Geometry and Topology
in Electronic Structure Theory

Raffaele Resta

Notes subject to ongoing editing
This version run through LaTeX on 16–Feb–22 at 17:0

Contents

0

Chapter 1

Introduction

1.1 About the present Notes

These Notes grew up as background material for an advanced undergraduate course delivered at the University of Trieste since March 2013 onwards. Even at that time, there was no lack of background material, in the form of comprehensive review papers, all of them state-of-the-art at the time of their publication. Some of these reviews were authored or coauthored by me [1, 2, 3, 4, 5, 6, 7, 8, 9], and many more by outstanding colleagues. A non exhaustive (and not updated) list of some relatively recent reviews is: [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. To these I add the outstanding book by David Vanderbilt, appeared in 2018 [24].

The present Notes, posted at `http://www-dft.ts.infn.it/~resta/`, have been continuously edited, updated, and augmented from 2013 to the present days. Some Chapters have been completely rewritten for the 2021 release; given the very long time span in which the editing has proceeded some repetitions are surely present. It is also quite probable that some Sections are not mutually consistent, particularly in the notations and symbols adopted.

Despite my best efforts, it is not possible to keep the pace with the most recent developments, and with my viewpoint on them. These Notes will therefore remain—probably in perpetuum—a kind of "work in progress".

The present Notes include some Appendices whose topics are neither "geometrical" nor "topological", although some of their ingredients are calculated by means of Berry phases and other geometrical tools. However, the Appendix topics are regularly included in the course for which the present Notes are planned.

It is easy to guess these Notes will continue to be plagued by many typos and even more serious errors. I kindly ask any reader spotting such occurrences to notify me at `resta@iom.cnr.it`.
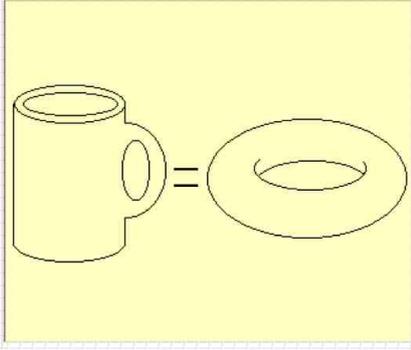
Figure 1.1: The hallmark of topology, as in many popular presentations. Hundreds of figures like this, and even some very perspicuous videos, can be downloaded from the internet. The two closed surfaces ("two-dimensional compact manifolds") have the same topological invariant $g = 1$, which measures the number of handles.

1.2   What topology is about

Topology is defined as a branch of mathematics that describes properties which remain unchanged under smooth deformations; such properties are usually labelled by integer numbers, named topological invariants. The concepts and tools belonging to topology are continuity and connectivity, open and closed sets, neighborhoods, and the like.

Differentiability, or even a metric structure, are not needed; theorems are proved under very general hypotheses, and are therefore very powerful, being applicable to very diverse frameworks. The tradeoff is that proofs, and even definitions, look clumsy and obscure to readers with the mathematical background of a typical condensed matter physicist. The good news is that the topological properties most relevant for electronic structure theory can be formulated in the more familiar language of differential geometry.

At its simplest level, topology is the branch of mathematics used to classiify the shapes of three-dimensional objects. Many introductions to topology start with the statement that, to a topologist, a coffee cup and a doughnut are the same thing, as in Fig. 1.1. Intuitively, the common feature of the two objects is the presence of one, and only one, handle. The mathematical definition of "handle" is coming soon.

1.2.1   Gauss-Bonnet theorem

We start with the simplest example, a sphere, and a tangent plane at a given point. In a local system of Cartesian coordinates on the plane the equation of the sphere is

$$z = R - \sqrt{R^2 - x^2 - y^2} \simeq \frac{x^2 + y^2}{2R}, \tag{1.1}$$

and the Hessian matrix is

$$H = \begin{pmatrix} \frac{\partial^2 z}{\partial x^2} & \frac{\partial^2 z}{\partial x \partial y} \\ \frac{\partial^2 z}{\partial y \partial x} & \frac{\partial^2 z}{\partial y^2} \end{pmatrix} = \begin{pmatrix} 1/R & 0 \\ 0 & 1/R \end{pmatrix}. \tag{1.2}$$
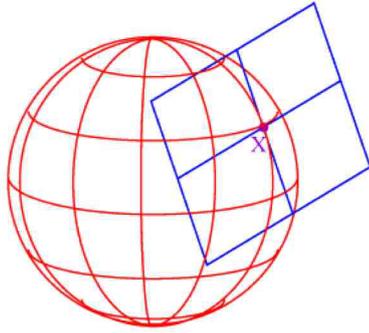
Figure 1.2: A sphere of radius R, and its tangent plane in a generic point. The Gaussian curvature in this trivial case is $\mathbf{\Omega} = 1/R^2$.

