mathbb{R}^d$ as a disjoint union of translates of the unit cube: $\mathbb{R}^d = \bigsqcup_{n \in \mathbb{Z}^d} [0,1)^d + n$. So $\mathbb{R}^d \cong \mathbb{Z}^d \times [0,1)^d$, and the Lebesgue measure corresponds to the product measure of the Lebesgue measure on $[0,1)^d$ and the counting measure on $\mathbb{Z}^d$. We may replace $[0,1)^d$ by $[0,1]^d$ because they differ by a set of measure 0. So we get a unitary operator

$$L^2(\mathbb{R}^d, \mathrm{d}x) \cong \ell^2(\mathbb{Z}^d) \otimes L^2([0,1]^d, \mathrm{d}x).$$

In addition, the Fourier transform is a unitary operator from $\ell^2(\mathbb{Z}^d)$ to $L^2(\mathbb{T}^d)$, and it induces a unitary operator

$$\ell^2(\mathbb{Z}^d) \otimes L^2([0,1]^d, \mathrm{d}x) \to L^2(\mathbb{T}^d) \otimes L^2([0,1]^d, \mathrm{d}x) \cong L^2(\mathbb{T}^d \times [0,1]^d).$$

The composite unitary operator from $L^2(\mathbb{R}^d)$ to $L^2(\mathbb{T}^d \times [0,1]^d)$ is the Bloch transform. □

The definition of the Bloch transform continues to make sense for $x \in \mathbb{R}^d$ instead of $x \in [0,1]^d$, and we shall now extend the definition in this way. The resulting function on $\mathbb{T}^d \times \mathbb{R}^d$ satisfies

$$(\mathcal{B}f)(k, x+n) = \exp(2\pi \mathrm{i} k \cdot n) \cdot (\mathcal{B}f)(k,x) \tag{3.1}$$

for all $k \in \mathbb{T}^d$, $x \in \mathbb{R}^d$, $n \in \mathbb{Z}^d$. This says that the Bloch transform simultaneously turns all translation operators for $n \in \mathbb{Z}^d$ into multiplication operators. This is useful because the Hamiltonian strongly commutes with the translations. A similar strategy is used to determine the spectrum of, say, the hydrogen atom: first one diagonalises the angular momentum operators, which strongly commute with the Hamiltonian because of rotational symmetry.

A function that satisfies the twisted periodicity condition $\varphi(x+n) = \exp(2\pi \mathrm{i} k \cdot n) \cdot \varphi(x)$ for all $x \in \mathbb{R}^d$, $n \in \mathbb{Z}^d$ for some $k \in \mathbb{T}^d$ is also called a *Bloch wave*. Let $\hat{k} \in \mathbb{R}^d$ lift k. Then we may rewrite a Bloch wave φ as $\exp(2\pi \mathrm{i} \hat{k} \cdot x) \cdot u(x)$ with a $\mathbb{Z}^d$-periodic function $u = \exp(-2\pi \mathrm{i} \hat{k} \cdot x) \cdot \varphi(x)$. on $\mathbb{R}^d$. Summing up, a Bloch wave modulates the plane wave $\exp(2\pi \mathrm{i} \hat{k} \cdot x)$ of momentum $2\pi \hat{k}$ by a periodic factor.

We now rewrite our Schrödinger operator with $\mathbb{Z}^d$-periodic potential using the Bloch transform. For this purpose, we must discuss some unbounded operator technicalities. A Schrödinger operator is unbounded and therefore not defined on all functions. In particular, we need some differentiability for the Laplace operator to be defined. We prefer to work with a core, that is, a subset of the domain that is dense in the graph norm. The smooth functions with compact support or the Schwartz functions form such a core. The actual self-adjoint Schrödinger operator is the closure of the operator on this core given by the same formula. The Bloch transform of a Schwartz function on $\mathbb{R}^d$ is a smooth function $\psi\colon \mathbb{T}^d \times \mathbb{R}^d \to \mathbb{C}$ satisfying the above twisted periodicity condition in the second variable. Conversely, one may show that any such smooth function is the Bloch transform of a unique Schwartz function on $\mathbb{R}^d$. Thus the smooth Bloch functions become a core of the Bloch transform of our Schrödinger operator. The Schrödinger operator acts on a smooth function $\psi\colon \mathbb{T}^d \times \mathbb{R}^d \to \mathbb{C}$ by the same formula as (2.1):

$$(H\psi)(k,x) \coloneqq -\sum_{j=1}^{d} \frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x_j^2}(k,x) + V(x). \tag{3.2}$$

This operator leaves the k-variable untouched: there is a family of operators $(H_k)_{k\in\mathbb{T}^d}$ such that $(H\psi)(k,x) = H_k(x \mapsto \psi(k,x))$. Actually, it may seem at first sight that the family of operators H_k is constant because the formula in (3.2) is the same for all k. This is not the case, however, because the boundary conditions are an integral part of an unbounded operator. Each H_k is a self-adjoint operator on $L^2([0,1]^d)$, and the formula defining it is the same for all k, but the domains are different: the smooth functions that satisfy the boundary condition given by (3.1) for a given $k \in \mathbb{T}^d$ form a core for H_k.

We are going to prove that the operators H_k for $k \in \mathbb{T}^d$ are self-adjoint and depend continuously on k. The relevant notion of continuity here is the continuity of the resolvents $(\lambda - H_k)^{-1}$ for fixed $\lambda \in \mathbb{C} \setminus \mathbb{R}$; the operator $(\lambda - H_k)^{-1}$ exists for all $\lambda \in \mathbb{C} \setminus \mathbb{R}$ because H_k is self-adjoint.

Theorem 3.3. *Let $V\colon \mathbb{R}^d \to \mathbb{R}$ be a $\mathbb{Z}^d$-periodic, bounded Borel function. Rewrite the Schrödinger operator with potential V in terms of operators H_k on Bloch wave functions for $k \in \mathbb{T}^d$ as above. If $k \in \mathbb{T}^d$, then the closure of the unbounded operator H_k is self-adjoint and bounded below by a uniform constant, and its resolvent $(\lambda - H_k)^{-1}$ for $\lambda \in \mathbb{C} \setminus \mathbb{R}$ is compact. The map $k \mapsto (\lambda - H_k)^{-1}$ for fixed $\lambda \in \mathbb{C} \setminus \mathbb{R}$ is norm-continuous.*

Proof. We first discuss the easy case when the potential is 0. Let H_k^0 be the operator in (3.2) with $V = 0$ and given $k \in \mathbb{R}^d$. The functions $\exp(2\pi \mathrm{i}(k+n) \cdot x)$ for $n \in \mathbb{Z}^d$ are Bloch waves with the same quasi-momentum $k + n \equiv k \in \mathbb{T}^d$. They form an orthonormal

basis of the space of $L^2([0,1]^d)$. They are eigenfunctions of H_k^0:

$$-\sum_{j=1}^{3}\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x_j^2}\exp(2\pi\mathrm{i}(k+n)\cdot x)$$
$$=-\frac{\hbar^2}{2m}(2\pi\mathrm{i}(k+n))^2\exp(2\pi\mathrm{i}(k+n)\cdot x)=\frac{2\pi^2\hbar^2}{m}(k+n)^2\exp(2\pi\mathrm{i}(k+n)\cdot x).$$

So H_k^0 has an orthonormal basis of eigenvectors $\big(\exp(2\pi\mathrm{i}(k+n)\cdot x)\big)_{n\in\mathbb{Z}^d}$ with eigenvalues $(k+n)^2\cdot 2\pi^2\hbar^2/m$. These eigenvectors and eigenvalues depend continuously on $k\in\mathbb{R}^d$. A simple argument shows now that the resolvents of H_k^0 are continuous functions of $k\in\mathbb{R}^d$. Since H_k^0 is periodic, so are the resolvents. Hence this continuous function descends to a continuous function on $\mathbb{T}^d$. The eigenvalues of H_k^0 are non-negative, and for any $C>0$, there are only finitely many eigenvalues $n\in\mathbb{Z}^d$ with $(k+n)^2\cdot 2\pi^2\hbar^2/m\le C$. So H_k^0 is a positive operator with compact resolvent, that is, the operators $(\lambda-H_k^0)^{-1}$ for $\lambda\notin\sigma(H_k^0)$ are compact. We have verified the assertions of Theorem 3.3 in the case $V=0$. And the other claims about H are also easy to check in this case. By the way, the spectrum of H^0 is equal to $[0,\infty)$.

Now we are treat the operators H_k with a bounded potential as perturbations of H_k^0. The operator norm $\|(t\mathrm{i}-H_k^0)^{-1}\|$ for $t\in\mathbb{R}$ is equal to the spectral radius of $(t\mathrm{i}-H_k^0)^{-1}$ because this operator is normal. Since the spectrum of H_k^0 is contained in $[0,\infty)$, this spectral radius is at most $|t|^{-1}$, which goes to 0 as $t\to\infty$. Therefore, given a bounded Borel function $V\in L^\infty(\mathbb{R}^d)$, we have $\|(t\mathrm{i}-H_k^0)^{-1}V\|<1$ for $|t|>\|V\|_\infty$. Assume that t is chosen such. Then the Neumann series $\sum((\mathrm{i}t-H_k^0)^{-1}V)^n$ converges absolutely, and

$$\sum_{n=0}^{\infty}((\mathrm{i}t-H_k^0)^{-1}V)^n\cdot(\mathrm{i}t-H_k^0)^{-1}=(1-(\mathrm{i}t-H_k^0)^{-1}V)^{-1}\cdot(\mathrm{i}t-H_k^0)^{-1}=(\mathrm{i}t-H_k^0-V)^{-1}.$$

Hence $(\mathrm{i}t-H_k^0-V)^{-1}$ is compact for $|t|>\|V\|_\infty$, and it depends continuously, even analytically on k because H_k^0 does so and the Neumann series converges absolutely. This implies that the operator H_k^0+V is self-adjoint and that its resolvent is compact; that is, $\mathrm{i}t-H_k^0-V$ remains surjective and $(\mathrm{i}t-H_k^0-V)^{-1}$ remains compact for all $\lambda\in\mathbb{C}\setminus\mathbb{R}$. Assume $V(x)\ge c$ for almost all $x\in\mathbb{R}^d$ and let $\psi\in L^2(\mathbb{R}^d)$ satisfy $\|\psi\|=1$. Then

$$\langle\psi\,|\,H_k\psi\rangle=\langle\psi\,|\,H_k^0\psi\rangle+\langle\psi\,|\,V\cdot\psi\rangle\ge 0+\langle\psi\,|\,c\cdot\psi\rangle=c.$$

So $H_k\ge c$ is bounded below uniformly in k. □

We take some of the assertions made in Theorem 3.3 as the description of an important class of self-adjoint operators. Let H be a Hamiltonian acting on a Hilbert space of the form $L^2(\mathbb{T}^d,\mathcal{K})$ by a family of self-adjoint operators H_k, with each H_k acting on $\mathcal{K}$; so $\mathcal{K}=L^2([0,1]^d)$ in the discussion above. We also assume that each H_k has compact resolvent and that these resolvents depend on k by a norm-continuous function. Let $\mathbb{K}(\mathcal{K})$ denote the C*-algebra of compact operators on $\mathcal{K}$ and let $\mathrm{C}(\mathbb{T}^d,\mathbb{K}(\mathcal{K}))$ denote the C*-algebra of norm-continuous functions $\mathbb{T}^d\to\mathbb{K}(\mathcal{K})$. This is a closed subalgebra of

$\mathbb{B}(L^2(\mathbb{T}^d, \mathcal{K}))$, and our assumptions on H amount to $(\lambda - H)^{-1} \in \mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$ for all $\lambda \in \mathbb{C} \setminus \mathbb{R}$. We briefly say that H is a self-adjoint operator *affiliated with* $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$. (This notation is a special case of the general theory in [18]).

Lemma 3.4. *If H is affiliated with $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$, then the spectrum of H is the union of the spectra of the operators H_k for $k \in \mathbb{T}^d$.*

Proof. Let $\mathbb{K}(\mathcal{K})^+ := \mathbb{K}(\mathcal{K}) \oplus \mathbb{C} \cdot \mathrm{id}_{\mathcal{K}}$. We claim that an element S of $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K})^+)$ is invertible if and only if it is invertible pointwise. First, assume S to be invertible pointwise. Then the pointwise inverses belong to $\mathbb{K}(\mathcal{K})^+$ and depend on $k \in \mathbb{T}^d$ in a norm-continuous fashion. Therefore, S is invertible. Conversely, assume that S_k is not invertible for some $k \in \mathbb{T}^d$. Let $\varepsilon > 0$. Then there is $\psi \in \mathcal{K}$ with $\|\psi\| = 1$ and $\|S_k(\psi)\| < \varepsilon$. Since S_k is norm-continuous, it is strongly continuous. So there is an open neighbourhood U of k with $\|S_{k'}(\psi)\| < \varepsilon$ for all $k' \in U$. Let $h \in \mathrm{C}_0(U)$ be such that $\|h\|_{L^2} = 1$. Then $h(k) \cdot \psi \in L^2(\mathbb{T}^d, \mathcal{K})$ is a unit vector with $\|S(h(k) \cdot \psi)\| < \varepsilon$. Since such vectors $h(k) \cdot \psi$ exist for any $\varepsilon > 0$, it follows that S is not invertible.

The above claim applies to the resolvent $(\lambda - H)^{-1}$ of H for any $\lambda \in \mathbb{C} \setminus \mathbb{R}$. Since the map $z \mapsto (\lambda - z)^{-1}$ is a bijection from $\mathbb{C} \setminus \{\lambda\}$ to $\mathbb{C} \setminus \{0\}$, the spectra of $(\lambda - H)^{-1}$ and H determine each other. Hence the spectrum of H must also be the union of the spectra of the operators H_k. □

The resolvent $(\lambda - H_k)^{-1}$ of a self-adjoint operator is always normal because H_k is self-adjoint. It is compact as well if H is affiliated with $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$. Therefore, $(\lambda - H_k)^{-1}$ has an orthonormal basis of eigenvectors, and the eigenvalues, counted with multiplicity, converge to 0. The eigenvectors of $(\lambda - H_k)^{-1}$ are also eigenvectors of H_k, and the eigenvalues of H_k go to ∞ in absolute value if those of $(\lambda - H_k)^{-1}$ go to 0. The eigenvalues of H_k are real because H_k is self-adjoint.

We also assume now that the operator H is bounded below. This is true for Schrödinger operators with bounded periodic potential by Theorem 3.3. Then the eigenvalues of H_k accumulate only at $+\infty$. Therefore, we may order them in an increasing way, and each H_k has an orthonormal basis of eigenvectors $(\psi_{k,n})_{n \in \mathbb{N}}$, such that $H_k(\psi_{k,n}) = E_{k,n} \cdot \psi_{k,n}$, $E_{k,n} \le E_{k,n+1}$ for all $n \in \mathbb{N}$, and $\lim_{n\to\infty} E_{k,n} = +\infty$. The norm continuity of the resolvents $(\lambda - H_k)^{-1}$ implies that the functions $E_n \colon \mathbb{T}^d \to \mathbb{R}$, $k \mapsto E_{k,n}$, are continuous (we prove this in Section 3.2). Since $\mathbb{T}^d$ is compact and connected, the range of the continuous function E_n is a compact interval $[b_n, c_n]$ in $\mathbb{R}$. Lemma 3.4 implies

$$\sigma(H) = \{E_{k,n} : k \in \mathbb{T}^d,\ n \in \mathbb{N}\} = \bigcup_{n \in \mathbb{N}} [b_n, c_n].$$

We have arranged $E_{k,n} \le E_{k,n+1}$ for $n \in \mathbb{N}$. This implies $b_n \le b_{n+1}$ and $c_n \le c_{n+1}$ for $n \in \mathbb{N}$. If $c_n \ge b_{n+1}$, then $[b_n, c_n] \cup [b_{n+1}, c_{n+1}] = [b_n, c_{n+1}]$. Continuing like this, we may merge overlapping intervals until we reach a gap where $c_n < b_{n+1}$. In this way, we write $\sigma(H)$ as a disjoint union of intervals. It may happen that infinitely many of the intervals overlap and give an unbounded interval $[d, \infty)$. In fact, physicists expect this to happen for large n. So we would expect that the spectrum of H is the disjoint union of finitely many (possibly zero) compact disjoint intervals $[d_n, e_n]$, $n = 0, \dots, N-1$, and

an unbounded interval $[d_N, \infty)$, with $d_n \le e_n < d_{n+1}$ for all $0 \le n < N$; from what we have proven, an infinite sequence of compact intervals, that is, the case $N = \infty$, is also possible. The disjoint intervals $[d_n, e_n]$ and $[d_N, \infty)$ (or maybe also the intervals $[b_n, c_n]$) are called *energy bands* of the system. The open intervals in between, (e_n, d_{n+1}) for $0 \le n < N$, are called *band gaps*.

So the spectrum of a Schrödinger operator with a bounded periodic potential is a union of energy bands with band gaps in between. Now we recall the Fermi energy μ. At temperature 0, all states of energy below the Fermi energy are filled with electrons. The energy bands $[d_n, e_n]$ with $e_n \le \mu$ are called *valence bands*. They are filled with electrons. Energy bands with $e_n > \mu$ are called conduction bands. Now we distinguish two cases.

Assume first that $\mu \in [d_n, e_n)$ for some n. Then some but not all of the states in $[d_n, e_n]$ are filled with electrons. We need not much energy to lift one of the electrons above the Fermi energy into an unoccupied state of the form $\int w(k) \exp(2\pi i k \cdot x) \psi_{k,M}(x)$, where $w(k)$ is a continuous function supported in a region where the eigenvalue $E_{k,M}$ of $\psi_{k,M}$ belongs to (μ, e_n). The function $\exp(2\pi i k \cdot x)$ describes a wave travelling with constant momentum. The factor $\psi_{k,M}(x)$ modifies this by some $\mathbb{Z}^d$-periodic pattern. Macroscopically, this function is essentially constant because the crystal unit is so small. So it does not affect the transport properties of the state. The factor w mixes waves of different momentum and is needed to produce a square-integrable function. So it does not cost much energy to excite an electron into a state in which it can move freely through the material. Thus our material conducts electricity. We have a *conductor* or a *metal*. The argument also justifies calling an energy band with $e_n > \mu$ a *conduction band*.

Assume next that μ lies in a band gap, $e_n < \mu < d_{n+1}$. Then all the filled states have energy at most e_n, and the energy $d_{n+1} - e_n$ is needed to lift an electron into one of the unoccupied energy bands. Unless the band gap is very small, this makes it impossible for electrons to reach a conduction band. So the material is an *insulator*. In the following, we are mostly interested in insulators, that is, in the case when the Fermi energy lies in a band gap.

Remark 3.5. In the discussion above, the merging of the intervals $[b_n, c_n]$ is very natural. In a more careful analysis, it may be better to keep these intervals separate. Assume that the spectrum of H contains two intervals $[b_n, \mu + \varepsilon_1]$ and $[\mu - \varepsilon_2, c_{n+1}]$, where μ is the Fermi energy and $\varepsilon_1, \varepsilon_2 > 0$ are very small. Then there are both filled and unfilled states near the Fermi energy. Nevertheless, the material is only a semimetal in this case. Roughly speaking, in a semimetal, there are only very few electron states whose energy is just below the Fermi energy, and there are only very few conduction states with energy just above the Fermi energy. So the material conducts electricity poorly.

To understand the behaviour of materials when the Fermi energy lies near a band gap, we must study finite-volume approximations and allow for small disorder (say, a small, random perturbation of the potential). A finite-volume approximation is just a finite matrix, so it has a finite set of eigenvalues. When the volume gets large, these eigenvalues will be close to the spectrum of the periodic Hamiltonian, and any point in the latter spectrum is close to some eigenvalue. The distribution of these discrete eigenvalues in the spectrum is not uniform, however, if we also add some small disorder

to the system. The "Lifschitz tail" phenomenon says that the density of the states of a disordered, finite-volume approximation near a band edge of the periodic Hamiltonian is very small.

3.2 Continuity of eigenfunctions

We continue to study a Hamiltonian affiliated with $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$, such as a Schrödinger operator with periodic, bounded potential as in Theorem 3.3. We claimed that the eigenvalues $E_{n,k}$ depend continuously on k. We are going to prove this, and we also study the continuity of the eigenfunctions $\psi_{n,k}$. Actually, continuity of $\psi_{n,k}$ is not the right question. The eigenfunction is unique up to scalar multiple if and only if the eigenvalue has multiplicity 1. So we can only expect the images of the eigenfunctions $\psi_{n,k}$ in the projective space of $L^2(\mathbb{R}^d)$ to be continuous, and this can only happen for those k where the eigenvalue $E_{n,k}$ has multiplicity 1. This statement is indeed true. But we aim at a more elegant statement. First, mapping $\psi_{n,k}$ to projective space has the same effect as taking the rank-1 projection $|\psi_{n,k}\rangle\langle\psi_{n,k}|$. Secondly, this is equal to the spectral projection of H_k for the singleton $\{E_{n,k}\}$ provided the eigenspace has multiplicity 1. So we should rather ask about continuity of spectral projections. The following lemma asserts this in the appropriate generality:

Lemma 3.6. *Let H be described by a norm-continuous family $(H_k)_{k\in\mathbb{T}^d}$ of self-adjoint operators with compact resolvent as in Theorem 3.3. Let $a, b \in \mathbb{R}$ satisfy $a < b$ and let $k_0 \in \mathbb{T}^d$ be such that $a, b \notin \sigma(H_{k_0})$. Let $\chi_{(a,b)}(H_k)$ be the spectral projection of H_k for the interval (a, b) for $k \in \mathbb{T}^d$. There is an open neighbourhood $U \subseteq \mathbb{T}^d$ of k_0 such that $a, b \notin \sigma(H_k)$ for all $k \in U$. And then the map $k \mapsto \chi_{(a,b)}(H_k)$ is norm-continuous on U, and its image consists of finite-rank projections.*

Proof. The operators $(\mathrm{i}+H_k)^{-1}$ form a norm-continuous function from U to the C*-algebra of compact operators on $L^2([0,1]^d)$. Let $x \coloneqq (\mathrm{i}+a)^{-1}$. Since $a \notin \sigma(H_{k_0})$, the operator $(\mathrm{i}+H_{k_0})^{-1} - x$ is invertible. Since the set of invertible operators on $\mathcal{K}$ is open, there is a neighbourhood U of k_0 so that $(\mathrm{i}+H_k)^{-1} - x$ is invertible for $k \in U$. This means that $a \notin \sigma(H_k)$ for $k \in U$. An analogous argument arranges $a, b \notin \sigma(H_k)$ for $k \in U$. Let $D \coloneqq \mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$. We may rewrite the functional calculus for H in terms of $(\mathrm{i}+H)^{-1} \in D$. Therefore, if $f\colon \mathbb{R} \to \mathbb{R}$ is a Borel function that is continuous on the spectrum of $H|_U$ and vanishes at ∞, then $f(H) \in D$. The characteristic function of (a, b) fulfils this because it is only discontinuous at a, b, and these two points do not belong to the spectrum of $H|_U$ by assumption. So $\chi_{(a,b)}(H_k)$ is norm-continuous in k and compact. Since these operators are projections, being compact means that the projections have finite rank. □

If the eigenvalue $E_{n,k}$ has multiplicity 1, then we may choose $E_{n-1,k} < a < E_{n,k} < b < E_{n+1,k}$, so that the projection $\chi_{(a,b)}(H_k)$ has rank 1. Then Lemma 3.6 gives an open neighbourhood U of k such that $k' \mapsto \chi_{(a,b)}(H_{k'})$ is continuous on U. Since its rank is 1 at k, the rank is 1 in a neighbourhood of k. Shrinking U if necessary, we arrange that the rank is 1 in all of U. That is, H_k for $k \in U$ has exactly one eigenvalue in the

interval (a, b). The continuity of $\chi_{(a,b)}(H_{k'})$ says, in addition, that the eigenfunction for this eigenvalue may be chosen continuously.