The Gaussian curvature $\mathbf{\Omega}$ is by definition the determinant of the Hessian at the tangency point. It is obviously constant and equal to $1/R^2$ at any point of the sphere; notice that the orientation of the z axis (either inwards or outwards) is irrelevant. The integral of $\mathbf{\Omega}$ over the whole closed surface is 4π.

Next we consider a smooth (i.e. twice differentiable) surface of arbitrary shape: the Gaussian curvature is defined as the determinant of the Hessian at the tangent plane, similarly to what we did for the sphere:

$$\mathbf{\Omega} = \det \begin{pmatrix} \frac{\partial^2 z}{\partial x^2} & \frac{\partial^2 z}{\partial x \partial y} \\ \frac{\partial^2 z}{\partial y \partial x} & \frac{\partial^2 z}{\partial y^2} \end{pmatrix}. \tag{1.3}$$

In general, $\mathbf{\Omega}$ can be positive, negative (at a saddle point), or zero (e.g. for a plane or a cylinder).

So far, we have looked at a $2d$ surface in a $3d$ space; but Gauss' great intuition—the 'theorema egregium"—is that the concept of curvature is intrinsic to the $2d$ surface itself. An ant crawling on the surface can measure the curvature in two equivalent ways; either via parallel transport, or measuring angles. Suppose we join three points on a surface by the shortest paths (geodesics). Then the sum of the angles in the triangle is π on a flat surface, greater than π if $\mathbf{\Omega}$ is positive (like on the sphere), and smaller than π if $\mathbf{\Omega}$ is negative. The angular mismatch per unit area defines indeed the Gaussian curvature.

The Gauss-Bonnet theorem states that for any closed smooth surface

$$\frac{1}{2\pi} \int_S d\sigma \, \mathbf{\Omega} = 2(1 - g), \tag{1.4}$$

where g is a nonnegative integer, called the "genus" of the surface. Surfaces which can be continuously deformed into each other (i.e. "homeomorphic") have the same genus. For the sphere and any surface homeomorphic to it $g = 0$; both the coffee cup and the doughnut, Fig. 1.1 have $g = 1$; a double-handle cup has $g = 2$. The genus is thus the mathematical definition for the number of handles. Notice also

3

that an ant crawling on a nontrivial $(g > 0)$ surface may walk around some closed paths that cannot be smoothly shrunk to a point on the surface.

1.2.2 Euler characteristic

We have considered smooth surfaces so far, but topological invariants are based on the more general condition of continuity, and—to the delight of mathematicians—unsurprisingly can be defined even for pathological surfaces ("manifolds" in topology-speak). The simplest non smooth case addresses polyhedra, where the Gaussian curvature is either zero (on the faces) or singular (at vertices and edges).

The Euler characteristic is defined as $\chi = V - E + F$, where V is the number of vertices, E is the number of edges, and F is the number of faces. If we address the set of regular polyhedra (tetrahedron, cube, octahedron, dodecahedron, icosahedron) it is easily verified that $\chi = 2$. All these surfaces can be continuously deformed into ("are homeomorphic to") each other, and into a sphere. In fact there is a one-to-one relationship between the Euler characteristic and the genus: $\chi = 2(1 - g)$. Polyhedra can also have $\chi \neq 2$, like the doughnut-shaped one shown in Fig. 1.4.

1.3 Electronic wavefunctions

In the domain of electronic structure, the typical object addressed via geometrical and/or topological concepts is the electronic ground state of some system. Whenever an observable effect has the nature of a topological invariant, i.e. it is an integer number, two remarkable features occur. (1) The observable is measurable in principle with infinite precision (10^{-9} is actually attained for the quantum Hall effect). (2) The observable is very robust under even strong variations of the sample conditions; a very disruptive perturbation is needed to switch from one
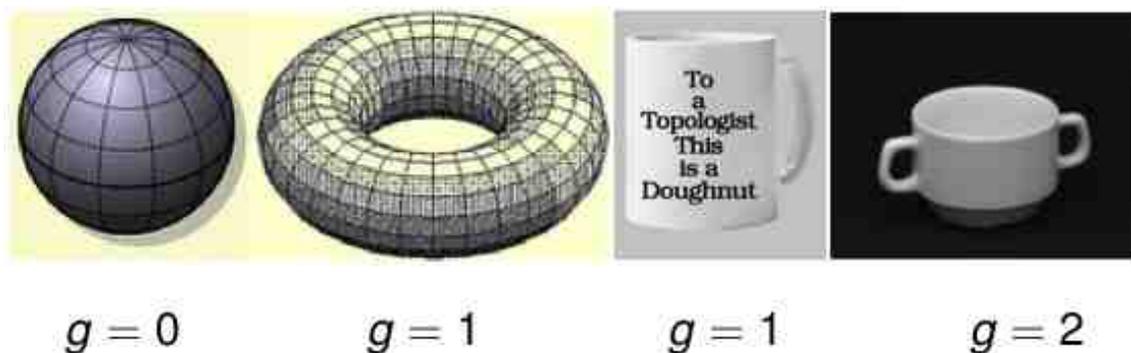
Figure 1.3: Some surfaces and their genus

4
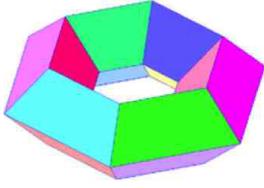
Figure 1.4: A doughnut shaped polyhedron. This surface has Euler characteristic $\chi = 0$ or, equivalently, genus $g = 1$.

integer to another. Topology concerns mostly insulators: in this case the disruptive perturbation amounts to crossing a metallic state.

These Notes are entirely devoted to physical properties having a topological and/or geometrical character. I am not sure of always using the right semantics. Loosely speaking, I would use the term "topological" for something which is quantized, and "geometrical" for something which is not. The framework and the mathematical tools are often the same for quantized and nonquantized quantities, the former frequently occurring as special cases of the latter.

The Berry phase is the typical geometrical quantity which is not quantized, although it can be quantized in high-symmetry cases. The macroscopic polarization of a solid is a Berry phase, and is obviously (from an experimental viewpoint) a nonquantized observable. Nonetheless, there are aspects of the modern theory of polarization that I would define topological. The same applies to other geometrical properties considered in this Notes. It is reassuring that even other authors often use "geometrical" and "topological" as synonymous, and that an illustrious author like Michael Berry confesses an original mistake about the semantics [25].

Finally, a few words about the many calculations cited and sometimes briefly outlined here. Unless otherwise stated, the term "first-principle calculations", when referred to a condensed matter system, means density functional calculations; independent-electron eigenfunctions and eigenvalues are the Kohn-Sham (KS) ones. Despite these Notes mostly address a computational physics readership, no technical details are given (basis sets, pseudopotentials, functionals...); they are obviously detailed in the original literature, while the focus here is on the physical properties.

1.4   Units

We use Gaussian electromagnetic units throughout: these have the advantage (at variance with SI units) that electric and magnetic fields have the same dimensions. Furthermore, the nasty ε_0 and μ_0 disappear; SI formulæ are converted by setting $4\pi\varepsilon_0 = 1$ and $4\pi/\mu_0 = 1$. Gaussian units are elegant and simple, and symmetric in dealing with electric and magnetic fields; they somewhat hide the fact that magnetic energies are typically much smaller—by a factor $1/137^2$—of electrical energies.

For a single particle, the Newton equation of motion and the Hamiltonian read,

respectively

$$M\frac{d\mathbf{v}}{dt} = \mathbf{f} = Q\left(\mathbf{E} + \frac{1}{c}\mathbf{v}\times\mathbf{B}\right), \tag{1.5}$$

$$H = \frac{1}{2M}\left(\mathbf{p} - \frac{Q}{c}\mathbf{A}(\mathbf{r})\right)^2 + Q\Phi(\mathbf{r}). \tag{1.6}$$

Generally, Gaussian electromagnetic units are associated to mechanical cgs units, but this is no means necessary. In electronic structure theory, it is expedient to associate Gaussian electromagnetic units with atomic units (a.u.), defined as $e^2 = 1$, $m_e = 1$, $\hbar = 1$. The unit of energy is the hartree (1 Ha = 2 Ry = 27.21 eV). In the present Notes the electron charge is $-e$, with $e > 0$; this sign choice agrees with most (but not all) the recent literature. For instance, the very popular review of Ref [1] adopts the symbol e for the "algebraic" electron charge ($e < 0$).