Lemma 3.7. *Assume that H is affiliated with $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$ and bounded below. Then the eigenvalues $E_{n,k}$ defined above depend continuously on k.*

Proof. Since H is bounded below, $\chi_{(-\infty,b)}(H_k)$ has finite rank for all $b \in \mathbb{R}$. The argument above shows that this rank is locally constant at k if $b \notin \sigma(H_k)$. Now choose $a < E_{n,k} < b$ so that $|b - a| < \varepsilon$. We have seen above that there is a neighbourhood U of k so that $a, b \notin \sigma(H_{k'})$ for $k' \in U$. And the ranks of $\chi_{(-\infty,b)}(H_{k'})$ and $\chi_{(-\infty,a)}(H_{k'})$ are constant in U. It follows that the nth eigenvalue of $H_{k'}$ satisfies $a < E_{n,k'} < b$ for all $k' \in U$. Since $|b - a| < \varepsilon$, this expresses the continuity of $E_{n,k}$ as a function in k. □

If the spectrum of H is not bounded below, then it may be impossible to order the eigenvalues continuously throughout $\mathbb{T}^d$. This is related to the spectral flow:

Example 3.8. Let $d = 1$ and identify $\mathbb{T}^1 = [0, 1]/\sim$ with $0 \sim 1$. We may choose H so that the spectrum of H_k is $k + \mathbb{Z}$ for all $k \in [0, 1]$. This is the same set for $k = 0$ and $k = 1$. For $k = 0$, it is natural to take $E_{n,0} = n$. The only way to label the eigenvalues continuously for $k > 0$ is by setting $E_{n,k} \coloneqq n + k$. But at $k = 1$, this gives $E_{n,1} = n + 1 \neq n$. So we do not get a consistent labelling of the eigenvalues throughout $\mathbb{T}$. The problem is that given any threshold $\lambda \in \mathbb{R} \setminus \mathbb{Z}$, exactly one eigenvalue will cross it when we move along the loop H_k for $k \in [0, 1]$. So the spectral flow is 1. This prevents us from labelling the eigenvalues throughout $\mathbb{T}$.

A consistent labelling of the eigenvalues exactly as above is possible if H has a spectral gap, say, $t \notin \sigma(H)$. Then we may let $E_{n,k}$ for $n \geq 0$ and $n < 0$ be the eigenvalues above t in increasing order and below t in decreasing order, respectively. The same argument as above shows that these are continuous functions. Problems may only occur if $\sigma(H) = \mathbb{R}$. Then we may define continuous functions $E_{n,\sqcup} \colon \mathbb{R}^d \to \mathbb{R}$ for $n \in \mathbb{Z}$ that label the eigenvalues of H_k in increasing order. The spectral flow in the direction $x \in \mathbb{Z}^d$ is defined as $E_{n,x} - E_{n,0}$; this is the same for all $n \in \mathbb{Z}$. There is a consistent definition of $E_{n,k}$ for $k \in \mathbb{T}^d$ if and only if $E_{n,x} - E_{n,0} = 0$ for all $x \in \mathbb{Z}^d$. Actually, the case where $\sigma(H) = \mathbb{R}$ gives us only conductors, so it is irrelevant for the study of topological insulators.

The most interesting object for topological phases is the projection $\chi_{(-\infty,\mu)}(H)$ for the Fermi energy μ. This is the projection onto the filled bands, which describes the ground state of our many-particle system. The projection $\chi_{(-\infty,\mu)}(H)$ is described by the projection-valued function $k \mapsto \chi_{(-\infty,\mu)}(H_k)$. If the operators H_k are bounded below and have compact resolvent, then the projections $\chi_{(-\infty,\mu)}(H_k)$ have finite rank. If μ lies in a spectral gap of H, then these projections depend continuously on k.

3.3 Tight binding models

The one-particle approximation treated above is only an approximation. We made the Ansatz that the effective Hamiltonian in the one-particle model is a Schrödinger operator with a bounded, periodic potential. This assumption seems unjustified because spin-orbit

coupling, the source of the known topological insulators, is a relativistic effect. Band theory only needs that the Hamiltonian is affiliated with $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$ for some Hilbert space $\mathcal{K}$. Many articles in physics describe concrete materials by models where $\mathcal{K}$ has finite dimension. In fact, the dimension of $\mathcal{K}$ is quite small – such as 2 or 4 – in most concrete models. An electron has infinitely many degrees of freedom in each cell of the crystal. But since the Hamiltonian has compact resolvent, there are only finitely many states in each unit cell whose energy lies below a given bound. We do not lose important information when we restrict to such a finite-dimensional subspace. The tight-binding approximation is an Ansatz to choose such a subspace and a corresponding Hamiltonian. The Ansatz only has a finite number of degrees of freedom, and there are algorithms to predict these from theory. Actually, these algorithms have variants that differ in which physical effects they take into account, such as spin-orbit coupling. In the hands of the experts, some of these algorithms predict the Hamiltonian well enough to detect whether a given material is an insulator and what its topological phase is.

In a tight-binding model, we first consider a single atom in our material and compute some of its eigenfunctions. Let a unit cell of our crystal have atoms of type i_1 at position $y_i \in \mathbb{R}^d$, $1 \le i \le k$, and let $\varphi_{t_i,n}$ for $n \in \mathbb{N}$ be eigenfunctions of the Schrödinger operator for an atom of type t_i at 0. Choose a finite set of indices (y_i, t_i, n_i), $1 \le i \le N$, and let $\mathcal{K}$ be the N-dimensional subspace spanned by the functions $\varphi_{t_i,n_i}(x - y_i)$. We think of $\varphi_{t_i,n_i}(x - y_i)$ as describing an electron tightly bound to the atom at y_i in the orbital n_i. The other atoms in the crystal and their electrons also affect this electron. We include this in our Hamiltonian by hopping terms $t_{i,j,R} = \langle \varphi_{t_i,n_i}(x - y_i) \,|\, H(\varphi_{t_j,n_j}(x - y_j - R)) \rangle$, where $R \in \mathbb{Z}^d$ is a vector in the translation lattice of the crystal. In practice, we only use a small number of orbitals to get a low-dimensional space $\mathcal{K}$, and we only consider hopping terms for atoms that are close to each other; often, we only consider the interaction between nearest neighbours or perhaps next-to-nearest neighbours. Thus we put $t_{i,j,R} = 0$ for $R \gg 0$.

Now we work in the Hilbert space $\mathcal{H} = \ell^2(\mathbb{Z}^d, \mathbb{C}^N)$ and use the translation-invariant Hamiltonian H' on $\mathcal{H}$ that has the matrix coefficients $t_{i,j,R-S}$ in the standard basis of $\mathcal{H}$. This is the best approximation to H on $\ell^2(\mathbb{Z}^d, \mathbb{C}^N)$ because

$$\langle \delta_S \otimes e_i \,|\, H'(\delta_R \otimes e_j) \rangle = t_{i,j,R-S} = \langle \varphi_{t_i,n_i}(x - y_i - S) \,|\, H(\varphi_{t_j,n_j}(x - y_j - R) \rangle.$$

Since $t_{i,j,R} = 0$ for $R \gg 0$, the operator H' is a finite sum of operators of the form $S_R \otimes H'_R$ for $R \in \mathbb{Z}^d$, where $H'_R \in \mathbb{M}_N(\mathbb{C})$ has matrix coefficients $t_{i,j,-R}$. Now we identify the Hilbert space $\ell^2(\mathbb{Z}^d, \mathbb{C}^N)$, with $L^2(\mathbb{T}^d, \mathbb{C}^N)$ by the Fourier transform. Then the operator S_R becomes pointwise multiplication with $\exp(-2\pi \mathrm{i} R \cdot k)$. So H'_R is unitarily equivalent to the operator on $L^2(\mathbb{T}^d, \mathbb{C}^N)$ of the form $(\hat{H}\psi)(k) = \hat{H}(k)\psi(k)$ for $k \in \mathbb{T}^d$, $\psi \in L^2(\mathbb{T}^d, \mathbb{C}^N)$, with

$$\hat{H} \colon \mathbb{T}^d \to \mathbb{M}_N(\mathbb{C}), \qquad t \mapsto \sum_{R \in \mathbb{Z}^d} \exp(2\pi \mathrm{i} t \cdot R) H'_R.$$

So $\hat{H}$ is a matrix-valued trigonometric polynomial. So $\hat{H} \in \mathrm{C}(\mathbb{T}^d, \mathbb{M}_N(\mathbb{C}))$. In particular, H' is bounded, and the resulting operators $\hat{H}_k$ for $k \in \mathbb{T}^d$ have exactly N eigenvalues. So there is no need to pass to resolvents in this case to discuss continuity or self-adjointness.

Many articles in physics and materials science do not consider a "continuum model" on $L^2(\mathbb{R}^d)$, but start directly with a Hamiltonian acting on $\ell^2(\mathbb{Z}^d, \mathbb{C}^N)$ We now consider some examples.

3.3.1 Graphene

Graphene is a single layer of carbon atoms arranged in a hexagonal lattice (see Figure 1). Thus it is a 2-dimensional material. It was first produced and observed by Hanns-Peter

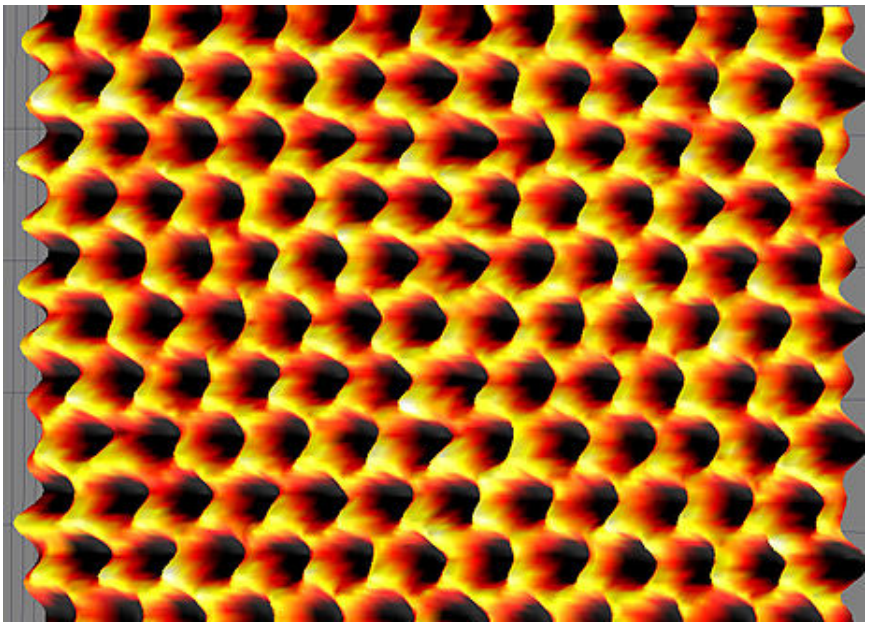


Figure 1: Scanning probe miscroscopy image of the graphene lattice taken from Wikipedia

Boehm in 1962 and received the name "graphene" in 1986. Let $\Lambda_0 \subseteq \mathbb{R}^2$ be the set of vertices of the hexagon lattice of graphene. We shall derive a tight-binding model for graphene based on the symmetries of Λ_0. The filled and empty circles in Figure 2 are the points in Λ_0, and the gray lines between them show the hexagon lattice. We we choose the coordinates in $\mathbb{R}^2$ so that two neighbours in Λ_0 are at the positions $(0,0)$ and $(1,0)$. Thus the length scale is normalised so that neighbours in Λ_0 have distance 1. Each vertex in Λ_0 has three neighbours. In our coordinate system, the three neighbours of $(0,0) \in \Lambda_0$ are

$$a_1 \coloneqq (1,0), \qquad a_2 \coloneqq \frac{1}{2}(-1,\sqrt{3}), \qquad a_3 \coloneqq \frac{1}{2}(-1,-\sqrt{3}).$$

These points of Λ_0 have six neighbours besides $(0,0)$ at $\pm b_i$ for $i = 1,2,3$, where

$$b_1 \coloneqq a_2 - a_3 = (0,\sqrt{3}), \qquad b_2 \coloneqq a_3 - a_1 = \frac{1}{2}(-3,-\sqrt{3}), \qquad b_3 \coloneqq a_1 - a_2 = \frac{1}{2}(3,-\sqrt{3}).$$

Notice that $a_1 + a_2 + a_3 = 0$ and $b_1 + b_2 + b_3 = 0$. Let

$$\Lambda = \mathbb{Z}b_1 + \mathbb{Z}b_2 + \mathbb{Z}b_3 \subseteq \mathbb{R}^2.$$

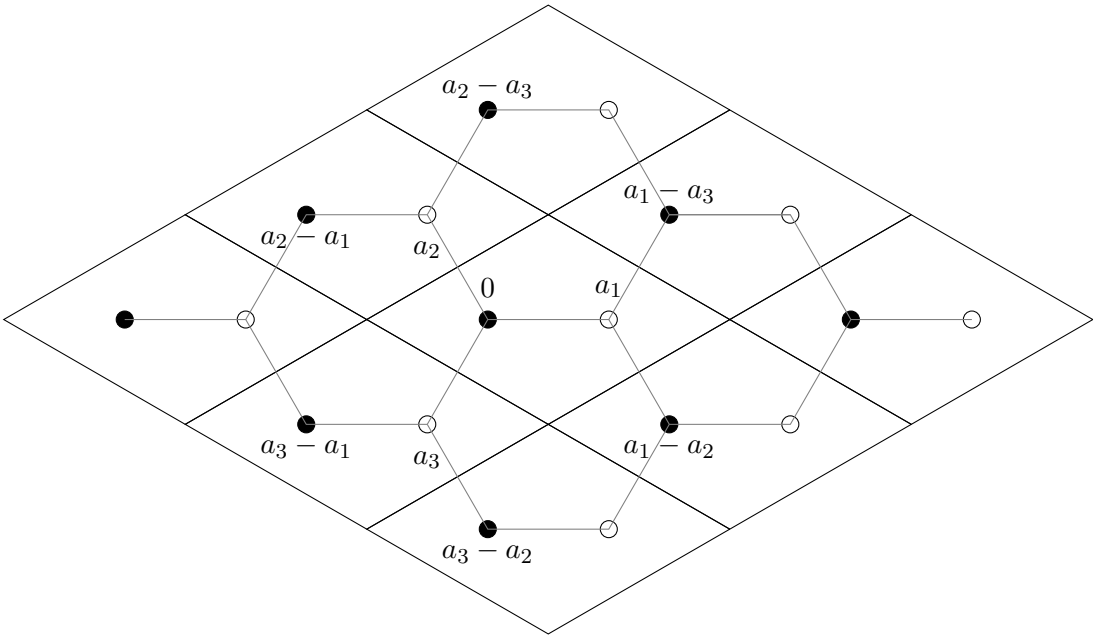


Figure 2: The graphene lattice in gray and the unit cells of the crystal in black.

Actually, any two of b_1, b_2, b_3 already generate this lattice because $b_1 + b_2 + b_3 = 0$. Each hexagon in our lattice has six neighbours, and the distance between their midpoints are $\pm b_i$ for $i = 1, 2, 3$. Hence the set Λ_0 is invariant under translations by $\pm b_j$. This implies that Λ_0 is invariant under translations by Λ, that is, $\Lambda_0 + \Lambda = \Lambda_0$.

It is clear that any of the hexagons in our lattice is a possible unit cell for the lattice Λ. We prefer a diamond unit cell as drawn in Figure 2. Namely, let $C \subseteq \mathbb{R}^2$ be the set of all $x \in \mathbb{R}^2$ such that the point in Λ_0 closest to x is $(0, 0)$ or $(1, 0)$. This subset C is one of the diamonds in Figure 2. Its translate $x + C$ for $x \in \Lambda$ tesselate the plane. If $x \in \Lambda$, then $(x + C) \cap \Lambda_0 = \{x, x + a_1\}$ has exactly two points – much fewer than a hexagon. This is why we prefer the diamonds. We see also that Λ_0 is a disjoint union of two cosets of the lattice Λ:

$$\Lambda_0 = \Lambda \sqcup (\Lambda + a_1).$$

In Figure 2, the points in Λ and $\Lambda + a_1$ are marked by full circles and empty circles, respectively. We briefly say that points in Λ and $\Lambda + a_1$ have types A and B, respectively. Nearest neighbours in Λ_0 always have different type. The midpoints of the hexagons in the lattice are at the points of $\Lambda - a_1$, which is disjoint from Λ_0.

The subset Λ_0 itself is not a subgroup of $\mathbb{R}^2$. For instance, $a_2 + a_3 = -a_1 \notin \Lambda_0$ although $a_2, a_3 \in \Lambda_0$. When we cut the diamonds in Figure 2 into two triangles by the vertical diagonal, we get a tesselation of the plane by equilateral triangles. These triangles are the "Voronoi cells" of the set Λ_0. These triangles, however, fail to be the "cells" of a lattice of translations. This cannot be because the triangles centred at the empty and full circles have different directions in the plane.

Let G be the symmetry group of Λ_0. It consists of all isometric maps $\Lambda_0 \to \Lambda_0$ or, equivalently, of all (affine) isometries $\mathbb{R}^2 \to \mathbb{R}^2$ that map Λ_0 onto itself.

Proposition 3.9. *The group G is generated by the translations T_x for $x \in \Lambda$ and by the 12-element dihedral group D_6, acting such that the origin is the midpoint $-a_1$ of one of the hexagons. The group G acts transitively on Λ_0, on the pairs $(x, y) \in \Lambda_0^2$ with $\|x - y\| = 1$, and on the pairs $(x, y) \in \Lambda_0^2$ with $\|x - y\| = \sqrt{3}$.*

Proof. The dihedral group D_6 consists of the six rotations around $-a_1$ by the angles $2\pi k/6$, $k \in \mathbb{Z}/6$, and the reflections at the six lines that go through $-a_1$ and either a vertex or the midpoint of an edge of the hexagon centred at $-a_1$. These isometries map Λ_0 onto itself. So do the translations by $x \in \Lambda$. Hence products $T_x h$ belong to G. We claim that any element $g \in G$ is a product $T_x h$ for some $x \in \Lambda$ and some $h \in D_6$. First, g maps $-a_1$ to the midpoint of some other hexagon: $g(-a_1) = x - a_1$ for some $x \in \Lambda$. Thus $T_{-x} g$ fixes $-a_1$. In the shifted coordinate system with origin at $-a_1$, $T_{-x} g$ is a linear isometry. Since it fixes $-a_1$ and maps Λ_0 onto itself, it must permute the six vertices of the hexagon around $-a_1$. Hence it is an element $h \in D_6$. So $g = T_x h$.

Our proof already shows that the group G acts transitively on Λ_0. The subgroup of G that fixes $0 \in \Lambda_0$ is the dihedral group D_3. It acts transitively on the three neighbours a_1, a_2, a_3 of 0 and also on the six next-to-nearest neighbours $\pm b_1, \pm b_2, \pm b_3$ of 0. This implies the transitivity claims in the proposition. □

Proposition 3.9 shows that a translation T_x belongs to G if and only if $x \in \Lambda$. That is, Λ is the largest group of translations that preserves Λ_0.

Now we build a tight-binding model for graphene. We choose to consider only one orbital or electron state for each carbon atom. Thus the Hilbert space of our model is $\ell^2(\Lambda_0)$, where the basis vector δ_x for $x \in \Lambda_0$ means a suitable orbital around the atom at position x, describing an electron state tightly bound to that atom. The isometry group G of Λ_0 acts on $\ell^2(\Lambda_0)$ by permuting the basis vectors δ_x for $x \in \Lambda_0$. We assume that our Hamiltonian commutes with this representation of G.

Since $\Lambda_0 = \Lambda \sqcup (\Lambda + a_1)$, we may identify $\ell^2(\Lambda_0)$ with $\ell^2(\Lambda, \mathbb{C}^2)$, where the Dirac distributions δ_x and δ_{x+a_1} for $x \in \Lambda_0$ become $\delta_x \otimes e_1$ and $\delta_x \otimes e_2$ in $\ell^2(\Lambda, \mathbb{C}^2)$, respectively. Choosing a lattice basis for Λ, we may further identify $\Lambda \cong \mathbb{Z}^2$ as a group and arrive at the Hilbert space $\ell^2(\mathbb{Z}^2, \mathbb{C}^2)$. We have worked in the latter Hilbert space when we developed the general theory. Since this distorts the geometry of the hexagon lattice, we now stay in the Hilbert spaces $\ell^2(\Lambda_0)$ or $\ell^2(\Lambda, \mathbb{C}^2)$.

We describe H through a matrix $(H_{x,y})_{x,y \in \Lambda_0}$. If $x, y \in \Lambda_0$, then $H_{x,y} = \overline{H_{y,x}}$ because H is self-adjoint, and $H_{g(x),g(y)} = H_{x,y}$ for $g \in G$ because H is G-equivariant. In addition, we assume that there is $R > 0$ so that $H_{x,y} = 0$ for $\|x - y\| > R$. In fact, we assume $H_{x,y} = 0$ if $\|x - y\| > \|b_j\| = \sqrt{3}$, that is, we only allow nearest and next-to-nearest neighbour hopping terms.