The speed of light in a.u. is $c = 137$. This immediately hints at why the largest atomic number Z in the periodic table is $Z \simeq 100$: in fact the core electrons have (in a.u.) energies of the order of Z^2, hence velocities of the order of Z.

1.5 Symbols

I am faced here with two contrasting issues: adopting the symbols most currently used in the literature, and adopting different symbols for different objects. This proved to be near to impossible in a work of the present kind, if baroque symbols are to be ruled out. When I started writing the present Notes in 2013 I adopted $A, \mathbf{A}, \mathcal{A}, \boldsymbol{\mathcal{A}}, \mathscr{A}$, all with a different meaning; similarly, I use $P, \hat{P}, \mathbf{P}, \mathcal{P}, \mathscr{P}$. After much editing over the years, I found unavoidable to use—in different Chapters or Sections—the same symbol for different objects or even different symbols for the same object. For instance, depending on the context, the symbol P may indicate a projector or, otherwise, a one-dimensional electrical polarization. Therefore caution is in order when extrapolating a given symbol from its own context.

1.6 Gauge and flux

We consider here a simple exercise which plays the role of a very important paradigm; it illustrates basic concepts and results which are going to reappear several times all along the present Notes.

We address the single-particle Hamiltonian

$$H = \frac{1}{2m}(\mathbf{p} + \frac{e}{c}\mathbf{A})^2 + V(\mathbf{r}), \tag{1.7}$$

where the vector potential $\mathbf{A}$ is independent of space and time. It is usually said that $\mathbf{A}$ is a pure gauge, meaning with this that it does not affect the fields:

$$\mathbf{B} = \nabla \times \mathbf{A} \equiv 0, \qquad \mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{A}}{\partial t} \equiv 0. \tag{1.8}$$

1.6.1 Classical mechanics

Let us first adopt a classical viewpoint. The Hamilton equation of motions are

$$\dot{\mathbf{p}} = -\frac{\partial H}{\partial \mathbf{r}} = -\nabla V(\mathbf{r}) \tag{1.9}$$

$$\dot{\mathbf{r}} = \frac{\partial H}{\partial \mathbf{p}} = \frac{1}{m}(\mathbf{p} + \frac{e}{c}\mathbf{A}). \tag{1.10}$$

From these we get

$$\mathbf{p} = m\dot{\mathbf{r}} - \frac{e}{c}\mathbf{A}, \tag{1.11}$$

which leads to the Newton equation of motion

$$m\ddot{\mathbf{r}} = -\nabla V(\mathbf{r}). \tag{1.12}$$

The bottom line looks quite obvious: a pure gauge has no effect. A basic tenet of classical mechanics is that the equations of motion can always be directly expressed in terms of the forces (i.e. the fields), while the potentials—scalar and vector—are auxiliary quantities, devoid of physical meaning.

1.6.2 Quantum mechanics, open boundary conditions

Next we switch to quantum mechanics. It is expedient to rewrite Eq. (1.7) as

$$H(\boldsymbol{\kappa}) = \frac{1}{2m}(\mathbf{p} + \hbar\boldsymbol{\kappa})^2 + V(\mathbf{r}), \qquad \boldsymbol{\kappa} = \frac{e}{c\hbar}\mathbf{A}, \tag{1.13}$$

where $\boldsymbol{\kappa}$, having the dimensions of an inverse length, will be referred to as "twist" in the following. The Schrödinger equation is

$$H(\boldsymbol{\kappa})|\psi_n(\boldsymbol{\kappa})\rangle = \epsilon_n(\boldsymbol{\kappa})|\psi_n(\boldsymbol{\kappa})\rangle. \tag{1.14}$$

The eigenvectors and eigenvalues of the Schrödinger equation depend on the boundary conditions assumed.