Since G acts transitively on Λ_0 and on the sets of pairs $(x, y) \in \Lambda_0^2$ with $\|x - y\| = 1$ and $\|x - y\| = \sqrt{3}$, the symmetry condition of H implies that there are constants E, t, z such that $H_{x,x} = E$ for all $x \in \Lambda_0$, $H_{x,y} = t$ for all $x, y \in \Lambda_0$ with $\|x - y\| = 1$, and $H_{x,y} = z$ for all $x, y \in \Lambda_0$ with $\|x - y\| = \sqrt{3}$. These constants E, t, z must be real in

order for H to be self-adjoint. If we allowed hopping to third neighbours, the possible Hamiltonians would get more complicated because the group G does not act transitively on such pairs any more. Let $H_{E,t,z} \in \mathbb{B}(\ell^2(\Lambda_0))$ denote the Hamiltonian built above for given $E, t, z \in \mathbb{R}$. We write H if there is no need to mention these constants.

To apply our general theory, we transfer H first to a self-adjoint operator H' on the Hilbert space $\ell^2(\Lambda, \mathbb{C}^2)$, and then identify the latter with $L^2(\mathbb{T}^2, \mathbb{C}^2)$ by a suitable Fourier transform. We describe H' through a 2×2-matrix of operators $(H'_{i,j})_{i,j=1,2}$ on $\ell^2(\Lambda)$:

Lemma 3.10. *The Hamiltonian on $\ell^2(\Lambda_0)$ corresponds to the following operator on $\ell^2(\Lambda, \mathbb{C}^2)$:*

$$\begin{pmatrix} E + \sum_{j=1}^3 z(S_{b_j} + S_{-b_j}) & t(S_0 + S_{-b_2} + S_{b_3}) \\ t(S_0 + S_{b_2} + S_{-b_3}) & E + \sum_{j=1}^3 z(S_{b_j} + S_{-b_j}) \end{pmatrix}.$$

Proof. Since H is Λ-invariant, each matrix entry $H'_{i,j}$ is a linear combination of the translation operators S_x for $x \in \Lambda$. The diagonal terms $H_{x,x} = E$ in H correspond to $E := E \cdot \mathrm{id}_{\ell^2(\Lambda)}$ in both diagonal entries of H'. The nearest neighbour hopping terms $H_{x,y} = t$ for $\|x - y\| = 1$ correspond to entries in the off-diagonal entries of H' because nearest neighbours have different type. Figure 2 shows that the three nearest neighbours of an atom in Λ_0 of type A are the atoms of type B in the same cell and in the cells translated by $-b_3 = a_2 - a_1$ and $b_2 = a_3 - a_1$, while the three nearest neighbours of an atom of type B are the atoms of type A in the same cell and in the cells translated by $+b_3$ and $-b_2$. So we get the terms $t \cdot (S_0 + S_{-b_2} + S_{b_3})$ in H'_{12} and $t \cdot (S_0 + S_{b_2} + S_{-b_3}) = t \cdot (S_0 + S_{-b_2} + S_{b_3})^*$ in H'_{21}. The next-to-nearest neighbours of an atom again have the same type, so that the terms $H_{x,y}$ for $(x, y) \in \Lambda_0^2$ with $\|x - y\| = \sqrt{3}$ contribute to the diagonal terms in H'. Since the relative positions of the six next-to-nearest neighbours of any vertex are $\pm b_j$ for $j = 1, 2, 3$, the contribution is $z \cdot \sum_{j=1}^3 (S_{b_j} + S_{-b_j})$. This gives the formula asserted in the lemma. □

Next, we turn H' into an operator of multiplication with a 2×2-matrix-valued function on $L^2(\mathbb{T}^2, \mathbb{C}^2)$. Since we did not normalise the translation lattice Λ, the relevant torus is the quotient $\mathbb{R}^2/\Lambda^*$ for the dual lattice

$$\Lambda^* = \{k \in \mathbb{R}^2 : k \cdot b_1, k \cdot b_2 \in \mathbb{Z}\} = \{k \in \mathbb{R}^2 : k \cdot (-b_2), k \cdot b_3 \in \mathbb{Z}\}.$$

We choose the sign in the Fourier transform so that the operator S_λ on $\ell^2(\Lambda)$ for $\lambda \in \Lambda$ corresponds to the operator of pointwise multiplication by $\exp(2\pi \mathrm{i} k \cdot \lambda)$ on $L^2(\mathbb{R}^2/\Lambda^*)$. With this convention, the Hamiltonian H' described in Lemma 3.10 corresponds to the matrix-valued function

$$H \colon \mathbb{R}^2/\Lambda^* \to \mathbb{M}_2(\mathbb{C}), \qquad k \mapsto$$
$$\begin{pmatrix} E + \sum_{j=1}^3 2z \cos(2\pi k \cdot b_j) & t(1 + \exp(-2\pi \mathrm{i} k \cdot b_2) + \exp(2\pi \mathrm{i} k \cdot b_3)) \\ t(1 + \exp(2\pi \mathrm{i} k \cdot b_2) + \exp(-2\pi \mathrm{i} k \cdot b_3)) & E + \sum_{j=1}^3 2z \cos(2\pi k \cdot b_j) \end{pmatrix}.$$

To get simple formulas for the off-diagonal terms, we express k in the dual basis of $-b_2, b_3$. We claim that it consists of the vectors

$$-\frac{2}{3} a_3 = \frac{1}{3}(1, -\sqrt{3}), \qquad -\frac{2}{3} a_2 = \frac{1}{3}(1, +\sqrt{3}).$$

Write $k \in \mathbb{R}^2$ as $k = -\frac{2}{3}a_3 x - \frac{2}{3}a_2 y$ with $x, y \in \mathbb{R}$. The claim says that

$$k \cdot (-b_2) = x, \qquad k \cdot b_3 = y,$$

which is verified directly. This implies $k \cdot b_1 = x - y$ because $b_1 = -b_2 - b_3$. So we may rewrite the diagonal entries of $H(k) = H(x, y)$ as

$$H_{11}(x, y) \coloneqq E + 2z(\cos(2\pi(x - y)) + \cos(2\pi x) + \cos(2\pi y)))$$

and the off-diagonal entries as

$$H_{12} = t(1 + \exp(2\pi \mathrm{i} x) + \exp(2\pi \mathrm{i} y)), \qquad H_{21} = \overline{H_{12}}.$$

This function is manifestly $\mathbb{Z}^2$-periodic.

To simplify the computation of the spectrum, we express the operator $H_{E,t,z}$ studied above by the simpler operator $K \coloneqq H_{0,1,0}$. This has $H_{x,x} = 0$ and $H_{x,y} = 0$ for $\|x - y\| = \sqrt{3}$ and $H_{x,y} = 1$ for $\|x - y\| = 1$. The (x, y)-matrix entry of K^2 is the number of paths x, w, y in the hexagon, that is, the number of $w \in \Lambda_0$ with $\|x - w\| = \|w - y\| = 1$. There is exactly one such path if x, y are next-to-nearest neighbours, there are three such paths if $x = y$, and there is no such path otherwise. Thus $K^2 = H_{3,0,1}$. Since $H_{1,0,0}$ is the identity operator and $H_{E,t,z}$ depends linearly on E, t, z, we may write

$$H_{E,t,z} = z \cdot K^2 + t \cdot K^1 + (E - 3z) \cdot K^0.$$

So we get the spectrum of $H_{E,t,z}$ by applying the quadratic function $h \mapsto zh^2 + th + E - 3z$ to the spectrum of K. Now K corresponds to the function

$$K(x, y) = \begin{pmatrix} 0 & 1 + \exp(2\pi \mathrm{i} x) + \exp(2\pi \mathrm{i} y) \\ 1 + \exp(-2\pi \mathrm{i} x) + \exp(-2\pi \mathrm{i} y) & 0. \end{pmatrix}$$

For given $x, y \in \mathbb{R}/\mathbb{Z}$, the two eigenvalues of this 2×2-matrix are

$$\pm|1 + \exp(2\pi \mathrm{i} x) + \exp(2\pi \mathrm{i} y)|.$$

This vanishes if and only if the three points 1, $\exp(2\pi \mathrm{i} x)$ and $\exp(2\pi \mathrm{i} y)$ on the unit circle are equally spaced, that is, the three third roots of unity; this corresponds to the two points $(x, y) = (1/3, 2/3)$ and $(x, y) = (2/3, 1/3)$. And the absolute value above is maximal and equal to 3 if $x = y = 0$. So K has the spectrum $[-3, 3]$. And the spectrum of $H_{E,t,z}$ is $\{zh^2 + th + E - 3z : h \in [-3, 3]\}$, which is always an interval. so there are no gaps in the spectrum. Graphene is, in fact, described in first approximation by the simplest Hamiltonian $K = H_{0,1,0}$ in the family. Since 0 is in its spectrum, it describes a conductor.

The operator K has an obvious chiral symmetry, namely,

$$\Xi = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

This symmetry is destroyed, however, when $E \neq 0$ or $z \neq 0$. So it is not a fundamental symmetry of our model. It explains nicely, however, why the spectrum of K is symmetric. Since the matrix coefficients of $H_{E,t,z}$ are all real, it commutes with complex conjugation on $\ell^2(\Lambda_0)$. This gives a time-reversal symmetry Θ with square $+1$, which commutes with the symmetry group G as well.

3.3.2 Allowing different atoms at sites A and B

Now we generalise the construction above. We still consider a hexagon lattice with one orbital for each atom, but we allow two different atom types at the sites of type A and B. So we still consider a self-adjoint Hamiltonian H on the Hilbert space $\ell^2(\Lambda_0)$. But now we only ask H to be invariant under certain subgroups of G.

First, we assume H to be invariant under the subgroup G_2 of G that consists of those isometries of Λ_0 that preserve the decomposition $\Lambda_0 = \Lambda \sqcup (\Lambda + a_1)$. This still contains the same translation group Λ, but the group D_6 in G is now replaced by the subgroup D_3 of index 2. By construction, the action of G_2 on Λ_0 has two orbits Λ and $(\Lambda + a_1)$. The group G acts transitively on pairs $(x, y) \in \Lambda_0$ with $\|x - y\| = 1$ and on pairs with $\|x - y\| = \sqrt{3}$ by Proposition 3.9. Therefore, the subgroup G_2 of index 2 has at most two orbits on both types of pairs. There are also at least two orbits. So the action on pairs $(x, y) \in \Lambda_0$ with $\|x - y\| = 1$ has two orbits, namely, the pairs with $x, y - a_1 \in \Lambda$ and those with $x - a_1, y \in \Lambda$. And the action on $(x, y) \in \Lambda_0$ with $\|x - y\| = \sqrt{3}$ has two orbits as well, namely, those pairs with $x, y \in \Lambda$ and those with $x, y \in \Lambda + a_1$.

As a result, there is now more flexibility for our Hamiltonian. Namely, there must be $E_A, E_B \in \mathbb{R}$ such that $H_{x,x} = E_A$ for $x \in \Lambda$ and $H_{x,x} = E_B$ for $x \in \Lambda + a_1$. There must be $t \in \mathbb{C}$ with $H_{x,y} = t$, $H_{y,x} = \bar{t}$ for $x \in \Lambda$, $y \in \Lambda + a_1$ with $\|x - y\| = 1$. And there must be $z_A, z_B \in \mathbb{R}$ with $H_{x,y} = z_A$ for $x, y \in \Lambda$ with $\|x - y\| = \sqrt{3}$, and $H_{x,y} = z_B$ for $x, y \in \Lambda + a_1$ with $\|x - y\| = \sqrt{3}$. Here we have also used the constraint $H_{x,y} = \overline{H_{y,x}}$ for $x, y \in \Lambda_0$ to conclude that E_A, E_B, z_A, z_B are real and that the hopping terms for $\|x - y\| = 1$ are t and $\bar{t}$ for some $t \in \mathbb{C}$. The corresponding Hamiltonian on the Hilbert space $\ell^2(\Lambda, \mathbb{C}^2)$ becomes

$$H' = \begin{pmatrix} E_A + \sum_{j=1}^3 z_A(S_{b_j} + S_{-b_j}) & \bar{t}(S_0 + S_{-b_2} + S_{b_3}) \\ t(S_0 + S_{b_2} + S_{-b_3}) & E_B + \sum_{j=1}^3 z_B(S_{b_j} + S_{-b_j}) \end{pmatrix}.$$

This may be translated to a function on $\mathbb{R}^2/\Lambda^*$ as above.

We may use a coordinate transformation to reduce to the case of a real parameter t. The operator

$$U = \begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix}$$

for $\lambda \in \mathrm{U}(1)$ is unitary. We compute

$$UH'U^* = \begin{pmatrix} E_A + \sum_{j=1}^3 z_A(S_{b_j} + S_{-b_j}) & \overline{\lambda t}(S_0 + S_{-b_2} + S_{b_3}) \\ \lambda t(S_0 + S_{b_2} + S_{-b_3}) & E_B + \sum_{j=1}^3 z_B(S_{b_j} + S_{-b_j}) \end{pmatrix}.$$

Letting $\lambda = \bar{t}/|t|$, we see that H' is unitarily equivalent to the operator of the same form with $t\bar{t}/|t| = |t|$ instead of t. So it suffices to treat the case $t \in [0, \infty)$. Assuming this, we may express H' in terms of the operator $K = H_{0,1,0}$ and its chiral symmetry Ξ:

$$H' = \frac{E_A + E_B - 3(z_A + z_B)}{2} \cdot 1 + \frac{E_A - E_B - 3(z_A - z_B)}{2} \cdot \Xi \\ + t \cdot K + \frac{z_A + z_B}{2} K^2 + \frac{z_A - z_B}{2} K^2 \Xi.$$

Since K and Ξ anticommute, $K^2\Xi = -K\Xi K = \Xi K^2$.

We are particularly interested in the case where $E_A = -E_B$ and $z_A = z_B = 0$. Then the function on $\mathbb{R}^2/\Lambda^*$ corresponding to H' simplifies to

$$H(x,y) \coloneqq \begin{pmatrix} E_A & t(1+\exp(2\pi\mathrm{i}x)+\exp(2\pi\mathrm{i}y)) \\ t(1+\exp(-2\pi\mathrm{i}x)+\exp(-2\pi\mathrm{i}y)) & -E_A \end{pmatrix}.$$

This matrix has trace 0 and determinant $-E_A^2 - t^2|1+\exp(2\pi\mathrm{i}x)+\exp(2\pi\mathrm{i}y)|^2$. So its spectrum consists of the two points

$$\pm\sqrt{E_A^2 + t^2|1+\exp(2\pi\mathrm{i}x)+\exp(2\pi\mathrm{i}y)|^2}.$$

Letting x, y vary, the spectrum splits into two disjoint energy bands $\pm[E_A, \sqrt{E_A^2+3t^2}]$, and 0 is in a spectral gap. Assuming that the Fermi energy is at 0, we have now got an insulator. It will turn out, however, that this insulator is topologically trivial. To get something topologically non-trivial, we weaken our symmetry constraint further. Namely, we let $G_4 \subseteq G_2$ be the subgroup of orientation-preserving isometries in G_2. This has index 2 in G_2 and hence index 4 in G. The group G_4 is generated by the translations T_x for $x \in \Lambda$ and by the rotation around 0 with angle $2\pi/3$. We still consider only nearest and next-to-nearest neighbour hoppings. Now we find, first, that there are $E_A, E_B \in \mathbb{R}$ with

$$H_{x,x} = \begin{cases} E_A & \text{if } x \in \Lambda, \\ E_B & \text{if } x \in \Lambda + a_1; \end{cases}$$

secondly, that there is $t \in \mathbb{C}$ with

$$H_{x,y} \coloneqq \begin{cases} t & \text{if } x \in \Lambda, y-x \in \{a_1,a_2,a_3\}, \\ \bar{t} & \text{if } x \in \Lambda + a_1, x-y \in \{a_1,a_2,a_3\}; \end{cases}$$

third, that there are $z_A, z_B \in \mathbb{C}$ with

$$H_{x,y} \coloneqq \begin{cases} z_A & \text{if } x \in \Lambda, x-y \in \{b_1,b_2,b_3\}, \\ \overline{z_A} & \text{if } x \in \Lambda, y-x \in \{b_1,b_2,b_3\}, \\ z_B & \text{if } x \in \Lambda + a_1, x-y \in \{b_1,b_2,b_3\}, \\ \overline{z_B} & \text{if } x \in \Lambda + a_1, y-x \in \{b_1,b_2,b_3\}; \end{cases}$$

and $H_{x,y} = 0$ for all other $x, y \in \Lambda_0$, that is, if $\|x-y\| > \sqrt{3}$. Any choice of the constants E_A, E_B, t, z_A, z_B gives a self-adjoint operator on $\ell^2(\Lambda_0)$ that commutes with G_4. As above, this Hamiltonian is unitarily equivalent to the one for the parameters $E_A, E_B, |t|, z_A, z_B$. So we may assume without loss of generality that $t \geq 0$.

The operator H' becomes the operator on $L^2(\mathbb{T}^2, \mathbb{C}^2)$ that acts on a function of $(x,y) \in \mathbb{R}^2/\mathbb{Z}^2$ by pointwise application by $\hat{H}(x,y) \in \mathbb{M}_2(\mathbb{C})$ with

$$\begin{aligned}
\hat{H}_{11}(x,y) &= E_A + 2\,\mathrm{Re}\big(z_A \exp(2\pi\mathrm{i}(x-y)) + z_A \exp(-2\pi\mathrm{i}x) + z_A \exp(2\pi\mathrm{i}y)\big), \\
\hat{H}_{12}(x,y) &= t \cdot \big(1+\exp(2\pi\mathrm{i}x)+\exp(2\pi\mathrm{i}y)\big), \\
\hat{H}_{21}(x,y) &= t \cdot \big(1+\exp(-2\pi\mathrm{i}x)+\exp(-2\pi\mathrm{i}y)\big), \\
\hat{H}_{22}(x,y) &= E_B + 2\,\mathrm{Re}\big(z_B \exp(2\pi\mathrm{i}(x-y)) + z_B \exp(-2\pi\mathrm{i}x) + z_B \exp(2\pi\mathrm{i}y)\big).
\end{aligned}$$

3.3.3 Comparison to the literature

Some of the Hamiltonians above have already been considered by Semenoff [14] and Haldane [4]. More precisely, these authors describe the same Hamiltonians on the Hilbert space $\ell^2(\Lambda_0)$, but give different formulas for the corresponding functions $\mathbb{R}^2 \to \mathbb{M}_2(\mathbb{C})$. We now study this discrepancy.

Semenoff studies the generalisation of graphene with two atoms at the sites A and B, but considering only nearest neighbour hopping. He normalises the energy so that $E_A + E_B = 0$. This may be done by subtracting a constant from the Hamiltonian. He uses different notation and different names for the constants. In our notation, his Hamiltonian acting on $\ell^2(\Lambda_0)$ is the G_2-invariant Hamiltonian described above with $E_A + E_B = 0$, $t \in \mathbb{R}$, and $z_A = z_B = 0$. However, the corresponding function $H\colon \mathbb{R}^2 \to \mathbb{M}_2(\mathbb{C})$ has the form

$$H(k) = \begin{pmatrix} E_A & t \cdot \sum_{j=1}^3 \exp(2\pi \mathrm{i} k \cdot a_j) \\ t \cdot \sum_{j=1}^3 \exp(-2\pi \mathrm{i} k \cdot a_j) & -E_A \end{pmatrix}$$

(beware that the letters a and b are switched in the notation of [14]). Since $0 = a_1 - a_1$, $-b_2 = a_1 - a_3$, $b_3 = a_1 - a_2$, we may write

$$S_0 + S_{-b_2} + S_{b_3} = \sum_{j=1}^3 S_{a_1 - a_j} = S_{a_1} \sum_{j=1}^3 S_{-a_j}. \tag{3.3}$$

So the Fourier transform of $S_0 + S_{-b_2} + S_{b_3}$ in our formula differs from the expression used in the corresponding matrix entry in [14] by the factor $\exp(-2\pi \mathrm{i} k \cdot a_1)$. This extra factor does not change the spectrum: it only depends on the absolute value of $\tilde{h}(k)$.

The source of the difference in the formulas is the following. Atoms of type A and B sit at the points of Λ_0 and $\Lambda_0 + a_1$, respectively. So hopping from an atom of type A to one of type B involves a translation S_{a_1}, whereas hopping from an atom of type B to one of type A involves the inverse translation S_{-a_1}. The Fourier transform is defined in [4, 14] and other articles so that this translation is present. In contrast, we treat the two atoms in a crystal cell as two internal degrees of freedom in the same cell. So we do not see a hopping between the two atoms in a cell as a translation, and we do not put a factor of S_{-a_1} into our formulas. This is why the factors $\exp(\pm 2\pi \mathrm{i} k \cdot a_1)$ in the off-diagonal terms of the Hamiltonian are missing in our computations.

The extra factor in [14] causes several qualitative differences because it is not Λ^*-periodic. It is only periodic under the dual of the lattice generated by the vectors a_1, a_2, a_3. The vector a_1 together with the lattice Λ generates the same lattice as the vectors a_1, a_2, a_3; of course, one of a_1, a_2, a_3 is redundant. The quotient $(a_1\mathbb{Z} + \Lambda)/\Lambda$ has three elements because $a_1 \notin \Lambda$ and $3 \cdot a_1 = b_3 - b_2 \in \Lambda$. Thus the Hamiltonian in [14] lives on the quotient of $\mathbb{R}^2$ by the lattice $(a_1\mathbb{Z} + \Lambda)^*$, which is strictly smaller than Λ^*. Three suitable copies of a fundamental domain for Λ^* form a fundamental domain for $(a_1\mathbb{Z} + \Lambda)^*$. We have found above that there are exactly two points in $k \in \mathbb{R}^2/\Lambda^*$ where $|1 + \exp(-2\pi \mathrm{i} k \cdot b_2) + \exp(2\pi \mathrm{i} k \cdot b_3)| = 0$. These correspond to six such points in $\mathbb{R}^2/(a_1\mathbb{Z} + \Lambda)^*$. This is why Semenoff finds six "Dirac cones" in the spectrum of $H(k)$ in the Brillouin zone (the fundamental domain), while we find only two.