The so-called open boundary conditions (OBCs) require that the bound eigenstates are square-integrable over $\mathbb{R}^3$. Let $|\psi_n(0)\rangle$ be a nondegenerate eigenstate of the "untwisted" Hamiltonian within OBCs. Then the state $\mathrm{e}^{-i\boldsymbol{\kappa}\cdot\mathbf{r}}|\psi_n(0)\rangle$

obviously obeys OBCs as well, and also obeys Eq. (1.14) with a $\boldsymbol{\kappa}$-independent eigenvalue. Therefore it coincides with the n-th eigenstate $|\psi_n(\boldsymbol{\kappa})\rangle$ of the twisted Hamiltonian; notice that this eigenstate is arbitrary by a $\boldsymbol{\kappa}$-dependent phase factor.

We conclude that a pure gauge within OBCs affects the wavefunction, but does not affect any of the observable quantities, such as expectation values, density, and current. We spell this, in jargon, by saying that the "twist" is easily "gauged away" within OBCs.

1.6.3 Quantum mechanics, periodic boundary conditions

We assume periodic boundary conditions (PBCs) over a cubic box of side L, i.e. we require the eigenstates of Eq. (1.14) to be Born-von-Kàrmàn periodic with period L over x, y, and z at any given $\boldsymbol{\kappa}$. Each Cartesian coordinate is therefore equivalent to an angle, e.g. $\varphi_x = 2\pi x/L$.

If $|\psi_n(0)\rangle$ is an eigenstate of the untwisted Hamiltonian within PBCs, then the state $\mathrm{e}^{-i\boldsymbol{\kappa}\cdot\mathbf{r}}|\psi_n(0)\rangle$ obeys Eq. (1.14) with a $\boldsymbol{\kappa}$-independent eigenvalue, but for a general $\boldsymbol{\kappa}$ it *does not* obey PBCs, and therefore in general does not coincide with the genuine eigenstate $|\psi_n(\boldsymbol{\kappa})\rangle$. Within PBCs the spectrum of Eq. (1.14) depends on the twist $\boldsymbol{\kappa}$ in a nontrivial way.

If $|\psi_n(\boldsymbol{\kappa})\rangle$ is an eigenstate of Eq. (1.14) within PBCs with eigenvalue $\epsilon_n(\boldsymbol{\kappa})$, then the auxiliary state $|\tilde{\psi}_n(\boldsymbol{\kappa})\rangle = \mathrm{e}^{i\boldsymbol{\kappa}\cdot\mathbf{r}}|\psi_n(\boldsymbol{\kappa})\rangle$ obeys the untwisted ($\boldsymbol{\kappa}=0$) Schrödinger equation, and quasi-periodical (a.k.a. "twisted" or "skewed") boundary conditions: at any two opposite faces of the cube the wavefunction differs by a $\boldsymbol{\kappa}$-dependent phase factor.

In other words the problem can be formulated in two equivalent ways: either the Hamiltonian is $\boldsymbol{\kappa}$-dependent, as in Eq. (1.14), and the boundary conditions are $\boldsymbol{\kappa}$-independent; or the Hamiltonian is $\boldsymbol{\kappa}$-independent but the boundary conditions are "twisted" in a $\boldsymbol{\kappa}$-dependent way.

1.6.4 Example: Free particle in 1d

For the sake of simplicity, we consider Eq. (1.14) in 1d, and with $V \equiv 0$:

$$\frac{\hbar^2}{2m}\left(-i\frac{d}{dx}+\kappa\right)^2|\psi_n(\kappa)\rangle = \epsilon_n(\kappa)|\psi_n(\kappa)\rangle. \qquad (1.15)$$

The eigenfunctions within PBCs and the spectrum are

$$\langle x|\psi_n(\kappa)\rangle \propto \mathrm{e}^{i\frac{2\pi n}{L}x}, \qquad n \in \mathbb{Z}, \qquad (1.16)$$

$$\epsilon_n(\kappa) = \frac{\hbar^2}{2m}\left(\frac{2\pi n}{L}+\kappa\right)^2, \qquad (1.17)$$

where the nontrivial κ-dependence is perspicuous. The velocity operator can be written as

$$v = \frac{1}{\hbar}\frac{\partial H}{\partial \kappa},$$ (1.18)

and the Hellmann-Feynman theorem yields

$$\langle \psi_n | v | \psi_n \rangle = \frac{1}{\hbar}\frac{d\epsilon_n(\kappa)}{d\kappa}.$$ (1.19)