The goal of Haldane [4] was to produce topological phases. His model was intended only as a toy model. By now, an experimental setup has been found that is governed by this Hamiltonian (see [6]). In Haldane's time, topological phases appeared as an analogue of the quantum Hall effect that occurs without magnetic field. A crucial observation by him was that this requires breaking the time-reversal symmetry of the graphene model. That is, we need a model with complex hopping terms, and complex t does not count because the phase of this parameter may be gauged away. So his Hamiltonian can only be invariant under the group G_4. In fact, the Hamiltonians he considers are those we built above with the restrictions $E_B = -E_A$ and $z_B = \overline{z_A}$; he writes M instead of E_A and t_1 instead of t, and he writes the parameter z_A in polar coordinates as $t_2 \cdot \cos(\varphi) + \mathrm{i}t_2 \sin(\varphi)$. Moreover, Haldane uses the Pauli matrices to express 2×2-matrices:

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma_2 = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, \qquad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = -\mathrm{i}\sigma_1\sigma_2 = \mathrm{i}\sigma_2\sigma_1.$$

In addition, I in [4] denotes the unit matrix. The matrices σ_i for $i = 1, 2, 3$ and I form a basis for $\mathbb{M}_2(\mathbb{C})$. With this notation, Haldane's Hamiltonian is

$$H_{\text{Haldane}}(k) \coloneqq 2t_2 \cos\varphi \left(\sum_{j=1}^{3} \cos(k \cdot b_j) \right) \cdot I + t_1 \left(\sum_{j=1}^{3} \cos(k \cdot a_j)\sigma_1 + \sin(k \cdot a_j)\sigma_2 \right) + \left(M - 2t_2 \sin\varphi \sum_{j=1}^{3} \sin(k \cdot b_j) \right) \cdot \sigma_3. \quad (3.4)$$

(The online course "Topology in Condensed Matter" specialises further to the case when z_A is purely imaginary, that is, $\cos\varphi = 0$.) Compared to our formula, (3.4) differs only by renaming the constants and by an extra factor $\exp(\pm 2\pi \mathrm{i} k \cdot a_1)$ in the off-diagonal terms. The source of this extra factor is the same as in [14].

The graphene Hamiltonian introduced above is refined in [19] in order to estimate the effect of spin-orbit coupling. It is argued in [7] that this effect opens a gap in the spectrum. The estimate of the size of the gap in [7] is found to be wrong by orders of magnitude in [19]: it is so small as to be negligible. The computation in [19] starts with a model with 8 orbitals for each atom, giving 16 degrees of freedom for each unit cell. Many of these orbitals have energies so far from the Fermi level that they can be safely ignored. The approximate Hamiltonian that is reached in the end has two orbitals per atom, corresponding to spin up and spin down versions of the orbital in the computation above. These two orbitals do not interact with each other. Both are described by Haldane models as above with different constants depending on the spin.

3.3.4 The Bernevig–Hughes–Zhang model

Bernevig, Hughes and Zhang [1] were the first to propose a material that behaves like a topological insulator in the "quantum spin Hall state". Their material is a thin layer of Hg Te in between Cd Te, which they call a quantum well; the system behaves differently depending on the thickness of the Hg Te-layer. It is again a 2-dimensional material.

Their model starts with six states per unit cell, which are the spin-up and spin-down components of three orbitals. It is argued, however, that the two components of one orbital have energies far from the other four and may safely be ignored. So their model has four degrees of freedom per unit cell, modelled on the Hilbert space $\ell^2(\mathbb{Z}^2, \mathbb{C}^4)$; this is isomorphic to $L^2(\mathbb{T}^2, \mathbb{C}^4)$ by Fourier transform.

Two of the four bands are symmetric and two are antisymmetric with respect to the third z-variable. This makes certain matrix coefficients of the Hamiltonian vanish; namely, it has the block diagonal form

$$H_{\text{eff}} = \begin{pmatrix} H_1 & 0 \\ 0 & H_2 \end{pmatrix}$$

with $H_1, H_2 \in \mathbb{B}(\ell^2(\mathbb{Z}^2, \mathbb{C}^2))$. The model also has a Fermionic time-reversal symmetry Θ with $\Theta^2 = -1$. It is acting only on the $\mathbb{C}^4$-factor (constant on $\mathbb{Z}^2$) and has the form

$$\Theta\colon \mathbb{C}^4 \to \mathbb{C}^4, \qquad (\xi_1, \xi_2, \eta_1, \eta_2) \mapsto \overline{(-\eta_1, -\eta_2, \xi_1, \xi_2)},$$

as in Example 2.13. The time-reversal symmetry forces $H_2 = \overline{H_1}$. For the function on $\mathbb{T}^2$, this means that $H_2(k) = \overline{H_1(-k)}$. Rotation by $2\pi/4$, mapping $(x, y) \mapsto (-y, x)$, is a symmetry of the lattice $\mathbb{Z}^2$. This is a twisted symmetry of the matrix H_1, that is, the latter commutes with the operator

$$RU\colon \ell^2(\mathbb{Z}^2, \mathbb{C}^2) \to \ell^2(\mathbb{Z}^2, \mathbb{C}^2), \qquad (RUf)(x, y) = \begin{pmatrix} 1 & 0 \\ 0 & -\mathrm{i} \end{pmatrix} \cdot f(-y, x).$$

When we restrict attention to next-neighbour hopping terms, the symmetry constraints above force H_1 to have the form

$$H_1 = \begin{pmatrix} \alpha S_0 + \beta(S_x + S_{-x} + S_y + S_{-y}) & \gamma(S_x - S_{-x} - \mathrm{i}(S_y - S_{-y})) \\ \overline{\gamma}(S_x - S_{-x} + \mathrm{i}(S_y - S_{-y})) & \delta S_0 + \varepsilon(S_x + S_{-x} + S_y + S_{-y}) \end{pmatrix}$$

with constants $\alpha, \beta, \delta, \varepsilon \in \mathbb{R}$ and $\gamma \in \mathbb{C}$. The phase of γ may be changed by conjugating H_1 (and hence H) with a suitable diagonal matrix. So it is no loss of generality to assume $\gamma \in \mathrm{i}\mathbb{R}$ as well. The Hamiltonian above is the model Hamiltonian considered in [1] except for the naming of the constants. First, in [1] 2×2-matrix valued functions are written as linear combinations of Pauli matrices. Secondly, the discussion in [1] starts with the second Taylor polynomial at 0 of the function $\mathbb{T}^2 \to \mathbb{M}_2(\mathbb{C})$ corresponding to H. The parameters are chosen as the Taylor coefficients of the coefficients of $H(k)$ in the Pauli matrix basis. Namely, the model in [1] is

$$\begin{aligned} H_1 = (C - 2D) \cdot (S_0 \otimes I) + D \cdot (S_x + S_{-x} + S_y + S_{-y}) \otimes I \\ + \frac{A}{2\mathrm{i}} \cdot (S_x - S_{-x}) \otimes \sigma_1 + \frac{A}{2\mathrm{i}} \cdot (S_y - S_{-y}) \otimes \sigma_2 \\ (M - 4B + B(S_x + S_{-x} + S_y + S_{-y})) \otimes \sigma_3. \end{aligned} \tag{3.5}$$

More precisely, we should look at [1, Equations (2) and (6)]. They describe the periodic function $H(k)\colon \mathbb{T}^2 \to \mathbb{M}_4(\mathbb{C})$ through sines and cosines of the quasi-momenta k_x, k_y as

$$H_1(k) = \big(C - 2D\big(2 - \cos(k_x) - \cos(k_y)\big)\big) \cdot I + A\sin(k_x)\sigma_1 + A\sin(k_y)\sigma_2 \\ - 2B\big(2 - (M/2B) - \cos(k_x) - \cos(k_y)\big)\sigma_3,$$

with real constants A, B, C, D, M. This function corresponds to the operator on $\ell^2(\mathbb{Z}^2, \mathbb{C}^2)$ in (3.5).

3.3.5 The Su–Schrieffer–Heeger model

The Su–Schrieffer–Heeger model [15, 16] describes the electric properties of polyacetylene. This is a linear polymer, that is, a long chain of CH joined with alternating simple and double bonds. Being a long chain, this is a 1-dimensional material. The tight-binding model for it has two orbitals for each crystal cell. So the Hilbert space is $\ell^2(\mathbb{Z}, \mathbb{C}^2)$. This system has the chiral symmetry Ξ given by the diagonal matrix with entries $1, -1$. The model has only nearest-neighbour hopping terms. The most general form of such a Hamiltonian is

$$\begin{pmatrix} 0 & aS + b + cS^* \\ \overline{a}S^* + \overline{b} + \overline{c}S & 0 \end{pmatrix}$$

The Su–Schrieffer–Heeger model has $a = 1$, $b = \mathrm{i}m$, $c = 0$ for some $m \in \mathbb{R}$ (we follow [11]). In terms of Pauli matrices, this becomes

$$H = \frac{1}{2}(\sigma_1 + \mathrm{i}\sigma_2) \otimes S + \frac{1}{2}(\sigma_1 - \mathrm{i}\sigma_2) \otimes S^* + m\sigma_2 \otimes 1.$$

The corresponding matrix-valued function on $\mathbb{R}/\mathbb{Z}$ is

$$H(k) = \begin{pmatrix} 0 & \exp(2\pi\mathrm{i}k) + \mathrm{i}m \\ \exp(-2\pi\mathrm{i}k) - \mathrm{i}m & 0 \end{pmatrix}. \tag{3.6}$$

The eigenvalues of this matrix are

$$E_\pm(k) := \pm|\exp(2\pi\mathrm{i}k) + \mathrm{i}m| = \pm\sqrt{m^2 + 1 + 2m\sin(2\pi k)}.$$

The value 0 is an eigenvalue for some $k \in [0, 1]$ if and only if $|m| = 1$. Then the spectrum is the interval $[|m| - 1, |m| + 1]$. If $|m| > 1$, the spectrum of H is the union of the two intervals $\pm[|m| - 1, |m| + 1]$. So there is a spectral gap $(-|m| - 1, |m| + 1)$ around 0. If $|m| < 1$, then the spectrum of H is the union of the two intervals $\pm[1 - |m|, |m| + 1]$, and there is a spectral gap $\big(-(1 - |m|), 1 - |m|\big)$ around 0.

The Hamiltonian H has an obvious chiral symmetry, which explains why the spectrum is symmetric around 0.

4 Topological phases

In this section, we are going to define the equivalence relation of homotopy on the set of Hamiltonians that describe insulators. This is used by many physicists to define "topological phases" (see, for instance, [8]). I prefer coarser equivalence relations such as "stable homotopy" instead of homotopy. Since these are more complicated, I discuss homotopy first.

We fix a Hilbert space $\mathcal{K}$ and consider self-adjoint operators on $\ell^2(\mathbb{Z}^d, \mathcal{K}) \cong L^2(\mathbb{T}^d, \mathcal{K})$. A homotopy between such operators should be a family $(H_t)_{t\in[0,1]}$ of self-adjoint operators that depends continuously on t. Here the right notion of continuity is the norm continuity of resolvents: the map $t \mapsto (\lambda - H_t)^{-1}$ is continuous in the norm topology on bounded operators for $\lambda \in \mathbb{C} \setminus \bigcup_{t\in[0,1]} \sigma(H_t)$; if this holds for one $\lambda \notin \bigcup_{t\in[0,1]} \sigma(H_t)$, then it holds for all such λ. For instance, the rescaling homotopy $(1-t)\cdot H$ is a homotopy from any self-adjoint operator H to the zero operator. We are only interested in self-adjoint operators with some extra properties: they should be bounded below, affiliated with $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$ (see Theorem 3.3) and have a spectral gap at the Fermi energy μ. There are ways to tune the Fermi energy, so that it is quite reasonable to allow homotopies to also shift the Fermi energy. To simplify the setup, however, we shift the Fermi energy to 0 by subtracting μ from the Hamiltonian. This normalisation allows us to restrict attention to the case where the Fermi energy is 0. Thus we describe experiments that tune the Fermi energy by changing the Hamiltonian instead.

Definition 4.1. Let H_0, H_1 be self-adjoint operators on $L^2(\mathbb{T}^d, \mathcal{K})$ that are affiliated with $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$, bounded below, and invertible. We call H_0, H_1 homotopic if there is a homotopy $(H_t)_{t\in[0,1]}$ between them (continuous in the topology of norm convergence of resolvents) such that all H_t are self-adjoint operators affiliated with $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$, bounded below, and invertible.

I am not aware of a mathematically precise treatment of topological phases in the situation where $\mathcal{K}$ has infinite dimension. We are going to discuss the subtle issues that arise in this case. The realistic Hamiltonians that occur in this case are unbounded. We also assume them to be bounded below – otherwise the classification breaks down, as we shall see.

4.1 The case of infinite-dimensional fibre Hilbert space

Let H be a self-adjoint, bounded below, invertible operators affiliated with $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$. Let $-c < 0$ be a lower bound on H, that is, $H + c \geq 0$. Then the spectral projection $\chi_{(-\infty,0)}(H)$ is equal to $\chi_{(-c,0)}(H)$. This belongs to $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$ because it is given by functional calculus with a continuous function on the spectrum of H that vanishes at $\pm\infty$. So we get a projection $\chi_{(-\infty,0)}(H)$ in $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$. If $(H_t)_{t\in[0,1]}$ is a homotopy of Hamiltonians as in Definition 4.1, then the family of projections $t \mapsto \chi_{(-\infty,0)}(H_t)$ is norm continuous. We call such a norm-continuous map a homotopy of projections, and two projections in $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$ are called homotopic if they are connected by a norm-continuous homotopy of projections.

Theorem 4.2. *Any projection in* $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$ *is of the form* $\chi_{(-\infty,0)}(H)$ *for a self-adjoint, bounded below, invertible operators affiliated with* $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$*. Let* H_0, H_1 *be self-adjoint, bounded below, invertible operators affiliated with* $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$*. Then* H_0 *and* H_1 *are homotopic if and only if the corresponding spectral projections* $\chi_{(-\infty,0)}(H_0)$ *and* $\chi_{(-\infty,0)}(H_1)$ *are homotopic.*

Proof. We first prove the statements in the easy case where $\dim \mathcal{K} < \infty$. Then any linear operator on $\mathcal{K}$ is compact, and a Hamiltonian affiliated with $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$ is bounded. Let $P \in \mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$ be a projection. Then $F \coloneqq 1 - 2P$ satisfies $F = F^*$ and $F^2 = 1 - 4P + 4P^2 = 1$. So F is invertible with $F^{-1} = F$, and the spectrum of F is $\{\pm 1\}$. The spectral projection $\chi_{(-\infty,0)}(F)$ is P. We have already seen above that a homotopy between two invertible Hamiltonians H_0 and H_1 generates a homotopy between the spectral projections $P_t \coloneqq \chi_{(-\infty,0)}(H_t)$ for $t = 0, 1$. Conversely, assume that P_0 and P_1 are homotopic through a norm-continuous homotopy of projections $(P_t)_{t\in[0,1]}$. Then $F_t \coloneqq 1 - 2P_t$ is a homotopy of invertible, self-adjoint operators affiliated with $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$ that connects $F_0 = \operatorname{sign}(H_0)$ and $F_1 = \operatorname{sign}(H_1)$. The operator F_t is homotopic to H_t by functional calculus with the functions $x \mapsto s \cdot x + (1 - s)\operatorname{sign}(x)$, which also gives only invertible, self-adjoint operators affiliated with $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$. Concatenating these homotopies gives a homotopy from H_0 to H_1.

Now we treat the more difficult case $\dim \mathcal{K} = \infty$. Let $A \coloneqq \mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$. First let $P \in A$ be a projection. Let $P^\perp \coloneqq 1 - P$. The corner $P^\perp A P^\perp$ is separable and hence contains a strictly positive element $x \in P^\perp A P^\perp$. The corresponding operator on $P^\perp \cdot L^2(\mathbb{T}^d, \mathcal{K})$ is positive and injective. So its inverse is a densely defined unbounded operator, which is positive and hence self-adjoint, and affiliated with $P^\perp A P^\perp$. Let $H \coloneqq -P + x^{-1}$. This is a self-adjoint, bounded below, invertible operator affiliated with $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$, and $\chi_{(-\infty,0)}(H) = P$.

Now let H_0 and H_1 be self-adjoint, bounded below, invertible operator affiliated with $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$. Let $P_t \coloneqq \chi_{(-\infty,0)}(H_t)$ for $t = 0, 1$, and let $(P_t)_{t\in(0,1)}$ be a homotopy of projections in $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$ between P_0 and P_1. This is a projection in $\mathrm{C}([0,1] \times \mathbb{T}^d, \mathbb{K}(\mathcal{K}))$. As above, we may find a self-adjoint, bounded below, invertible operator affiliated with $\mathrm{C}([0,1] \times \mathbb{T}^d, \mathbb{K}(\mathcal{K}))$ with spectral projection $(P_t)_{t\in[0,1]}$. And this is a homotopy between some self-adjoint, bounded below, invertible operator H_0' and H_1' that are affiliated with $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$ and that have the same spectral projections P_0 and P_1 as H_0 and H_1. So it remains to prove that H_0 and H_0' are homotopic and that H_1 and H_1' are homotopic. This reduces the general case to the case where H_0 and H_1 have the same spectral projections $P \coloneqq \chi_{(-\infty,0)}(H_0) = \chi_{(-\infty,0)}(H_1)$. We claim that

$$H_t \coloneqq \left((1-t)H_0^{-1} + tH_1^{-1}\right)^{-1}$$

is a norm continuous path of self-adjoint, bounded below, invertible operators affiliated with $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$ that connects H_0 and H_1. First, both H_0^{-1} and H_1^{-1} commute with P. And PH_0P and PH_1P are strictly negative elements in the unital C*-algebra PAP. Then $PH_i^{-1}P$ is the inverse of H_i in PAP, which is again strictly negative. The linear homotopy between these strictly negative elements is strictly negative in PAP, and

hence again invertible. The inverse is PH_tP. So this is a norm continuous path of invertible elements in PAP. The operators $P^\perp H_j P^\perp$ for $j = 0, 1$ are strictly positive, self-adjoint, unbounded operators on $P^\perp L^2(\mathbb{T}^d, \mathcal{K})$ with resolvent in $P^\perp A P^\perp$. Thus $P^\perp H_j^{-1} P^\perp \in P^\perp A P^\perp$ for $j = 0, 1$ are strictly positive elements of $P^\perp A P^\perp$. Then the linear path between them is a norm continuous homotopy of strictly positive elements in $\mathrm{C}([0,1], P^\perp A P^\perp)$. Taking the inverse then gives a well defined norm continuous homotopy of unbounded self-adjoint operators on $P^\perp L^2(\mathbb{T}^d, \mathcal{K})$; this path is exactly $P^\perp H_t P^\perp$. Putting the positive and negative parts together gives the result. □

If $\dim \mathcal{K} < \infty$, then the main idea in the proof of Theorem 4.2 is to replace an invertible, self-adjoint operator H by $\operatorname{sign}(H)$. This operation is called *spectral flattening* because it mostly forgets the spectrum; only the projections on the positive and negative parts of the spectrum remain. The proof shows that H is homotopic to $\operatorname{sign}(H)$. Our assumptions in the case $\dim \mathcal{K} = \infty$ are such that the same result remains true, although spectral flattening now leaves the class of allowed Hamiltonians: the operator $\operatorname{sign}(H)$ is bounded and therefore does not have compact resolvent if $\dim \mathcal{K} = \infty$. The proof shows that a positive unbounded operator with compact resolvent can always be found that goes with the spectral projection to make an allowed Hamiltonian with the given negative spectral projection.

Assuming Hamiltonians to be bounded below is reasonable because of Theorem 3.3. Nevertheless, we would rather avoid this assumption because it rules out Hamiltonians with $\dim \mathcal{K} = \infty$ and with a chiral or a particle-hole symmetry. We need H to be bounded below in order for the spectral projection $\chi_{(-\infty,0)}(H)$ to belong to $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$. In fact, we get a rather trivial homotopy relation if we allow Hamiltonians that are not bounded below during the paths:

Proposition 4.3. *Assume* $\dim \mathcal{K} = \infty$. *Let H be a Hamiltonian that is bounded below. Then there is a continuous path of Hamiltonians $(H_t)_{t\in(0,1)}$ that are self-adjoint, invertible, and affiliated with* $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$, *bounded below – but not uniformly – and such that H_0 and $H_1 \geq 0$.*

Proof. This proof uses a non-trivial result, namely, the Kasparov Stabilisation Theorem for Hilbert modules over C*-algebras. We shall use this theory without further explanations because it only gives us a "no-go theorem", so no positive results in the future will depend on it. Let $P = \chi_{(-\infty,0)}(H)$. Then $P \cdot \mathrm{C}_0([0,1) \times \mathbb{T}^d, \mathcal{K})$ is a Hilbert module over the C*-algebra $\mathrm{C}([0,1] \times \mathbb{T}^d)$. By the Kasparov Stabilisation Theorem, there is a unitary operator

$$U \colon P \cdot \mathrm{C}_0([0,1) \times \mathbb{T}^d, \mathcal{K}) \oplus \mathrm{C}([0,1] \times \mathbb{T}^d, \mathcal{K}) \to \mathrm{C}([0,1] \times \mathbb{T}^d, \mathcal{K}).$$

Let x and y be strictly positive, compact operators on $P \cdot \mathrm{C}_0([0,1) \times \mathbb{T}^d, \mathcal{K})$ and $\mathrm{C}([0,1] \times \mathbb{T}^d, \mathcal{K})$, respectively. Then $x \oplus y$ is a strictly positive operator on the direct sum, which gives a strictly positive operator on $\mathrm{C}([0,1] \times \mathbb{T}^d, \mathcal{K})$ by conjugation with U. Let $L \coloneqq (U((-x) \oplus y)U^*)^{-1}$. This is a well defined, unbounded self-adjoint, invertible

element affiliated with $\mathrm{C}([0,1]\times\mathbb{T}^d,\mathbb{K}(\mathcal{K}))$. Its inverse is $U((-x)\oplus y)U^*$, which is a norm continuous function on the interval $[0,1]$ with values in $\mathrm{C}(\mathbb{T}^d,\mathbb{K}(\mathcal{K}))$ because

$$\mathrm{C}([0,1]\times\mathbb{T}^d,\mathbb{K}(\mathcal{K}))\cong\mathrm{C}([0,1],\mathrm{C}(\mathbb{T}^d,\mathbb{K}(\mathcal{K}))).$$

Hence it is equivalent to a homotopy of self-adjoint, invertible operators affiliated with $\mathrm{C}(\mathbb{T}^d,\mathbb{K}(\mathcal{K}))$. The negative spectral projections of these invertible Hamiltonians are $U_t(P\oplus 0)U_t^*$ for $t\in[0,1)$ and 0 for $t=1$. Since these spectral projections are all compact, the corresponding unbounded operators cannot help being bounded below. But the lower bound will go to $-\infty$ for $t\nearrow 1$. This is how the resolvent may remain norm continuous although the spectral projections manifestly fail to be norm continuous. The starting point of the homotopy above is a self-adjoint, invertible, bounded below operator affiliated with $\mathrm{C}(\mathbb{T}^d,\mathbb{K}(\mathcal{K}))$ with spectral projection $U_0(P\oplus 0)U_0^*$.

By Theorem 4.2, it now remains to show that the projections P and $U_0(P\oplus 0)U_0^*$ are homotopic. In fact, we will show that we may choose U_0 so that $P=U_0(P\oplus 0)U_0^*$, after replacing P by a homotopic projection, which is allowed by Theorem 4.2. Here we use some basic ideas from the K-theory of C*-algebras. Identify $\mathcal{K}\cong\ell^2(\mathbb{N})$. There are $n\in\mathbb{N}$ and $P'\in\mathrm{C}(\mathbb{T}^d,\mathbb{M}_n(\mathbb{C}))$ with $\|P'-P\|<1/100$, say. Then the spectrum of P' is close enough to $\{0,1\}$ that functional calculus with a continuous function turns P' into a projection close to P' and hence to P. These nearby projections are homotopic. So we may assume without loss of generality that already $P\in\mathrm{C}(\mathbb{T}^d,\mathbb{M}_n(\mathbb{C}))$. Let $P^\perp=1_n-P$ be the complementary projection in $\mathrm{C}(\mathbb{T}^d,\mathbb{M}_n(\mathbb{C}))$. Then reordering the summands gives a unitary operator

$$\begin{aligned}\mathrm{C}(\mathbb{T}^d,\mathcal{K})&=\bigoplus_{j\in\mathbb{N}}\mathrm{C}(\mathbb{T}^d,\mathbb{C}^n)\cong\bigoplus_{j\in\mathbb{N}}\Big(P\cdot\mathrm{C}(\mathbb{T}^d,\mathbb{C}^n)\oplus P^\perp\cdot\mathrm{C}(\mathbb{T}^d,\mathbb{C}^n)\Big)\\&\cong P\cdot\mathrm{C}(\mathbb{T}^d,\mathbb{C}^n)\oplus\bigoplus_{j\in\mathbb{N}}\Big(P^\perp\cdot\mathrm{C}(\mathbb{T}^d,\mathbb{C}^n)\oplus P\cdot\mathrm{C}(\mathbb{T}^d,\mathbb{C}^n)\Big)\cong P\cdot\mathrm{C}(\mathbb{T}^d,\mathbb{C}^n)\oplus\mathrm{C}(\mathbb{T}^d,\mathcal{K}).\end{aligned}$$

This unitary has the special feature that it maps $P\cdot\mathrm{C}(\mathbb{T}^d,\mathcal{K})$ onto the first direct summand $P\cdot\mathrm{C}(\mathbb{T}^d,\mathbb{C}^n)$. Now we may arrange in the construction of U that U_t is constant equal to this unitary for $t\in[0,\varepsilon]$. If we arrange this so, then $P=U_0(P\oplus 0)U_0^*$. □

Even in the finite-dimensional case, the invertibility assumption on Hamiltonians is crucial. We may connect any two bounded self-adjoint operators by the straight path $H_t\coloneqq tH_1+(1-t)H_0$. Therefore, there is a homotopy between any two self-adjoint operators. If, however, H_0 and H_1 correspond to insulators and are not homotopic as such, then for any homotopy $(H_t)_{t\in(0,1)}$ between them, there must be $t\in(0,1)$ so that H_t is not invertible, that is, $0\in\sigma(H_t)$. This is already a reason why homotopy is interesting. Most of the model Hamiltonians considered above contain some parameters, which may be modified in experiments. Assume that the Hamiltonian depends continuously on a parameter $a\in\mathbb{R}$ and that H_{a_0} and H_{a_1} for $a_0<a_1$ are invertible and not homotopic among invertible self-adjoint operators. Then there must be $t\in(a_0,a_1)$ for which $0\in\sigma(H_t)$ because otherwise H_{a_0} and H_{a_1} would be homotopic. If $0\in\sigma(H_t)$, then H_t is

a conductor (or at least a semimetal). Physicists say that two insulators are homotopic if and only if they may be "adiabatically connected" (see, for instance, [8]).

We are also interested in the following situation. Cut the space into two halfs and fill one with an insulator governed by H_0, the other with an insulator governed by H_1; often, H_1 is a "trivial" system describing the vacuum. Let H be the Hamiltonian of this total system. It is no longer periodic unless $H_0 = H_1$. So band theory as developed so far does not apply to it. We will treat such systems in Section 6. It may happen that H is not invertible although H_0 and H_1 are. That is, putting two insulators together may give a conductor. When this happens, we may wonder how robust the effect is: can we modify H near the boundary that separates the two regions so as to remove the spectrum near 0? Sometimes, it can be shown that this is impossible by topological arguments. Roughly speaking, there are topological invariants that differ for H_0 and H_1, and this implies that H cannot be invertible. It is not true that the mere lack of a homotopy already suffices for this. There is, however, a vague physical argument why the lack of a homotopy between H_0 and H_1 implies that a sufficiently "smooth" transition H between them cannot be invertible. Here we mean that the Hamiltonian is varying quite slowly: for some sufficiently large $\ell > 0$, the Hamiltonian is very close to the restriction of a periodic Hamiltonian in the strips $x_1 \in [a, a+\ell]$ for all $a \in \mathbb{R}$, and equal to the restriction of H_0 or H_1 for $x_1 \ll 0$ or $x_1 \gg 0$, respectively. Let H_a be a periodic Hamiltonian to which H is close in the strip $x_1 \in [a, a+\ell]$. Then H_a and H_b must be close for $a \approx b$ and so we get a homotopy between H_0 and H_1. This homotopy cannot consist entirely of insulators. So an operator close to H_a for some $a \in \mathbb{R}$ describes a conductor. Since H is close to H_a in the strip $x_1 \in [a, a+\ell]$, we expect that H cannot be invertible. This argument is vague, and a smooth transition between two materials is uncommon. So we will not pursue this idea further.

4.2 Some basic homotopy theory

We now return to the mathematical theory. We fix the Hilbert space $\mathcal{K}$. Theorem 4.2 says that the homotopy classification of Hamiltonians of insulators is equivalent to the homotopy classification of projections in $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$. A projection in $\mathrm{C}(\mathbb{T}^d, \mathbb{K}(\mathcal{K}))$ is simply a continuous map from $\mathbb{T}^d$ to the space of finite-rank, orthogonal projections on the Hilbert space $\mathcal{K}$. These orthogonal projections correspond bijectively to finite-dimensional, closed subspaces of $\mathcal{K}$. Since $\mathbb{T}^d$ is connected, the rank of these projections is constant, say, r. Physically, r is the number of filled bands in our model. By definition, the topological space of such orthogonal projections of rank r is the *Grassmann manifold* $\mathrm{Gr}(r, \dim \mathcal{K})$. This is a famous example of a compact manifold if $\dim \mathcal{K} < \infty$. It is an infinite-dimensional "manifold" if $\dim \mathcal{K} = \infty$. So homotopy classes of Hamiltonians of insulators with r filled bands and $\dim \mathcal{K}$ bands altogether are in bijection with homotopy classes of continuous maps $\mathbb{T}^d \to \mathrm{Gr}(r, \dim \mathcal{K})$.

Given two topological spaces X and Y, let $[X, Y]$ denote the set of homotopy classes of continuous maps $X \to Y$. So our problem is to describe the set $[\mathbb{T}^d, \mathrm{Gr}(r, n)]$ for $r \leq n \leq \infty$. This is a prototypical problem of homotopy theory. A basic idea in homotopy theory is to reduce the computation of $[X, Y]$ to similar groups where X is

simpler, namely, a sphere. In the case at hand, we may also simplify the target space $\mathrm{Gr}(r,n)$, replacing it by suitable unitary groups. Let $\mathrm{U}(n)$ be the group of unitary operators on an n-dimensional Hilbert space (*unitary group*), and let $\mathrm{SU}(n)$ be the subgroup of unitary matrices of determinant 1 (*special unitary group*). We assume for simplicity that $\dim \mathcal{K} < \infty$ and identify $\mathcal{K}$ with $\ell^2(\{1,\dots,\dim \mathcal{K}\})$.

Lemma 4.4. *Let $r \geq 0$. Let P_0 be the projection onto the subspace of $\mathcal{K}$ spanned by the first r basis vectors. The map $\mathrm{SU}(n) \to \mathrm{Gr}(r,n)$, $U \mapsto UP_0U^*$, descends to a diffeomorphism*

$$\mathrm{SU}(n) \,/\, \big(\mathrm{U}(r) \times \mathrm{U}(n-r) \cap \mathrm{SU}(n)\big) \to \mathrm{Gr}(r,n).$$

Proof. Let P_1 be any orthogonal projection on $\mathcal{K}$ of rank r. Choose bases for the image of P_1 and its orthogonal complement. This defines a unitary operator $U \in \mathrm{U}(\mathcal{H})$ that maps the image of P_0 onto the image of P_1. Equivalently, $P_1 = UP_0U^*$. Multiplying one of the basis vectors by a phase, we may arrange that $\det(U) = 1$. So the map $\mathrm{SU}(n) \to \mathrm{Gr}(r,n)$, $U \mapsto UP_0U^*$, is surjective. If $U, V \in \mathrm{SU}(n)$ satisfy $UP_0U^* = VP_0V^*$, then U^*V commutes with P_0. Equivalently, $U^*V \in \mathrm{U}(r) \times \mathrm{U}(n-r) \cap \mathrm{SU}(n)$. Hence our map descends to a bijection from $\mathrm{SU}(n) \,/\, \big(\mathrm{U}(r) \times \mathrm{U}(n-r) \cap \mathrm{SU}(n)\big)$ onto $\mathrm{Gr}(r,n)$. Now some further calculus shows, first, that the set of orthogonal projections in $\mathbb{M}_n(\mathbb{C})$ is a smooth submanifold. This is the original manifold structure on $\mathrm{Gr}(r,n)$. Secondly, the surjection $\mathrm{SU}(n) \to \mathrm{Gr}(r,n)$ is a submersion. This implies that the induced bijection from $\mathrm{SU}(n) \,/\, \big(\mathrm{U}(r) \times \mathrm{U}(n-r) \cap \mathrm{SU}(n)\big)$ onto $\mathrm{Gr}(r,n)$ is a diffeomorphism. □

The computation of $[\mathbb{T}^d, \mathrm{Gr}(r,n)]$ must be done by a recursion over d. This is because the d-torus contains the $d-1$-torus as a retract, that is, there are continuous maps $\mathbb{T}^{d-1} \to \mathbb{T}^d \to \mathbb{T}^{d-1}$ that compose to the identity map on $\mathbb{T}^{d-1}$. These continuous maps induce continuous maps $[\mathbb{T}^{d-1}, \mathrm{Gr}(r,n)] \to [\mathbb{T}^d, \mathrm{Gr}(r,n)] \to [\mathbb{T}^{d-1}, \mathrm{Gr}(r,n)]$ that compose to the identity map. As a result, any computation of $[\mathbb{T}^d, \mathrm{Gr}(r,n)]$ must contain the computation of $[\mathbb{T}^{d-1}, \mathrm{Gr}(r,n)]$. So to describe $[\mathbb{T}^d, \mathrm{Gr}(r,n)]$ in the physically interesting cases $d = 2, 3$, we must begin with the easier cases $d = 0$ and $d = 1$. The space $\mathbb{T}^0$ is a single point. So a continuous map $\mathbb{T}^0 \to Y$ is a single point in Y, and a homotopy of such maps is a continuous path between these points. As a result, $[\mathbb{T}^0, Y]$ is the set of path components of Y, which is usually denoted by $\pi_0(Y)$.

Lemma 4.5. *The groups $\mathrm{SU}(n)$ and $\mathrm{U}(r) \times \mathrm{U}(n-r) \cap \mathrm{SU}(n)$ and the space $\mathrm{Gr}(r,n)$ are path connected.*

Proof. Let $U \in \mathrm{SU}(n)$. We may choose a basis in which U is diagonal. Taking a logarithm for each eigenvalue gives a matrix $\log(U)$ with $\exp(\log U) = U$. Since $\det(U) = \exp(\mathrm{tr}(\log U)) = 1$, the trace of $\log(U)$ belongs to $2\pi\mathrm{i}\mathbb{Z}$. We may subtract this from one of the diagonal entries to get a matrix $\log(U)$ with $\mathrm{tr}\log(U) = 0$ and $\exp(\log U) = U$. Then $U_t \coloneqq \exp(t\log(U))$ is a smooth path from the unit matrix to U, consisting of matrices of determinant 1. This shows that $\mathrm{SU}(n)$ is path connected. For a matrix in $\mathrm{U}(r) \times \mathrm{U}(n-r) \cap \mathrm{SU}(n)$, we may choose the eigenbasis above so that $\log U$ commutes with the projection P_0 in the proof of Lemma 4.4. Then the homotopy above belongs to $\mathrm{U}(r) \times \mathrm{U}(n-r) \cap \mathrm{SU}(n)$. So this space is path connected as well. Now let $P_1 \in \mathrm{Gr}(r,d)$.

Then $P_1 = UP_0U^*$ for some $U \in \mathrm{SU}(n)$ by Lemma 4.4. And $P_t \coloneqq U_t P_0 U_t^*$ for the path built above is a smooth path from P_0 to P_1 in $\mathrm{Gr}(r,d)$. Thus $\mathrm{Gr}(r,d)$ is path connected as well. □

As a consequence, $[\mathbb{T}^0, \mathrm{Gr}(r,d)]$ has a single point.

Next we consider the case $d = 1$. The space $\mathbb{T}^1$ is a circle. We shall describe it as the quotient $[0,1]/\sim$ where $0 \sim 1$. We first study $[\mathbb{T}^1, Y]$ for a general space Y. Recall that $\mathbb{T}^0$ is a retract in $\mathbb{T}^1$. So $[\mathbb{T}^0, Y] = \pi_0(Y)$ is a retract of $[\mathbb{T}^1, Y]$. In other words, there is a disjoint union decomposition of $[\mathbb{T}^1, Y]$ into subsets indexed by $\pi_0(Y)$. To single out one of these subsets, we fix $y_0 \in Y$ and restrict attention to maps $f\colon \mathbb{T}^1 \to Y$ with $f([0]) = 1$. These are equivalent to continuous maps $f\colon [0,1] \to Y$ with $f(0) = f(1) = y_0$. Such a map is called a *loop* in Y based at y_0.

A *homotopy* between two loops f_0, f_1 is a continuous map $f\colon [0,1]^2 \to Y$ with $f(s,t) = f_s(t)$ for $s = 0, 1$, $t \in [0,1]$ and $f(s,0) = f(s,1) = y_0$ for all $s \in [0,1]$. Homotopy is an equivalence relation on loops, and the set of equivalence classes is denoted by $\pi_1(Y, y_0)$. Two loops may be concatenated by running through them one ofter the other: $f_0 * f_1(t) = f_0(2t)$ for $0 \le t \le 1/2$ and $f_0 * f_1(t) = f_1(2t-1)$ for $1/2 \le t \le 1$. Concatenating three loops like this is multiplicative up to a reparametrisation. So the induced multiplication on $\pi_1(Y, y_0)$ is associative. The constant loop $f(t) = y_0$ for $t \in [0,1]$ is a unit element for this operation. And any loop f has an inverse (up to homotopy of loops), namely, the loop defined by $f^{-1}(t) = f(1-t)$. Thus $\pi_1(Y, y_0)$ is a group; it is called the *fundamental group* of Y based at y_0. The homotopy relation on loops is not quite the same as the homotopy relation in the definition of $[\mathbb{T}^1, Y]$ because we require $f(s,0) = f(s,1) = y_0$ for all $s \in [0,1]$ for a homotopy of loops.

Let $f_0, f_1\colon [0,1]/\sim \to Y$ be two loops based at y_0. A homotopy between them as in the definition of $[\mathbb{T}^1, Y]$ is equivalent to a continuous map $f\colon [0,1]^2 \to Y$ with $f(s,t) = f_s(t)$ for $s = 0, 1$ and $t \in [0,1]$ and $f(s,0) = f(s,1)$ for all $s \in [0,1]$; but $f(s,0) = f(s,1)$ may differ from y_0. Hence the restriction of f to the faces $[0,1] \times \{t\}$ for $t = 0, 1$ is another loop $\gamma \in \pi_1(Y, y_0)$. We may reparametrise the function f on $[0,1]^2$ so that it becomes a homotopy between f_0 and the loop $\gamma * f_1 * \gamma^{-1}$. In fact, one can show that f_0, f_1 are equal in $[\mathbb{T}^1, Y]$ if and only if there is $\gamma \in \pi_1(Y, y_0)$ with $f_0 = \gamma * f_1 * \gamma^{-1}$ in $\pi_1(Y, y_0)$. Even more,

$$[\mathbb{T}^1, Y] \cong \bigsqcup_{y_0 \in \pi_0(Y)} \pi_1(Y, y_0)/\text{conjugation}, \tag{4.1}$$

where $\pi_1(Y, y_0)/\text{conjugation}$ denotes the set of conjugacy classes of the group $\pi_1(Y, y_0)$. The groups $\pi_1(Y, y_0)$ and $\pi_1(Y, y_1)$ for y_0, y_1 in the same conjugacy class are isomorphic, so that this set does not depend on the choice of y_0 in its path component.

A topological space is called *simply connected* if it is path connected and any two loops are homotopic (as loops). If Y is simply connected, then (4.1) implies that $[\mathbb{T}^1, Y]$ has only one point. Hence the following lemma completes the computation of $[\mathbb{T}^1, \mathrm{Gr}(d,n)]$:

Lemma 4.6. *The spaces* $\mathrm{SU}(n)$ *and* $\mathrm{Gr}(r,n)$ *are simply connected.*

Proof. We have already seen in Lemma 4.5 that these spaces are path connected. We may choose the base points at will and take $1 \in \mathrm{SU}(n)$ and $P_0 \in \mathrm{Gr}(r,n)$. We need a

new tool to compute the fundamental group of $\mathrm{SU}(n)$, namely, the long exact sequence of homotopy groups for a fibration. An example of a fibration is the surjective submersion $q\colon \mathrm{SU}(n) \to \mathrm{Gr}(r,n)$ used above. To explain how the long exact sequence works, we first apply it to this fibration and show that $\mathrm{Gr}(r,n)$ is simply connected provided $\mathrm{SU}(n)$ is simply connected.

Let $f\colon [0,1] \to \mathrm{Gr}(r,n)$ be a loop based at P_0, that is, $f(0) = f(1) = P_0$. Since the map q is a surjective submersion, we may lift f locally to a continuous map to $\mathrm{SU}(n)$. Since $[0,1]$ is compact, a finite number of such liftings produces a continuous map $\tilde{f}\colon [0,1] \to \mathrm{SU}(n)$ with $q \circ \tilde{f} = f$ and $\tilde{f}(0) = 1$. Since $f(0) = f(1)$, the other end point $\tilde{f}(1)$ must belong to the subgroup $G \coloneqq \mathrm{U}(r) \times \mathrm{U}(n-r) \cap \mathrm{SU}(n)$. Lemma 4.5 asserts that G is path connected. So there is a continuous path $g\colon [0,1] \to G$ with $g(0) = 1$ and $g(1) = \tilde{f}(1)$. Then $\tilde{f}(t) \cdot g(t)^{-1}$ is a loop in $\mathrm{SU}(n)$ based at y_0 with $q(\tilde{f}(t) \cdot g(t)^{-1}) = f(t)$. If $\mathrm{SU}(n)$ is simply connected, then this loop is homotopic to a trivial loop. Applying q to such a homotopy gives a homotopy between f and a trivial loop. So $\mathrm{Gr}(r,n)$ is simply connected.

Now let $q\colon Y \to Z$ be any surjective submersion between two path-connected, smooth, compact manifolds. Let $F \coloneqq q^{-1}(z_0) \subseteq Y$ for some $z_0 \in Z$ be its "fibre". Choose $y_0 \in F \subseteq Y$. The inclusion $F \hookrightarrow Y$ and $q\colon Y \to Z$ induce the solid maps in the following line:

$$\pi_1(F,y_0) \to \pi_1(Y,y_0) \to \pi_1(Z,z_0) \overset{\delta}{\dashrightarrow} \pi_0(F,y_0) \to \pi_0(Y,y_0) \to \pi_0(Z,z_0). \tag{4.2}$$

We are going to build the dashed arrow δ as well. The key point is that this sequence is "exact" where this makes sense. Namely, an element in $\pi_1(Y,y_0)$ goes to the neutral element of $\pi_1(Z,z_0)$ if and only if it comes from $\pi_1(F,y_0)$; an element in $\pi_1(Z,z_0)$ goes to the path component of y_0 in F if and only if it comes from $\pi_1(Y,y_0)$; and an element in $\pi_0(F)$ goes to the path component of y_0 in $\pi_0(Y)$ if and only if it comes from $\pi_1(Z,z_0)$. The construction of the map δ follows a key step in the proof above that $\mathrm{Gr}(r,n)$ is simply connected provided $\mathrm{SU}(n)$ is simply connected. Let $f\colon [0,1] \to Z$ be a loop in Z based at z_0. A local lifting construction gives a continuous map $\tilde{f}\colon [0,1] \to Y$ with $\tilde{f}(0) = y_0$ and $q \circ \tilde{f} = f$. So $\tilde{f}(1) \in q^{-1}(z_0) = F$ because f is a loop. Let $\delta([f])$ be the path component of $\tilde{f}(1)$ in $\pi_0(F)$. We should show that this depends only on the homotopy class of f and not on the choice of $\tilde{f}$. So let f_0, f_1 be two loops that are homotopic through a homotopy h. Let $\tilde{f}_0, \tilde{f}_1$ be liftings of f_0, f_1 as above. The local lifting construction also applies to the homotopy h, giving a continuous map $\tilde{h}\colon [0,1]^2 \to Y$ with $q \circ \tilde{h} = h$. Even more, we may choose $\tilde{h}$ with $\tilde{h}(0,t) = y_0$ for all $t \in [0,1]$ and $\tilde{h}(s,t) = \tilde{f}_t(s)$ for $t = 0, 1$ and all $s \in [0,1]$. Now $h(1,t)$ is a continuous path in the fibre F between $\tilde{f}_0(1)$ and $\tilde{f}_1(1)$. This shows that δ is well defined.

Now we have to check the various claims asserted by the exactness of the sequence (4.2). First, let $f\colon [0,1] \to Y$ be a loop in Y based at y_0 that goes to the neutral element in $\pi_1(Z,z_0)$. Then there is a homotopy from $q \circ f$ to the constant loop. Lifting this homotopy as above gives a homotopy from f to a loop with $q \circ f'(t) = z_0$ for all $t \in [0,1]$. In other words, f is homotopic to a loop in F. This is the exactness of (4.2) at $\pi_1(Y,y_0)$. Secondly, let $f\colon [0,1] \to Z$ be a loop in Z based at z_0. If f comes from a loop in Y,

then we may use this loop to define $\delta([f])$, showing that $\delta([f]) = z_0$. Conversely, assume that $\delta([f])$ is z_0. Let $\tilde f\colon [0,1] \to Y_0$ be a continuous map with $q \circ \tilde f = f$ and $\tilde f = y_0$. By assumption, there is a continuous map $g\colon [0,1] \to F$ with $g(0) = z_0$ and $g(1) = \tilde f(1)$. Concatenating $\tilde f$ and g^{-1} gives a loop $g^{-1} * \tilde f$ in Y with $q \circ (g^{-1} * \tilde f) = f * \mathrm{const}_{z_0}$. Since f is homotopic to $f * \mathrm{const}_{z_0}$, this means that $[f]$ comes from $\pi_1(Y, y_0)$. Finally, let $y \in F$. Then $[y] = [y_0] \in \pi_0(Y)$ if and only if there is a continuous map $\tilde f\colon [0,1] \to Y$ with $\tilde f(0) = y_0$ and $\tilde f(1) = y$. Now $q \circ \tilde f$ is a loop in Z because $q(y) = q(y_0)$, and $y = \delta([f])$ by construction. This finishes the proof of the exact sequence (4.2).