We have introduced PBCs as a basic framework of condensed matter physics. Many concepts (like the Bloch vector or the Fermi surface) make sense only within PBCs. But we also may regard this problem as if the electrons were confined to a circular rail of circumference L, as in Fig. 1.5. There is no field (electric or magnetic) on the rail, but a constant vector potential of intensity $A = c\hbar\kappa/e$ is present along the rail; eigenvectors and eigenfunctions *depend* on its value.

1.6.5 Flux and flux quantum

The constant vector potential A on the circular rail corresponds to a magnetic flux $\phi = LA$ threading the surface encircled by the rail, in a region *not visited* by the electronic system; it has been appropriately called by some authors "inaccessible flux".

We further observe that the spectrum, Eq. (1.17), is periodic in κ with period $2\pi/L$; alternatively, it is periodic in the flux ϕ with period $\phi_0 = 2\pi\hbar c/e = hc/e$, the elementary flux quantum. In cgs units $hc/e = 4.135 \times 10^{-7}$ gauss cm^2, while in SI units $\phi_0 = h/e = 4.136 \times 10^{-15}$ Wb. Notice also that, in the framework of superconductivity, the same symbol ϕ_0 indicates *one half* of this (it refers to electron pairs).

We stress that only the fractional part of the flux affects the results in a nontrivial way. This is perspicuous if we recast Eq. (1.17) as

$$\epsilon_n(\phi) = \frac{\hbar^2}{2m}\left(\frac{2\pi}{L}\right)^2\left(n + \frac{\phi}{\phi_0}\right)^2.$$ (1.20)
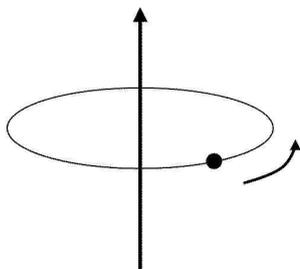
Figure 1.5: The electron motion is confined to a circular rail. A constant vector potential $A = c\hbar\kappa/e$ along the rail, as in Eq. (1.15), corresponds to vanishing fields (electric and magnetic), yet the spectrum *depends* on the "inaccessible flux" threading the surface encircled by the rail.

9

The flux breaks time-reversal (T) symmetry ($\kappa \to -\kappa$), and the spectrum is nondegenerate, except when $\phi = 0$ or $\phi = \phi_0/2$, the latter also called "π flux". In these two cases (and in these cases only) the eigenfunctions can be chosen as real.

When the flux is varied with time, an emf is induced along the loop. Using Eq. (1.19), the current is

$$I = -\frac{e}{L}v = -c\frac{d\epsilon_n}{d\phi}. \qquad (1.21)$$

This result is remarkable: it holds even in presence of a potential $V(x)$, and generalizes straightforwardly to N noninteracting electrons. It will be used in the discussion of the quantum Hall effect: see Eq. (2.20) below.

Chapter 2

Early discoveries

2.1 The Aharonov-Bohm effect: A paradox?

The Aharonov-Bohm effect is the paradigm for a measurable effect induced by an inaccessible flux. We anticipate that in many other phenomena such flux may be purely "geometrical" or "topological", without any relationship to a genuine magnetic field: this is e.g., the case considered in the next Section. It is only in the Aharonov-Bohm effect that one addresses indeed the inaccessible flux of a magnetic field, as present e.g. inside a solenoid. An interference experiment detects the presence of the flux even when the electronic motion is confined in the region outside the solenoid, where the magnetic field is zero. This seems paradoxical: something which "happens" in a region not visited by the quantum particle may affect some observable properties. Indeed, the founding fathers of quantum mechanics (in the 1920s) failed to notice such peculiar feature. It only surfaced more than 30 years afterwards in the milestone paper by Aharonov and Bohm [26], appeared in 1959, whose abstract states verbatim *"...contrary to the conclusions of classical mechanics, there exist effects of potential on charged particles, even in the regions where all the fields (and therefore the forces on the particles) vanish"*.