This exact sequence exists for any map with a suitable lifting property, the *homotopy lifting property*. Topologists call such maps *fibrations*. Ehresmann's Fibration Theorem says that proper surjective submersions between smooth manifolds are locally trivial fibrations, that is, they have the appropriate homotopy lifting property. The map is proper if preimages of compact subsets are again compact.

Now we are going to use the exact sequence (4.2) for suitable surjective submersions to compute the fundamental group of $\mathrm{SU}(n)$. The unit sphere in $\mathbb{C}^n$ is a $2n-1$-dimensional sphere $\mathbb{S}^{2n-1}$. The group $\mathrm{SU}(n)$ acts transitively on it, and the stabiliser of the point $e_1 = (1, 0, \dots, 0)$ is a subgroup isomorphic to $\mathrm{SU}(n-1)$. Hence (4.2) gives an exact sequence

$$\pi_1(\mathrm{SU}(n-1), 1) \to \pi_1(\mathrm{SU}(n), 1) \to \pi_1(\mathbb{S}^{2n-1}, e_1) \to \pi_0(\mathrm{SU}(n-1)).$$

Now we recall that the spheres $\mathbb{S}^k$ for $k \geq 2$ are simply connected. And $\mathrm{SU}(1)$ is the trivial group, hence simply connected. Since $2n - 1 \geq 2$ for $n \geq 2$, the exact sequence above implies by induction that $\pi_1(\mathrm{SU}(n), 1)$ is the trivial group for all $n \geq 1$. This finishes the proof. □

Remark 4.7. The group $\mathrm{U}(n)$ is not simply connected. The group extension $\mathrm{SU}(n) \hookrightarrow \mathrm{U}(n) \twoheadrightarrow \mathrm{U}(1)$ splits by a Lie group homomorphism. So $\mathrm{U}(n)$ is diffeomorphic to $\mathrm{SU}(n) \times \mathbb{S}^1$. Hence its fundamental group is isomorphic to $\mathbb{Z}$. This is why we were careful to work with $\mathrm{SU}(n)$ in the proofs above.

Now we come to the dimension $d = 2$, which is already physically relevant. Here we need the higher-dimensional homotopy group $\pi_2(\mathrm{Gr}(r, n))$. In general, the k-th *homotopy group* $\pi_k(Y, y_0)$ of a pointed topological space (Y, y_0) is the group of homotopy classes of base-point preserving maps $(\mathbb{S}^k, p) \to (Y, y_0)$, where it is understood that homotopies must also preserve base points. Briefly,

$$\pi_k(Y, y_0) \coloneqq [(\mathbb{S}^k, p), (Y, y_0)].$$

This reduces to our definition of the fundamental group $\pi_1(Y, y_0)$ for $k = 1$ and also gives the set $\pi_0(Y)$ for $k = 0$ when we understand that $\mathbb{S}^0 = \{\pm 1\}$ has two points, one of which is the base point. If $k \geq 2$, then $\pi_k(Y, y_0)$ is an Abelian group. Given a fibration $q\colon Y \to Z$ with fibre F, the exact sequence (4.2) extends to a long exact sequence

$$\begin{aligned} \cdots \to \pi_k(F, y_0) \to \pi_k(Y, y_0) \to \pi_k(Z, z_0) \\ \to \pi_{k-1}(F, y_0) \to \pi_{k-1}(Y, y_0) \to \pi_{k-1}(Z, z_0) \to \cdots . \end{aligned} \tag{4.3}$$

The homotopy groups of a contractible space vanish. In particular, $\mathbb{R}^n$ for $n \in \mathbb{N}$ has trivial homotopy groups.

The group $\pi_k(\mathbb{S}^n)$ is trivial for $n > k$ and is isomorphic to $\mathbb{Z}$ for $k = n$. Of course, the higher homotopy groups are trivial for discrete spaces, so that $\pi_k(\mathbb{S}^0) = 0$ for $k \geq 1$. The exponential map gives a locally trivial fibration $\mathbb{R} \to \mathbb{S}^1$ with discrete fibre. The long exact sequence (4.3) applied to this fibration shows that $\pi_k(\mathbb{S}^1) = 0$ for all $k > 1$ because $\mathbb{R}$ is contractible, so that $\pi_k(\mathbb{R}) = 0$ for all $k \in \mathbb{N}$.

For $k > n > 1$, the homotopy groups of spheres are quite complicated and need not be 0. Much work in algebraic topology has been put into understanding these groups. Besides the computations above, we mention one more computation in the case $k > n$, as another application of the long exact sequence (4.3). The group $\mathrm{SU}(2)$ is diffeomorphic to the sphere $\mathbb{S}^3$. The matrices that are diagonal in the standard basis give a subgroup that is diffeomorphic to $\mathbb{S}^1$, and the quotient group is the complex projective space $\mathbb{CP}^1$, which is diffeomorphic to the sphere $\mathbb{S}^2$. So we get a locally trivial fibration $\mathbb{S}^3 \to \mathbb{S}^2$ with fibre $\mathbb{S}^1$, called the *Hopf fibration*. Since $\pi_k(\mathbb{S}^1) = 0$ for $k \geq 2$, the long exact sequence (4.3) for the Hopf fibration implies $\pi_k(\mathbb{S}^3) \cong \pi_k(\mathbb{S}^2)$ for all $k \geq 3$. In particular,

$$\pi_3(\mathbb{S}^2) \cong \pi_3(\mathbb{S}^3) \cong \mathbb{Z}$$

is non-trivial. There are canonical maps $\pi_{n+k}(\mathbb{S}^n) \to \pi_{n+1+k}(\mathbb{S}^{n+1})$ for all k, n, *Freudenthal's suspension maps*. Freudenthal's theorem says that the suspension map is an isomorphism for $n > k + 1$. The common values of these homotopy groups are called the *stable homotopy groups* of the spheres. In particular, $\pi_{n+1}(\mathbb{S}^n) \cong \pi_4(\mathbb{S}^3)$ for $n \geq 3$. The suspension map $\pi_3(\mathbb{S}^2) \to \pi_4(\mathbb{S}^3)$ is not an isomorphism. It turns out that it is surjective and that its kernel is the subgroup generated by 2. So $\pi_{n+1}(\mathbb{S}^n) \cong \mathbb{Z}/2$ for $n \geq 3$. The last claims are proven in [17], which computes the homotopy groups $\pi_{n+k}(\mathbb{S}^n)$ for $k \leq 19$.

Next we study the homotopy groups of $\mathrm{SU}(n)$. The long exact sequence (4.3) applied to the fibration $\mathrm{SU}(n) \to \mathbb{S}^{2n-1}$ with fibre $\mathrm{SU}(n-1)$ shows $\pi_k(\mathrm{SU}(n)) \cong \pi_k(\mathrm{SU}(n-1))$ for $k \leq 2n - 3$. Equivalently,

$$\pi_k(\mathrm{SU}(n)) \cong \pi_k(\mathrm{SU}(\lceil (k+1)/2 \rceil)) \tag{4.4}$$

for all $n \geq \lceil (k+1)/2 \rceil$. The group $\mathrm{SU}(2)$ is diffeomorphic to the sphere $\mathbb{S}^3$. Hence

$$\pi_2(\mathrm{SU}(n)) = 0, \qquad \pi_3(\mathrm{SU}(n)) \cong \mathbb{Z}$$

for all $n \geq 2$ by (4.4). If $k = 4, 5$, then (4.4) implies $\pi_k(\mathrm{SU}(n)) \cong \pi_k(\mathrm{SU}(3))$ for $n \geq 3$. And (4.3) for our standard fibration for $n = 3$ contains the exact sequence

$$\begin{aligned} \cdots \to \pi_6(\mathbb{S}^5) \to \pi_5(\mathrm{SU}(2)) \to \pi_5(\mathrm{SU}(3)) \to \pi_5(\mathbb{S}^5) \\ \to \pi_4(\mathrm{SU}(2)) \to \pi_4(\mathrm{SU}(3)) \to \pi_4(\mathbb{S}^5) \to \cdots . \end{aligned}$$

Since $\mathrm{SU}(2) \cong \mathbb{S}^3$, the results on homotopy groups of spheres mentioned above turn this into an exact sequence

$$\cdots \to \mathbb{Z}/2 \to \pi_5(\mathbb{S}^3) \to \pi_5(\mathrm{SU}(3)) \to \mathbb{Z} \to \mathbb{Z}/2 \to \pi_4(\mathrm{SU}(3)) \to 0 \to \cdots .$$

As it turns out, the boundary map $\mathbb{Z}/2 \to \pi_5(\mathbb{S}^3)$ here is an isomorphism, and the boundary map $\mathbb{Z} \to \mathbb{Z}/2$ is non-zero. Hence $\pi_4(\mathrm{SU}(3)) = 0$ and $\pi_5(\mathrm{SU}(3)) \cong \mathbb{Z}$, and this implies

$$\pi_4(\mathrm{SU}(n)) = 0, \qquad \pi_5(\mathrm{SU}(n)) \cong \mathbb{Z} \qquad \text{for all } n \geq 3.$$

And $\pi_4(\mathrm{SU}(2)) \cong \mathbb{Z}/2$ and $\pi_5(\mathrm{SU}(2)) \cong \mathbb{Z}/2$. An important phenomenon is *Bott periodicity*, which asserts that

$$\pi_k(\mathrm{U}(n)) \cong \begin{cases} 0 & \text{if } k \text{ is even,} \\ \mathbb{Z} & \text{if } k \text{ is odd,} \end{cases}$$

for $k \in \mathbb{N}$ and $n \geq \lceil (k+1)/2 \rceil$. Further computations of homotopy groups of $\mathrm{SU}(n)$ may be found in [10].

Now we turn to the homotopy groups of $\mathrm{Gr}(r,n)$. This space is a point if $r = 0$ or $r = n$, so that we may ignore these two cases and assume $1 \leq r \leq n-1$. We want to use the fibration $\mathrm{SU}(n) \to \mathrm{Gr}(r,n)$, which has the fibre $G \coloneqq \mathrm{U}(r) \times \mathrm{U}(n-r) \cap \mathrm{SU}(n)$. There is an obvious group extension $\mathrm{SU}(r) \times \mathrm{SU}(n-r) \hookrightarrow G \twoheadrightarrow \mathrm{U}(1)$, which splits by a group homomorphism. Hence G is diffeomorphic to the product

$$G \cong \mathrm{SU}(r) \times \mathrm{SU}(n-r) \times \mathbb{S}^1.$$

The homotopy groups are compatible with products, so that

$$\pi_k(G) \cong \pi_k(\mathrm{SU}(r)) \times \pi_k(\mathrm{SU}(n-r)) \times \pi_k(\mathbb{S}^1).$$

In particular, $\pi_1(G) \cong \mathbb{Z}$ because $\mathrm{SU}(n)$ is simply connected for all $n \geq 1$ by Lemma 4.6. And the factor $\pi_k(\mathbb{S}^1)$ vanishes for $k \geq 2$. Recall that the groups $\pi_k(\mathrm{SU}(n))$ do not depend on n any more for $n \geq \lceil (k+1)/2 \rceil$. The isomorphism is induced by the inclusion map, and this is also the map that appears in the long exact sequence for the fibration $\mathrm{SU}(n) \to \mathrm{Gr}(r,n)$. Hence the groups $\pi_k(\mathrm{SU}(n-r))$ and $\pi_k(\mathrm{SU}(n))$ in the long exact sequence cancel if $n - r \geq \lceil (k+1)/2 \rceil$. Therefore, the boundary map induces an isomorphism

$$\pi_k(\mathrm{Gr}(r,n)) \cong \pi_{k-1}(\mathrm{SU}(r)) \tag{4.5}$$

if $n - r \geq \lceil (k+1)/2 \rceil$ and $k \geq 3$, and an isomorphism

$$\pi_2(\mathrm{Gr}(r,n)) \cong \pi_1(\mathrm{SU}(r)) \times \mathbb{Z} \cong \mathbb{Z} \tag{4.6}$$

if $k = 2$ and $n - r \geq 2$. In particular, $\pi_3(\mathrm{Gr}(r,n)) = 0$ for $1 \leq r \leq n-2$ because $\pi_2(\mathrm{SU}(n)) = 0$ for all $n \geq 2$. Taking orthogonal complements is a diffeomorphism $\mathrm{Gr}(r,n) \cong \mathrm{Gr}(n-r,n)$. Hence we get

$$\pi_2(\mathrm{Gr}(r,n)) \cong \mathbb{Z}, \qquad \pi_3(\mathrm{Gr}(r,n)) \cong 0$$

for $1 \leq r \leq n-1$ provided $n > 2$. If $n = 2$ and $r = 1$, then $\mathrm{Gr}(1,2) = \mathbb{CP}^1$ is diffeomorphic to the sphere $\mathbb{S}^2$. So $\pi_2(\mathrm{Gr}(1,2)) \cong \mathbb{Z}$ and $\pi_3(\mathrm{Gr}(1,2)) \cong \mathbb{Z}$. The last isomorphism is

exceptional because $\pi_3(\mathrm{Gr}(r,n)) \cong 0$ if $n \geq 3$ and $0 \leq r \leq n$. Bott periodicity implies that

$$\pi_k(\mathrm{Gr}(r,n)) \cong \begin{cases} \mathbb{Z} & \text{if } k \text{ is even,} \\ 0 & \text{if } k \text{ is odd,} \end{cases}$$

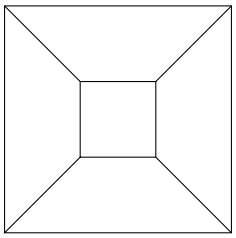
for $k \in \mathbb{N}$ and $r, n-r \geq \lceil (k+1)/2 \rceil$.

Next we compute $[\mathbb{T}^2, \mathrm{Gr}(r,n)]$ using the low-dimensional homotopy groups of $\mathrm{Gr}(r,n)$. We continue to assume $1 \leq r \leq n-1$ to rule out trivial cases.

Lemma 4.8. *Let Y be a simply connected topological space. Then there is a canonical bijection $[\mathbb{T}^2, Y] \cong \pi_2(Y)$.*

Proof. We describe $\mathbb{T}^2 = [0,1]^2/\!\sim_{\mathbb{T}}$ and $\mathbb{S}^2 = [0,1]^2/\!\sim_{\partial}$ and where $\sim_{\mathbb{T}}$ identifies $(0,t) \sim_{\mathbb{T}} (1,t)$ and $(s,0) \sim_{\mathbb{T}} (s,1)$ for all $s,t \in [0,1]$, whereas $\sim_{\partial}$ identifies all points on the boundary $\partial[0,1]^2$ with each other. So the identity map on $[0,1]^2$ induces a quotient map $\mathbb{T}^2 \to \mathbb{S}^2$, which induces a map $\pi_2(Y) \coloneqq [(\mathbb{S}^2,0),(Y,y_0)] \to [(\mathbb{T}^2,0),(Y,y_0)]$. Here we use the equivalence class of 0 as the base point in $\mathbb{S}^2$ and $\mathbb{T}^2$. Forgetting to preserve base points defines a forgetful map $[(\mathbb{T}^2,0),(Y,y_0)] \to [\mathbb{T}^2,Y]$. We claim that the composite map $\pi_2(Y) \to [\mathbb{T}^2,Y]$ is a bijection. This is a higher-dimensional analogue of the computation of $[\mathbb{T}^1,Y]$ in (4.1), under the simplifying assumption that the target space be simply connected.

Let $y_0 \in Y$ be a base point and let $f\colon [0,1]^2/\!\sim_{\mathbb{T}} \to Y$ be a continuous map. Then $f(0,0) = f(0,1) = f(1,0) = f(1,1)$. Since Y is connected, we may choose a path g from y_0 to this point. Concatenating the paths g, $f(t,0) = f(t,1)$ for $t \in [0,1]$, and g^{-1} gives a loop at y_0. Since Y is simply connected, this loop is homotopic to the constant loop at y_0. Hence there is a continuous map $h_1\colon [0,1]^2 \to Y$ with $h_1(0,t) = y_0$, $h_1(t,0) = h_1(t,1) = g(t)$ and $h_1(1,t) = f(t,0) = f(t,1)$ for all $t \in [0,1]$. Similarly, there is a continuous map $h_2\colon [0,1]^2 \to Y$ with $h_2(0,t) = y_0$, $h_2(t,0) = h_2(t,1) = g(t)$ and $h_2(1,t) = f(0,t) = f(1,t)$ for all $t \in [0,1]$. We may use these maps to define a continuous map on an enlarged square $[-1,2]^2$, decomposed as follows:



Here each quadrilateral is identified with the square by an affine map, and we use the maps h_1 on the left and right, h_2 on the top and bottom quadrilateral, and f in the square in the middle. The resulting function is constant equal to y_0 on the boundary of $[-1,2]^2$. Scaling gives a homotopy of maps $\mathbb{T}^2 \to Y$ from the original map f to the map on $[-1,2]^2$. The latter maps $\partial[0,1]^2$ to y_0, so it defines an element in $\pi_2(Y)$. Hence the map $\pi_2(Y) \to [\mathbb{T}^2,Y]$ is surjective.

To prove that this map is injective, take two maps $f_0, f_1 \colon [0,1]^2/{\sim_\partial} \to Y$ mapping all boundary points to y_0 and a homotopy between them as maps $h \colon [0,1]^2/{\sim_{\mathbb{T}}} \to Y$. This homotopy is a map defined on $[0,1]^3$. An argument as above gives a homotopy between h and a map $\tilde{h}$ on $[0,1]^3$ that is constant equal to y_0 on each one-dimensional face of $\partial[0,1]^3$. That is, the restrictions of $\tilde{h}$ to the six faces of the cube are elements of $\pi_2(Y)$. The continuous map $\tilde{h}$ witnesses that the oriented sum of these restrictions of $\tilde{h}$ vanishes in $\pi_2(Y)$. The new homotopy $\tilde{h}$ is still defined on $\mathbb{T}^2 \times [0,1]$. So its restrictions to the front and back faces and to the left and right faces are equal, respectively, and they cancel in the oriented sum in $\pi_2(Y)$. And the restrictions to the top and bottom faces are the given maps f_0 and f_1. Hence $[f_0] = [f_1]$ in $\pi_2(Y)$ as desired. □

Putting our computations together gives

$$[\mathbb{T}^2, \mathrm{Gr}(n,r)] \cong \mathbb{Z}$$

for all $n \in \mathbb{N}$ and $1 \leq r \leq n-1$. Thus there are non-trivial topological phases in dimension 2.

A similar computation would show $[\mathbb{T}^3, Y] \cong \pi_3(Y)$ if $\pi_k(Y) = 0$ for $k = 0, 1, 2$. But the spaces $\mathrm{Gr}(n,r)$ that we care about have non-trivial π_2 and (usually) trivial π_3. Hence we use a different idea to study $[\mathbb{T}^3, \mathrm{Gr}(n,r)]$. A map $\mathbb{T}^3 \to Y$ is equivalent to a map $[0,1]^3/{\sim_{\mathbb{T}}} \to Y$, where $\sim_{\mathbb{T}}$ describes periodic boundary conditions, $(0,s,t) \sim_{\mathbb{T}} (1,s,t)$, $(s,0,t) \sim_{\mathbb{T}} (s,1,t)$, $(s,t,0) \sim_{\mathbb{T}} (s,t,1)$ for all $s,t \in [0,1]$. If Y is simply connected, then we may replace any map $\mathbb{T}^3 \to Y$ by one that is constant equal to y_0 on all one-dimensional faces of the boundary $\partial[0,1]^3$, and we may restrict to homotopies with this property as well.

Lemma 4.9. *Let Y be a simply connected space with $\pi_3(Y) = 0$. A continuous map $\partial[0,1]^3 \to Y$ defines an element in $\pi_2(Y)$, and it extends to a continuous map $[0,1]^3 \to Y$ if and only if this class is 0. Any two such extensions are homotopic with a homotopy that is constant on the boundary.*

Proof. The cube $[0,1]^3$ is homeomorphic to the closed disk

$$D^3 = \{(x,y,z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 \leq 1\}$$

in $\mathbb{R}^3$, such that the boundary $\partial[0,1]^3$ is mapped onto the boundary ∂D^3 of D^3. One way to prove this is to shift and rescale the cube to map it onto the unit ball of the ∞-norm, which is a cube of side length 2 centred at 0. The map $x \mapsto \|x\|_2 \cdot \|x\|_\infty^{-1} \cdot x$ is a homeomorphism from the unit ball of the ∞-norm onto D^3, which is the unit ball of the 2-norm. The diffeomorphism above identifies the boundary $\partial[0,1]^3$ with $\mathbb{S}^2$. Thus a continuous map $\partial[0,1]^3 \to Y$ defines a map $f \colon \mathbb{S}^2 \to Y$. Since Y is path connected, we may modify this map in a little neighbourhood of the base point $*$ in $\mathbb{S}^2$ so that it becomes base-point preserving. This modification depends on the choice of a path between $f(*)$ and the base point in Y. Since Y is simply connected, the choice does not influence the resulting class in $\pi_2(Y)$.

A map on $\mathbb{S}^2$ extends to a map on D^3 if and only if it is homotopic to the constant map. So a map on $\partial[0,1]^3$ extends to $[0,1]^3$ if and only if the associated class in $\pi_2(Y)$ vanishes. Now consider two extensions $f_\pm\colon D^3 \to Y$. We may build a 3-sphere $\mathbb{S}^3$ by glueing two copies of D^3 along their boundaries $\mathbb{S}^2$. The two copies of D^3 become the northern and southern hemispheres and $\mathbb{S}^2$ becomes the equator in $\mathbb{S}^3$. So the two maps $f_\pm$ that coincide on the boundary combine to a map $\mathbb{S}^3 \to Y$. Since $\pi_3(Y) = 0$, this map extends to a continuous map on the closed unit disk $D^4 \subseteq \mathbb{R}^4$. Now the embeddings of D^3 into D^4 as the northern and southern hemispheres are homotopic through a homotopy that is constant on the boundary sphere $\mathbb{S}^2$. A particular homotopy that does it has the form

$$D^3 \to D^4, \qquad (x,y,z) \mapsto \left(x,y,z,t\cdot\sqrt{1-x^2-y^2-z^2}\right),$$

for $-1 \le t \le 1$. Hence a continuous map on D^4 that agrees with the maps $f_\pm$ on the two boundary hemispheres gives a homotopy between the two maps $f_\pm$ that is constant on the boundary $\mathbb{S}^2$. □

Proposition 4.10. *Let Y be a simply connected topological space with $\pi_3(Y) = 0$. Then there is a bijection between $[\mathbb{T}^3, Y]$ and the product $\pi_2(Y)^3$ of three copies of $\pi_2(Y)$.*

Proof. The three embeddings $\mathbb{R}^2 \to \mathbb{R}^3$ as the xy-, xz- and yz-planes induce three embeddings of the torus $\mathbb{T}^2$ into $\mathbb{T}^3$. We describe $\mathbb{T}^3$ as $[0,1]^3/\!\sim_{\mathbb{T}}$, where the equivalence relation $\sim_{\mathbb{T}}$ identifies opposite faces in the boundary $\partial[0,1]^3$. So the six boundary faces of the cube $[0,1]^3$ are only three different faces. And these are the images of the three embedded tori $\mathbb{T}^2 \subseteq \mathbb{T}^3$. Restricting a map on $\mathbb{T}^3$ to one of the embedded tori gives a map $[\mathbb{T}^3, Y] \to [\mathbb{T}^2, Y]$. Lemma 4.8 identifies the target of this map with $\pi_2(Y)$. Therefore, the three maps $[\mathbb{T}^3, Y] \to [\mathbb{T}^2, Y]$ combine to a map $[\mathbb{T}^3, Y] \to \pi_2(Y)^3$. We claim that this map is bijective. To begin with, three elements in $\pi_2(Y)$ may be represented by three maps $[0,1]^2 \to Y$ that are constant equal to the base point y_0 of Y on $\partial[0,1]^2$. We take the first of these maps on the front and back faces of the cube, the second on the top and bottom faces, and the third on the left and right faces of the cube. This gives a well defined map because all the three maps are constant equal to y_0 on the boundaries of the faces, where different faces meet. So we get a well defined map on $\partial[0,1]^3$. The map in $\pi_2(Y)$ associated to this map is an alternating sum of the maps on the six boundary faces, where opposite faces appear with different signs. Since we have used the same map on opposite faces, the class in $\pi_2(Y)$ always vanishes. By Lemma 4.9, we may extend our map on $\partial[0,1]^3$ to $[0,1]^3$. The extension preserves the equivalence relation $\sim_{\mathbb{T}}$, which only concerns the boundary values. So it defines a map on $\mathbb{T}^3$. Thus the map $[\mathbb{T}^3, Y] \to \pi_2(Y)^3$ is surjective. In addition, the uniqueness of the extension stated in Lemma 4.9 shows that this map is injective. □

The formulas for $\pi_3(\mathrm{Gr}(r,n))$ show that Proposition 4.10 applies to $\mathrm{Gr}(r,n)$ for $n \ge 3$ and $0 \le r \le n$, but not for $(r,n) = (1,2)$. So $[\mathbb{T}^3, \mathrm{Gr}(r,n)] = \mathbb{Z}^3$ if $1 \le r \le n-1$ and $3 \le n$. I think that $\pi_3(\mathrm{Gr}(1,2)) \cong \mathbb{Z}$ implies that $[\mathbb{T}^3, \mathrm{Gr}(1,2)] = \mathbb{Z}^4$.

5 Hamiltonians with symmetry

Our classification of Hamiltonians so far only takes into account translation symmetry. Now we assume that the system has some other symmetries and restrict homotopies accordingly. This leads to a very different classification. We first define the relevant notion of homotopy. Then we specialise to the simplest case of chiral symmetry. Other types of symmetry lead to equivariant as opposed to ordinary homotopy theory. We only touch upon this briefly here.

Assume that a group G is represented projectively on the Hilbert space $\mathcal{K}$ by unitary or anti-unitary operators. These induce unitary or anti-unitary operators on $L^2(\mathbb{T}^d, \mathcal{K})$, which again form a projective representation of G. Let $c\colon G \to \{\pm 1\}$ be a group homomorphism. We restrict attention to Hamiltonians that satisfy $gHg^{-1} = c(g)H$, that is, H commutes with elements of G with $c(g) = 1$ and anticommutes with elements of G with $c(g) = -1$.

I am not aware of a mathematically precise treatment of homotopy for Hamiltonians on an infinite-dimensional Hilbert space $\mathcal{K}$ and with chiral or particle-hole symmetries. The assumption that Hamiltonians are bounded below forbids this case.

Definition 5.1. Let H_0, H_1 be invertible, self-adjoint, bounded below operators on $L^2(\mathbb{T}^d, \mathcal{K})$ with compact resolvent that satisfy $gH_tg^{-1} = c(g)H_t$ for all $g \in G$ and $t = 0, 1$. A (*G-equivariant*) *homotopy* between H_0 and H_1 is a family of invertible self-adjoint operators $(H_t)_{t\in(0,1)}$ with compact resolvent that satisfy $gH_tg^{-1} = c(g)H_t$ for all $t \in (0, 1)$, such that the resolvents $(\lambda - H_t)^{-1}$ define a norm-continuous map $[0, 1] \to \mathbb{K}(L^2(\mathbb{T}^d, \mathcal{K}))$. If such a homotopy exists, we call H_0 and H_1 (*G-equivariantly*) *homotopic*.

Roughly speaking, if two Hamiltonians have some symmetries, we may restrict attention to homotopies consisting of Hamiltonians with the same symmetries. Which of the symmetries we require for homotopies is a choice. If perturbations of the Hamiltonian that break some particular symmetries are relevant, then we may disregard these symmetries. For instance, a magnetic field usually breaks time-reversal symmetry. Therefore, if your Hamiltonian is time-reversal symmetric, then we may choose to disregard this in the homotopy classification if external magnetic fields occur in experiments.

The following lemma is a variant of Theorem 4.2:

Lemma 5.2. *Let $\dim \mathcal{K} < \infty$ and let H_0, H_1 be invertible, self-adjoint operators on $L^2(\mathbb{T}^d, \mathcal{K})$ that satisfy $gH_tg^{-1} = c(g)H_t$ for all $g \in G$ and $t = 0, 1$. Let $F_t \coloneqq \operatorname{sign}(H_t)$. This a self-adjoint operator with $F_t^2 = 1$ ("self-adjoint involution") that satisfies $gF_tg^{-1} = c(g)F_t$ for all $g \in G$ and $t = 0, 1$. The operators H_0, H_1 are G-equivariantly homotopic among invertible self-adjoint operators on $L^2(\mathbb{T}^d, \mathcal{K})$ if and only if F_0, F_1 are G-equivariantly homotopic among self-adjoint involutions on $L^2(\mathbb{T}^d, \mathcal{K})$.*

Proof. The proof is essentially the same as for Theorem 4.2. The only new point is that $\operatorname{sign}(H_t)$ (anti)commutes with any operator that (anti)commutes with H_t because the sign function is odd. In contrast, the spectral projections used in Theorem 4.2 do not anticommute with operators that anticommute with H_t. □

5.1 Chiral symmetry

As a first case of symmetry, we consider Hamiltonians with a chiral symmetry. That is, we are given a unitary involution $\Xi\colon \mathcal{K} \to \mathcal{K}$, and we restrict attention to Hamiltonians that anti-commute with Ξ. In the general scheme above, this corresponds to the case where $G = \{\pm 1\}$ acts by unitary operators and $c\colon G \to \{\pm 1\}$ is the identity map. We assume that $\mathcal{K}$ is finite-dimensional.

The unitary involution Ξ has spectrum $\{\pm 1\}$. So it induces a decomposition $\mathcal{K} = \mathcal{K}_+ \oplus \mathcal{K}_-$ such that Ξ restricts to ± 1 on $\mathcal{K}_\pm$. Conversely, such a direct sum decomposition defines a unique self-adjoint involution. A self-adjoint operator that anticommutes with Ξ is a block matrix of the form

$$H = \begin{pmatrix} 0 & T^* \\ T & 0 \end{pmatrix}$$

in the decomposition $\mathcal{K} = \mathcal{K}_+ \oplus \mathcal{K}_-$, with some operator $T\colon \mathcal{K}_+ \to \mathcal{K}_-$. Thus a matrix-valued function $H\colon \mathbb{T}^d \to \mathbb{K}(\mathcal{K})$ that anti-commutes with $\mathrm{id} \otimes \Xi$ is equivalent to a matrix-valued function $H\colon \mathbb{T}^d \to \mathbb{K}(\mathcal{K}_+, \mathcal{K}_-)$. By spectral flattening (Lemma 5.2), we may replace invertible self-adjoint operators by self-adjoint unitaries (involutions). The operator H is unitary if and only if $H^2 = 1$, if and only if the corresponding operator T is unitary. So homotopy classes of invertible Hamiltonians in dimension d with chiral symmetry Ξ are naturally in bijection with the set $[\mathbb{T}^d, \mathrm{U}(\mathcal{K}_+, \mathcal{K}_-)]$ of homotopy classes of maps from $\mathbb{T}^d$ to the set of unitary operators $\mathcal{K}_+ \to \mathcal{K}_-$.

This set is empty unless $\dim \mathcal{K}_+ = \dim \mathcal{K}_-$. So we assume this and call the common dimension n. Then there are unitaries $\mathbb{C}^n \cong \mathcal{K}_+$ and $\mathbb{C}^n \cong \mathcal{K}_-$. Using these, we may identify $\mathrm{U}(\mathcal{K}_+, \mathcal{K}_-) \cong \mathrm{U}(n)$. So we must study the space $[\mathbb{T}^d, \mathrm{U}(n)]$. Fortunately, we have already done this when studying the case without symmetries. Recall that $\mathrm{U}(n) \cong \mathrm{SU}(n) \times \mathbb{S}^1$. So

$$[\mathbb{T}^d, \mathrm{U}(n)] \cong [\mathbb{T}^d, \mathrm{SU}(n)] \times [\mathbb{T}^d, \mathbb{S}^1].$$

We have seen in Section 4.2 that $\mathrm{SU}(n)$ is simply connected and that $\pi_2(\mathrm{SU}(n)) = 0$ and $\pi_3(\mathrm{SU}(n)) = \mathbb{Z}$ for all $n \geq 2$. Hence Lemma 4.8 implies that $[\mathbb{T}^2, \mathrm{SU}(n)]$ is a single point. And an argument as in the proof of Lemma 4.8 implies $[\mathbb{T}^3, \mathrm{SU}(n)] \cong \pi_3(\mathrm{SU}(n)) \cong \mathbb{Z}$ for $n \geq 2$. We are, in fact, more interested in the second factor $[\mathbb{T}^d, \mathbb{S}^1]$. This is the only piece if $n = 1$ because then $\mathrm{SU}(1) = \{1\}$ and $\mathrm{U}(1) = \mathbb{S}^1$.

Lemma 5.3. *There is a natural bijection between $[\mathbb{T}^d, \mathbb{S}^1]$ and the set of group homomorphisms $\mathbb{Z}^d \to \mathbb{Z}$.*

Proof. We first pick base points $x_0 \in \mathbb{T}^d$, $y_0 \in \mathbb{S}^1$ at will and restrict attention to base-point preserving maps and homotopies. A base-point preserving continuous map $X \to Y$ induces a homomorphism between the fundamental groups $\pi_1(X) \to \pi_1(Y)$, and homotopic maps induce the same homomorphism of fundamental groups. So we get a canonical map from $[(\mathbb{T}^d, x_0), (\mathbb{S}^1, y_0)]$ to $\mathrm{Hom}(\pi_1(\mathbb{T}^d), \pi_1(\mathbb{S}^1))$. Now recall that $\pi_1(\mathbb{T}^d) = \mathbb{Z}^d$, $\pi_1(\mathbb{S}^1) = \mathbb{Z}$. The spaces $\mathbb{T}^d$ have the special property that their universal coverings are contractible. Namely, these are the quotient maps $q_d\colon \mathbb{R}^d \to \mathbb{R}^d/\mathbb{Z}^d \cong \mathbb{T}^d$.

Since $\mathbb{T}^1 = \mathbb{S}^1$, the same is true for $\mathbb{S}^1$. The theory of covering spaces implies that two continuous maps between aspherical spaces are homotopic if and only if they induce the same maps between the fundamental groups. A crucial step here is a local lifting construction, showing that a map $\mathbb{T}^d \to \mathbb{S}^1$ lifts to a map $\mathbb{R}^d \to \mathbb{R}$, which is unique if we specify the value at one point of $\mathbb{R}^d$.

Now it remains to study what happens when we allow maps that do not preserve base points. Since $\mathbb{T}^d$ and $\mathbb{S}^1$ are connected, any map $\mathbb{T}^d \to \mathbb{S}^1$ is homotopic to a map that sends x_0 to y_0; the homotopy changes the map only on a very small neighbourhood of y_0, which we take to be homeomorphic to a closed ball in $\mathbb{R}^d$ and so that the image is contained in a very small, contractible open subset of $\mathbb{S}^1$. A homotopy between base-point preserving maps is, in particular, a base-point preserving map $\mathbb{T}^d \times [0,1] \to \mathbb{S}^1$, where we may take the base point in $\mathbb{T}^d \times [0,1]$ to be either $(x_0, 0)$ or $(x_0, 1)$. The fundamental groups for these two base points are the same, and the induced maps $\mathbb{Z}^d \to \mathbb{Z}$ on fundamental groups are also the same. So homotopic maps induce the same map on fundamental groups, even if we allow homotopies that do not preserve base points. □

Of course, $\mathrm{Hom}(\mathbb{Z}^d, \mathbb{Z}) \cong \mathbb{Z}^d$ by specifying the values of a homomorphism on the generators of $\mathbb{Z}^d$. So

$$[\mathbb{T}^1, \mathrm{U}(n)] \cong \mathbb{Z}, \qquad [\mathbb{T}^2, \mathrm{U}(n)] \cong \mathbb{Z}^2, \qquad [\mathbb{T}^3, \mathrm{U}(n)] \cong \mathbb{Z}^4.$$

In particular, non-trivial topological phases for Hamiltonians with chiral symmetry exist already for $d = 1$. Indeed, the Hamiltonians of the Su–Schrieffer–Heeger model studied in Section 3.3.5 for $m \in (-1, 1)$ and $m > 1$ are not homotopic. Let H_m denote this Hamiltonian. Recall that H_m has chiral symmetry and that it is invertible for $|m| \neq 1$. So $(H_m)_{m \in [a,b]}$ is a homotopy between H_a and H_b provided $a, b \in (-1, 1)$ or $a, b \in (-\infty, -1)$ or $a, b \in (1, \infty)$. If, say, $a < 1 < b$, then the homotopy $(H_m)_{m \in [a,b]}$ is not allowed because H_1 is not invertible.

To see that there really are phase transitions at ± 1, we need to understand the bijection $[\mathbb{T}^1, \mathrm{U}(n)] \cong \mathbb{Z}$. The Su–Schrieffer–Heger model has $n = 1$, where $\mathrm{U}(1) = \mathbb{S}^1$. Actually, for $n \geq 2$, the first step in the bijection $[\mathbb{T}^1, \mathrm{U}(n)] \cong \mathbb{Z}$ is the canonical map

$$[\mathbb{T}^1, \mathrm{U}(n)] \xrightarrow{\cong} [\mathbb{T}^1, \mathrm{U}(1)]$$

that simply composes a loop in $\mathrm{U}(n)$ with the projection $\det \colon \mathrm{U}(n) \to \mathrm{U}(1) \subseteq \mathbb{C}$ with kernel $\mathrm{SU}(n)$. This produces a loop in $\mathrm{U}(1)$ even if we start in the case $n \geq 2$. The homotopy group computations above show that applying the determinant does not forget any information for loops. Now we identify $\mathbb{T}^1$ and $\mathrm{U}(1)$ with the unit circle in $\mathbb{C}$. The canonical bijection between $[\mathbb{T}^1, \mathrm{U}(1)]$ and $\mathbb{Z}$ maps a loop in $\mathrm{U}(1)$ to its *winding number*. This counts with appropriate signs how often a loop in $\mathrm{U}(1) \subseteq \mathbb{C} \setminus \{0\}$ winds around the origin. For the Su–Schrieffer–Heeger model, the loop in question is the spectral flattening of the loop

$$[0,1] \to \mathbb{C} \setminus \{0\}, \qquad k \mapsto \exp(2\pi \mathrm{i} k) + \mathrm{i} m;$$

here we have taken the appropriate entry of the matrix $H(k)$ in (3.6). We only get a loop in $\mathbb{C} \setminus \{0\}$, not in $\mathrm{U}(1)$ because we have not yet spectrally flattened the Hamiltonian.

Spectral flattening amounts to the radial projecting $\mathbb{C} \setminus \{0\} \to \mathrm{U}(1)$, which does not change the winding number. So the relevant loop is a circle around the point $\mathrm{i}m$ of radius 1, and we run through the circle with constant speed in the positive (anti-clockwise) direction. This loop has winding number $+1$ around all points in the interior of the circle, and winding number 0 around points outside the circle. The point 0 lies in the circle if $m \in (-1, 1)$, and outside the circle otherwise. So our bijection $[\mathbb{T}^1, \mathrm{U}(1)] \xrightarrow{\cong} \mathbb{Z}$ maps the Su–Schrieffer–Heeger Hamiltonian H_m to $+1$ if $m \in (-, 1, 1)$, and to 0 if $m < -1$ or $m > 1$. In particular, the Hamiltonians for $m < -1$ and $m > 1$ are homotopic. One homotopy between H_m and M_{-m} for $m > 1$ would be to take

$$L_t(k) \coloneqq \begin{pmatrix} 0 & S_t(k)^* \\ S_t(k) & 0 \end{pmatrix}, \qquad S_t(k) = \exp(2\pi \mathrm{i} k) + \mathrm{i}m \cdot \exp(\pi \mathrm{i} t)$$

for $t \in [0, 1]$. The function $S_t(k)$ describes a circle around $\mathrm{i}m \cdot \exp(\pi \mathrm{i} t)$ of radius 1, and all these circles avoid the point 0.

5.2 Non-Unitary symmetries and equivariant homotopies

Now assume that our symmetry group G also contains some non-unitary symmetries. This case is qualitatively different. Let $H \in \mathrm{C}(\mathbb{T}^d, \mathbb{B}(\mathcal{K}))$. We write the function H as a Fourier series $H(z) = \sum_{x \in \mathbb{Z}^d} z^x H_x$ with $z^x = \prod_{j=1}^d z_j^{x_j}$ and $H_x \in \mathbb{B}(\mathcal{K})$ for $x \in \mathbb{Z}^d$, $z \in \mathbb{T}^d$. The corresponding operator on $\ell^2(\mathbb{Z}^d) \otimes \mathcal{K}$ is $\sum_{x \in \mathbb{Z}^d} S_x \otimes H_x$ with $(S_x f)(y) \coloneqq f(y - x)$ for $x, y \in \mathbb{Z}^d$, $f \in \ell^2 Z^d$). Let $g \in G$ be an anti-unitary symmetry acting on $\mathcal{K}$; so the action on $\ell^2(\mathbb{Z}^d, \mathcal{K})$ is by pointwise application of G. So g commutes with the translations S_x for all $x \in \mathbb{Z}^d$. We write the commutation relation $Hg = c(g) \cdot gH$ as $H = c(g) \cdot gHg^{-1}$ and compute

$$c(g) g H g^{-1} = \sum_{x \in \mathbb{Z}^d} S_x \otimes c(g) g H_x g^{-1}.$$

So the function on T^d corresponding to $c(g) g H g^{-1}$ has the Fourier coefficients $c(g) g H_x g^{-1}$. Its value at $z \in \mathbb{T}^d$ is therefore

$$\sum_{x \in \mathbb{Z}^d} z^x \cdot c(g) g H_x g^{-1} = c(g) \cdot g \Big(\sum_{x \in \mathbb{Z}^d} \overline{z}^x H_x \Big) g^{-1} = c(g) g \cdot H(\overline{z}) \cdot g^{-1}.$$

We replace z by $\overline{z}$ because g is conjugate-linear, so that $g \cdot (\mathrm{i} \cdot T) = -\mathrm{i} \cdot g \cdot T$. Hence the assumption that the operator H on $\ell^2(\mathbb{Z}^d, \mathcal{K})$ satisfies $H = c(g) g H g^{-1}$ corresponds to the condition $H(z) = c(g) g H(\overline{z}) g^{-1}$ for the function $H \colon \mathbb{T}^d \to \mathbb{B}(\mathcal{K})$. When we describe $\mathbb{T}^d$ as $\mathbb{R}^d / \mathbb{Z}^d$, then the complex conjugation above becomes the map $k \mapsto -k$ or, equivalently, $k \mapsto 1 - k$.

Hence the space of Hamiltonians is the space of continuous maps

$$f \colon \mathbb{T}^d \to \{a \in \mathbb{B}(\mathcal{K}) : a = a^*,\ a^2 = 1\} \tag{5.1}$$

that satisfies $f(z) = c(g) g f(z) g^{-1}$ if $g \in G$ is unitary and $f(z) = c(g) g f(\overline{z}) g^{-1}$ if $g \in G$ is anti-unitary. This is a G-equivariant map when we let G act on $\mathbb{T}^d$ by $g \cdot z \coloneqq z$ if g is

unitary and $g \cdot z := \overline{z}$ if g is anti-unitary and on $\mathbb{B}(\mathcal{K})$ by $T \mapsto c(g) g T g^{-1}$. We may first treat the unitary symmetries by shrinking the target space in (5.1) to

$$X := \{a \in \mathbb{B}(\mathcal{K}) : a = a^*,\ a^2 = 1,\ \text{and } c(g) g a g^{-1} = a \text{ for all unitary } g \in G\}. \tag{5.2}$$

The subgroup of unitary symmetries in G has index 2 when we assume that there is at least one non-unitary symmetry. Let $g_0 \in G$ be a chosen anti-unitary symmetry. Then g_0^2 is a unitary symmetry in G, which acts trivially on X. Hence $\mathbb{T}^d$ and X carry actions of the group $\mathbb{Z}/2$ generated by complex conjugation and the map $T \mapsto c(g_0) g_0 T g_0^{-1}$, respectively. The map in (5.1) is G-equivariant if and only if it is a $\mathbb{Z}/2$-equivariant map $\mathbb{T}^d \to X$. The space of homotopy classes of $\mathbb{Z}/2$-equivariant maps $Y \to X$ with $\mathbb{Z}/2$-equivariant homotopies is denoted by $[Y, X]^{\mathbb{Z}/2}$. So our problem is to describe the set $[\mathbb{T}^d, X]^{\mathbb{Z}/2}$, where the space X and the action of $\mathbb{Z}/2$ on it depend on the dimension of $\mathcal{K}$ and the symmetry group G. This problem is handled by equivariant homotopy theory, which is more complicated than ordinary homotopy because we also have to take into account the group actions.

We briefly consider a simple case. Assume that $G = \{1, \Theta\}$ has only time-reversal symmetry, with $\Theta^2 = \pm 1$. In this case, all symmetries commute with H. Hence we may do spectral flattening as in the previous section, replacing involutions by projections. To make our problem more concrete, we have to represent Θ by an anti-unitary operator on $\mathcal{K}$. If $\Theta^2 = 1$, then complex conjugation will do, and any other anti-unitary operator on $\mathbb{C}^n$ with $\Theta^2 = 1$ is unitarily equivalent to complex conjugation. If $\Theta^2 = -1$, then multiplication with i and Θ generate an action of the quaternion algebra on $\mathcal{K}$, making $\mathcal{K}$ a vector space over the quaternions $\mathbb{H}$. So the dimension of $\mathcal{K}$ is even, and we may choose an isomorphism $\mathcal{K} \cong \mathbb{R}^k \otimes_{\mathbb{R}} \mathbb{C}^2$ such that Θ acts by $1 \otimes \Theta_0$ with Θ_0 as in Example 2.13.

Now assume also that $d = 1$. What is a $\mathbb{Z}/2$-equivariant map $f\colon \mathbb{T} \to \mathbb{B}(\mathcal{K})$? The map $z \mapsto \overline{z}$ on the unit circle in $\mathbb{C}$ has the two fixed points ± 1, and it maps the upper semicircle bijectively onto the lower semicircle. Hence $f(\pm 1)$ must commute with Θ, and the restriction of f to the upper semicircle is a homotopy between $f(1)$ and $f(-1)$, and the restriction of f to the lower semicircle is determined by this and the constraint $f(\overline{z}) = \Theta f(z) \Theta^{-1}$.

In the case $\Theta^2 = 1$, an operator commutes with the complex conjugation Θ if and only if its matrix has real entries. So $f(\pm 1)$ are projections in $\mathbb{M}_n(\mathbb{R})$, which describe points in the real Grassmanian $\mathrm{Gr}_{\mathbb{R}}(r, n)$, where r is the rank of the projection. In the case $\Theta^2 = -1$, an operator commutes with the complex conjugation Θ if and only if is linear over the quaternions. An $\mathbb{H}$-linear orthogonal projection is again equivalent to an $\mathbb{H}$-linear subspace, so that $f(\pm 1)$ are points in the quaternionic Grassmanian $\mathrm{Gr}_{\mathbb{H}}(r/2, n/2)$; here $r/2$ is the rank of the projection over $\mathbb{H}$, which is half the rank as a complex matrix. As in the complex case, the real and quaternionic Grassmann manifolds are connected, that is, any two projections in them are homotopic. The map f on the upper semicircle also provides a homotopy between these two points in the complex Grassmannian, which is $\mathrm{Gr}_{\mathbb{H}}(r, n)$ in both cases. Such a homotopy exists if and only if the projections at ± 1 have the same rank. And then it is unique because $\mathrm{Gr}_{\mathbb{H}}(r, n)$ is simply connected. So all Hamiltonians with time-reversal symmetry in dimension 1 are in the same topological phase.

As it turns out, the set of homotopy classes of Hamiltonians with time-reversal symmetry Θ with $\Theta^2 = -1$ in dimension 2 is in bijection with $\mathbb{Z}/2$, that is, there are exactly two homotopy classes of Hamiltonians with this symmetry (see [2]).

6 The bulk–edge correspondence

The homotopy classification of Hamiltonians studied above is interesting because an non-trivial homotopy class may make an insulator into a conductor when a boundary or edge is produced. In general, a precise version of this phenomenon is called the bulk–edge or bulk–boundary correspondence. It relates the physical properties of the bulk (interior) and the boundary of a material. Consider a d-dimensional material, described in a tight-binding model by a Hamiltonian H on the Hilbert space $\ell^2(\mathbb{Z}^d, \mathbb{C}^n)$ for some $n \in \mathbb{N}$. Let G be some group of internal symmetries that is acting on $\mathbb{C}^n$ by unitary or anti-unitary operators that commute or anti-commute with the Hamiltonian H.

In reality, our chunks of materials are finite, so that we should replace the lattice $\mathbb{Z}^d$ by a finite subset of it. Then our Hamiltonian becomes a finite matrix, and we lose all structure. It is crucial to work on the infinite set $\mathbb{Z}^d$ to talk about periodicity and describe Hamiltonian through matrix-valued functions on $\mathbb{T}^d$. To describe a material with a boundary, we replace $\mathbb{Z}^d$ by $\mathbb{N} \times \mathbb{Z}^{d-1}$. The approximation is that we study the behaviour of the system near one of the boundaries, pretending that all other boundary faces are infinitely far away.

We start with the Hamiltonian on $\ell^2(\mathbb{Z}^d, \mathbb{C}^n)$ that describes the periodic material without boundary. This is described by a continuous function $H\colon \mathbb{T}^d \to \mathbb{M}_n(\mathbb{C})$ with $H(z) = H(z)^*$ for all $z \in \mathbb{T}^d$. Let $H(z) = \sum_{x\in\mathbb{Z}^d} H_x z^x$ be its Fourier series. Then the corresponding operator on $\ell^2(\mathbb{Z}^d, \mathbb{C}^n)$ is $H = \sum_{x\in\mathbb{Z}^d} S_x \otimes H_x$ with the translation operators S_x for $x \in \mathbb{Z}^d$. We also describe this operator as matrix–vector multiplication with the block matrix $(H_{x,y})_{x,y\in\mathbb{Z}^d}$, where $H_{x,y} = H_{x-y}$ for all $x, y \in \mathbb{Z}^d$. Now let $\hat{H}$ be the operator on $\ell^2(\mathbb{N} \times \mathbb{Z}^{d-1}, \mathbb{C}^n)$ that has the same block matrix components, now only for $x, y \in \mathbb{N} \times \mathbb{Z}^{d-1}$.

Let $I\colon \ell^2(\mathbb{N} \times \mathbb{Z}^{d-1}) \to \ell^2(\mathbb{Z}^d)$ be the canonical isometry and let $I_n \coloneqq I \otimes 1\colon \ell^2(\mathbb{N} \times \mathbb{Z}^{d-1}, \mathbb{C}^n) \to \ell^2(\mathbb{Z}^d, \mathbb{C}^n)$. Then

$$\hat{H} = I_n^* H I_n = \sum_{x\in\mathbb{Z}^d} I^* S_x I \otimes H_x$$

Let $x = (x_1, \dots, x_d) \in \mathbb{Z}^d$ and let $S_1, \dots, S_d$ be the shift operators on $\ell^2(\mathbb{Z}^d)$ in the coordinate directions. So $S_x = S_1^{x_1} \cdots S_d^{x_d}$. Then $I^* S_x I = \prod_{j=1}^d (I^* S_j^{x_j} I)$ because $I^* I = 1$. Let $\hat{S}_j = I^* S_j I$. This operator is the unilateral shift – an isometry – for $j = 1$ and unitary for $j = 2, \dots, d$. We interpret $S_1^x \coloneqq S_1^x$ for $x \geq 0$ and $S_1^x \coloneqq (S_1^*)^{-x}$ for $x < 0$. Then

$$\hat{H} = \sum_{x\in\mathbb{Z}^d} \hat{S}_1^{x_1} \hat{S}_2^{x_2} \cdots \hat{S}_d^{x_d} \otimes H_x.$$

The self-adjointness of $H(z)$ for $z \in \mathbb{T}^d$ is equivalent to $H_x^* = H_{-x}$ for $x \in \mathbb{Z}^d$ and to $H_{x,y}^* = H_{y,x}$ for all $x, y \in \mathbb{Z}^d$. This is equivalent to $\hat{H}$ being self-adjoint.

While the operator $\hat{H}$ above is the most obvious operator on $\ell^2(\mathbb{N} \times \mathbb{Z}^{d-1})$ associated to H, it need not be the physically correct one. The Hamiltonian in a tight binding model is an effective one, taking into accout the interaction between the many electrons and nuclei. Near the boundary, the physics of the material is different. The distances between

the nuclei may be different, and there are no electrons on one side. So one should expect that the true Hamilton operator of the system with boundary differs is $\hat{H} + K$ with an error term K. The matrix coefficients $K_{x,y}$ of K for $x, y \in \mathbb{N} \times \mathbb{Z}^{d-1}$ should decay for $x_1 \to \infty$ or $y_1 \to \infty$ and should be invariant under translations in $\{0\} \times \mathbb{Z}^{d-1}$. If we assume that $H_x = 0$ for $\|x\| > R$, then it is reasonable to also assume that $K_{x,y} = 0$ for $x_1 > R$ or $y_1 > R$. Another reason to expect this structure is that the operators $\hat{H}$ above do not form a *-algebra. While $\hat{S}_1^* \hat{S}_1 = 1$, the other composition $\hat{S}_1 \hat{S}_1^*$ is the projection onto the image of $\hat{S}_1$, which is the linear span of δ_x with $x_1 \geq 1$. Let $\hat{A}$ be the space of all operators on $\ell^2(\mathbb{N} \times \mathbb{Z}^{d-1}, \mathbb{C}^n)$ of the form $\hat{H} + K$, such there is $R > 0$ with $H_x = 0$ for $x \in \mathbb{Z}^d$ with $\|x\| > R$ and $K_{x,y} = 0$ for $x, y \in \mathbb{N} \times \mathbb{Z}^{d-1}$ with $\|x - y\| > R$ or $x_1 > R$ or $y_1 > R$. This is a *-algebra. Let $\mathrm{C}^*(\hat{A})$ be its norm closure.

The map sending $\hat{H} + K$ to H extends to a surjective *-homomorphism from $\mathrm{C}^*(\hat{A})$ onto $\mathrm{C}(\mathbb{T}^d, \mathbb{M}_n(\mathbb{C}))$ and that its kernel is isomorphic to $\mathbb{K}(\ell^2\mathbb{N}) \otimes \mathrm{C}(\mathbb{T}^{d-1}, \mathbb{M}_n(\mathbb{C}))$, where $\mathbb{K}(\ell^2\mathbb{N})$ denotes the C^*-algebra of compact operators on the Hilbert space $\ell^2\mathbb{N}$. The construction above lifts a self-adjoint element H of $\mathrm{C}(\mathbb{T}^d, \mathbb{M}_n(\mathbb{C}))$ to a self-adjoint element $\hat{H}$ of $\mathrm{C}^*(\hat{A})$. This lifting is not unique: two liftings differ by a self-adjoint element of $\mathbb{K}(\ell^2\mathbb{N}) \otimes \mathrm{C}(\mathbb{T}^{d-1}, \mathbb{M}_n(\mathbb{C}))$.

Recall that H describes an insulator if and only if H is invertible. Similarly, $\hat{H}$ describes an insulator if and only if $\hat{H}$ is invertible. If H is invertible, but not $\hat{H}$, then $\ker \hat{H}$ consists of functions that are supported near the boundary $\{0\} \times \mathbb{Z}^{d-1}$. These are conducting states of the half-space material that are localised near the boundary. Now the kernel of $\hat{H}$ depends on the choice of the lifting. We may certainly change the kernel of $\hat{H}$ by adding terms in $\mathbb{K}(\ell^2\mathbb{N}) \otimes \mathrm{C}(\mathbb{T}^{d-1}, \mathbb{M}_n(\mathbb{C}))$. The interesting phenomenon is that sometimes an invertible H does not have any invertible lifting $\hat{H}$ for topological reasons. In general, it may be shown that the existence of an invertible lifting of H is a homotopy invariant property. That is, if H is homotopic to H' and if H has an invertible lifting, then so does H'. To make things clearer, we now restrict attention to the one-dimensional case.

So let $d = 1$. Then $\mathrm{C}^*(\hat{A})$ is simply the algebra of $n \times n$-matrices over the Toeplitz C^*-algebra $\mathcal{T}$, which is the C^*-subalgebra of $\mathbb{B}(\ell^2\mathbb{N})$ generated by the unilateral shift, which we denote by $\hat{S}$. The extension above specialises to the well known C^*-algebra extension

$$\mathbb{K}(\ell^2\mathbb{N}) \rightarrowtail \mathcal{T} \twoheadrightarrow \mathrm{C}(\mathbb{T}).$$

To see interesting topological phases in dimension 1, we consider systems with chiral symmetries. In the simplest case, we have $n = 2$ and Ξ is the diagonal matrix with entries $+1, -1$. So the possible Hamiltonians are off-diagonal matrices

$$H = \begin{pmatrix} 0 & U^* \\ U & 0 \end{pmatrix}.$$

And this is invertible if and only if U is invertible. Spectral flattening reduces further to the case where U is unitary. An example of this type is the Su–Schrieffer–Heeger model in Section 3.3.5.

Example 6.1. Consider the case where $U = S_k$ is translation by $k \in \mathbb{N}$. Then

$$\hat{H} = \begin{pmatrix} 0 & (\hat{S}^*)^k \\ \hat{S}^k & 0 \end{pmatrix}.$$

The operator $\hat{S}^k$ is an isometry, hence injective, and the kernel of $(\hat{S}^*)^k$ is the orthogonal complement of the image of $\hat{S}^k$, which is spanned by the first basis vectors $\delta_0, \dots, \delta_{k-1}$. So the kernel of $\hat{H}$ has dimension k. We may easily enlarge this dimension as much as we like by adding a finite-rank operator to $\hat{H}$. For instance,

$$\hat{H} + K = \begin{pmatrix} 0 & \hat{S}^m(\hat{S}^*)^{k+m} \\ \hat{S}^{k+m}(\hat{S}^*)^m & 0 \end{pmatrix}$$

is another lifting of H, which differs from $\hat{H}$ by a finite-rank operator. Now $\hat{S}^m(\hat{S}^*)^{k+m}$ and $\hat{S}^{k+m}(\hat{S}^*)^m$ are partial isometries, and their kernels have dimension $k+m$ and m, respectively. But, as it turns out, any lifting of H with chiral symmetry has a kernel whose dimension is at least k. To explain this, we recall some fundamental results about the index of Fredholm operators.

Definition 6.2. An operator T on a Hilbert space V is called *Fredholm* if $\ker T$ and $\operatorname{coker} T := V/\operatorname{im} T$ are finite-dimensional. Then the *index* of T is

$$\operatorname{ind} T := \dim \ker T - \dim \operatorname{coker} T.$$

Theorem 6.3. *An operator T on a Hilbert space V is Fredholm if and only if there is an operator S, called* parametrix *for T, such that $S \cdot T - \mathrm{id}_V$ and $T \cdot S - \mathrm{id}_V$ are compact.*

Theorem 6.3 implies that an element $\hat{H}$ of $\mathbb{M}_n \otimes \mathcal{T}$ is Fredholm if and only if its image in the quotient C*-algebra $\mathbb{M}_n \otimes \mathcal{T} \;/\; \mathbb{M}_n \otimes \mathbb{K}(\ell^2\mathbb{N}) \cong \mathrm{C}(\mathbb{T}, \mathbb{M}_n)$ is invertible. In other words, $\hat{H}$ is Fredholm if and only if H is invertible. So the index of $\hat{H}$ is well defined for all of them. Actually, the index of all self-adjoint operators vanishes. The interesting index is the index of the off-diagonal entry $\hat{U}$ in $\hat{H}$ above.

Theorem 6.4. *Let T be a Fredholm operator. If K is compact, then $T + K$ is Fredholm and $\operatorname{ind}(T + K) = \operatorname{ind} T$. There is $\varepsilon > 0$ such that any operator S with $\|T - S\| < \varepsilon$ is Fredholm and satisfies $\operatorname{ind} T = \operatorname{ind} S$. If two Fredholm operators are homotopic in the space of Fredholm operators, then they have the same index.*

The second statement says that the map taking a Fredholm operator to its index is locally constant. Hence it is constant on the connected components of the space of Fredholm operators. This is the reason why homotopic Fredholm operators have the same index.

Corollary 6.5. *Let $U \in \mathrm{C}(\mathbb{T}, \mathbb{M}_n)$ be invertible. Then any lifting $\hat{U} \in \mathbb{M}_n \otimes \mathcal{T}$ of U is Fredholm. The index $\operatorname{ind} \hat{U}$ depends only on the class of U up to homotopy. It is equal to minus the winding number invariant of U.*

Proof. The first statement in Theorem 6.4 says that all liftings $\hat{U}$ of U have the same index. A norm-continuous homotopy of invertible elements in $\mathrm{C}(\mathbb{T}, \mathbb{M}_n)$ lifts to a norm-continuous homotopy of Fredholm operators in $\mathbb{M}_n \otimes \mathcal{T}$. So the third statement in Theorem 6.4 implies that liftings of homotopic invertible elements in $\mathrm{C}(\mathbb{T}, \mathbb{M}_n)$ have the same index. We have seen that any element of $\mathrm{C}(\mathbb{T}, \mathbb{M}_n)$ is homotopic to a function of the form

$$\upsilon_m \colon \mathbb{T} \to \mathbb{M}_n, \qquad z \mapsto \operatorname{diag}(z^m, 1, \dots, 1)$$

for $m \in \mathbb{Z}$. This is lifted to S^m if $m \geq 0$, and to $(S^*)^{-m}$ if $m < 0$. Direct computation shows that the index of S^m is $-m$ and the index of $(S^*)^{-m}$ is m. Hence the lifting of U has index $-m$ if U is homotopic to the function υ_m above. The number m here is the winding number of the function $\mathbb{T} \to \mathbb{C} \setminus \{0\}$, $z \mapsto \det U(z)$. □

Corollary 6.5 explains why all liftings of the operator H in Example 6.1 have index k. If $k \neq 0$, this implies that the operator cannot be invertible. In fact, an operator T has the property that $T + K$ is invertible for some compact operator K if and only if T is Fredholm with index 0. One implication follows from the theorems above: the index of an invertible operator is clearly 0, and T and $T + K$ have the same index. Conversely, if T has index 0, then one may choose a finite rank-operator K that vanishes on the orthogonal complement of $\ker T$ so that $T + K$ is invertible. Thus the index of U is exactly the obstruction to finding a lifting $\hat{U}$ that is invertible.

References

[1] B. Andrei Bernevig, Taylor L. Hughes, and Shou-Cheng Zhang, *Quantum Spin Hall Effect and Topological Phase Transition in HgTe Quantum Wells*, Science **314** (2006), no. 5806, 1757–1761, doi: 10.1126/science.1133734.

[2] Domenico Fiorenza, Domenico Monaco, and Gianluca Panati, $\mathbb{Z}_2$ *invariants of topological insulators as geometric obstructions*, Comm. Math. Phys. **343** (2016), no. 3, 1115–1157, doi: 10.1007/s00220-015-2552-0. MR **3488553**

[3] Daniel S. Freed and Gregory W. Moore, *Twisted equivariant matter*, Ann. Henri Poincaré **14** (2013), no. 8, 1927–2023, doi: 10.1007/s00023-013-0236-x. MR **3119923**

[4] F. D. M. Haldane, *Model for a quantum hall effect without landau levels: Condensed-matter realization of the "parity anomaly"*, Phys. Rev. Lett. **61** (1988), 2015–2018, doi: 10.1103/PhysRevLett.61.2015.

[5] M. Z. Hasan and C. L. Kane, *Colloquium: Topological insulators*, Rev. Mod. Phys. **82** (2010), no. 4, 3045–3067, doi: 10.1103/RevModPhys.82.3045.

[6] Gregor Jotzu, Michael Messer, Rémi Desbuquois, Martin Lebrat, Thomas Uehlinger, Daniel Greif, and Tilman Esslinger, *Experimental realization of the topological Haldane model with ultracold fermions*, Nature **515** (2014), 237, doi: 10.1038/nature13915.

[7] C. L. Kane and E. J. Mele, *Quantum spin Hall effect in graphene*, Phys. Rev. Lett. **95** (2005), 226801, doi: 10.1103/PhysRevLett.95.226801.

[8] Ricardo Kennedy and Charles Guggenheim, *Homotopy theory of strong and weak topological insulators*, Phys. Rev. B **91** (2015), no. 24, 245148, doi: 10.1103/PhysRevB.91.245148.

[9] Ricardo Kennedy and M. R. Zirnbauer, *Bott periodicity for* $\mathbb{Z}_2$ *symmetric ground states of gapped free-fermion systems*, Comm. Math. Phys. **342** (2016), no. 3, 909–963, doi: 10.1007/s00220-015-2512-8. MR **3465435**

[10] Mamoru Mimura, *Homotopy theory of Lie groups*, Handbook of algebraic topology, 1995, pp. 951–991, doi: 10.1016/B978-044481779-2/50020-1. MR **1361904**

[11] Emil Prodan and Hermann Schulz-Baldes, *Bulk and boundary invariants for complex topological insulators*, Mathematical Physics Studies, Springer, 2016. From K-theory to physics. doi: 10.1007/978-3-319-29351-6 MR **3468838**

[12] Michael Reed and Barry Simon, *Methods of modern mathematical physics. I. Functional analysis*, Academic Press, Inc., New York, 1980. MR **751959**

[13] ———, *Methods of modern mathematical physics. II. Fourier analysis, self-adjointness*, Academic Press, Inc., New York, 1975. MR **751959**

[14] Gordon W. Semenoff, *Condensed-matter simulation of a three-dimensional anomaly*, Phys. Rev. Lett. **53** (1984), 2449–2452, doi: 10.1103/PhysRevLett.53.2449.

[15] W. P. Su, J. R. Schrieffer, and A. J. Heeger, *Solitons in polyacetylene*, Phys. Rev. Lett. **42** (1979), no. 25, 1698–1701, doi: 10.1103/PhysRevLett.42.1698.

[16] ———, *Soliton excitations in polyacetylene*, Phys. Rev. B **22** (1980), 2099–2111, doi: 10.1103/PhysRevB.22.2099.

[17] Hirosi Toda, *Composition methods in homotopy groups of spheres*, Annals of Mathematics Studies, No. 49, Princeton University Press, Princeton, N.J., 1962. MR **0143217**

[18] Stanisław Lech Woronowicz, *Unbounded elements affiliated with* C^**-algebras and noncompact quantum groups*, Comm. Math. Phys. **136** (1991), no. 2, 399–432, available at `http://projecteuclid.org/euclid.cmp/1104202358`. MR **1096123**

[19] Yugui Yao, Fei Ye, Xiao-Liang Qi, Shou-Cheng Zhang, and Zhong Fang, *Spin-orbit gap of graphene: First-principles calculations*, Phys. Rev. B **75** (2007Jan), 041401, doi: 10.1103/PhysRevB.75.041401.