The paper was shocking, and its conclusions were challenged by several authors; nonetheless experimental validations appeared as early as 1960 [28, 29]. The main message of Ref. [26] is at the basis of many subsequent developments in electronic structure theory, many of them illustrated below in the present Notes. The Aharonov-Bohm effect is also at the root of the commercial SQUID technology [30].

It is remarkable that R. P. Feynman included the Aharonov-Bohm effect in his legendary lectures, delivered to the sophomore class at Caltech during the 1962-63 academic year [27]. In the final sentence about this topic, Feynman says: *"...$\mathbf{E}$ and $\mathbf{B}$ are slowly disappearing from the modern expression of physical laws; they are being replaced by $\mathbf{A}$ and Φ"*.

It is also remarkable and shameful that a paper bearing the title "Nonexistence of the Aharonov-Bohm effect" [31] was published as late as 1978. All challenges disappeared with the publication in 1984 of the celebrated paper by Michael Berry (now *Sir* Michael Berry [32]), where the eponymous phase made its first appearance [33].

2.2   Conical intersections in molecules

At the time Berry wrote his famous paper, only two occurrences of a geometrical phase (called Berry phase soon afterwards) in quantum mechanics were known to him: the Aharonov-Bohm effect and a somewhat exotic phenomenon occurring in molecular physics. Even the latter was known since the late 1950s [34, 35], and appropriately rebaptized in the late 1970s as "molecular Aharonov-Bohm effect" [36, 37]. In the subsequent years Berry phases were discovered in many branches of physics.

The smallest molecular system where the molecular Aharonov-Bohm effect is possible is a trimer, having three internal coordinates (e.g. the three internuclear distances), and the simplest trimers are of course the homonuclear ones, where symmetry plays a major role. I give a simple outline for this particular system: a dynamical Jahn-Teller effect, bearing the conventional symmetry label $E \otimes \varepsilon$.

We focus on a trimer of monovalent atoms, e.g. H_3 or Na_3, and we assume an independent-electron picture in the Born-Oppenheimer approximation. We start with the molecule in the equilateral configurations, Fig. 2.2. Two of the valence electrons occupy a totalsymmetric orbital, while the unpaired electron occupies the next available one, which has E symmetry and is doubly degenerate. In a simple tight-binding (alias minimal-basis LCAO) scheme, a possible basis in the
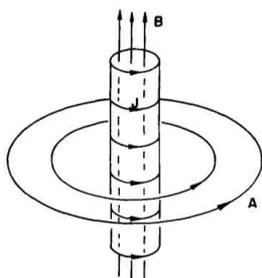
Fig. 15–6. The magnetic field and vector potential of a long solenoid.
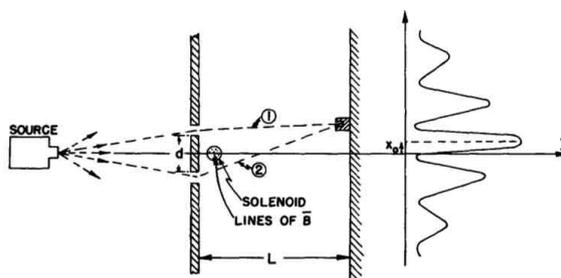
Fig. 15–7. A magnetic field can influence the motion of electrons even though it exists only in regions where there is an arbitrarily small probability of finding the electrons.

Figure 2.1: The Aharonov-Bohm interference experiment (From Ref. [27])

two-dimensional manifold is:

$$|1\rangle = \frac{1}{\sqrt{2}}(\ |B\rangle - |C\rangle\ )\ ; \qquad |2\rangle = \frac{1}{\sqrt{6}}(\ 2|A\rangle - |B\rangle - |C\rangle\ )\ , \qquad (2.1)$$

where A,B,C are atomic labels (as in the figure). This choice deserves an important comment. We are adopting OBCs, as appropriate for an isolated molecule, and the Hamiltonian is invariant under time reversal (no magnetic field, no spin-orbit interaction). These two conditions guarantee that the orbitals may always be chosen as real. They *may*, but they don't need: it may instead be convenient to choose a complex basis in the same two-dimensional degenerate manifold.

When we distort the molecule from its equilateral configuration, the doublet is linearly split: one of the two components is energetically favored, the molecule undergoes Jahn-Teller distortion, and the electronic ground state in the Born-Oppenheimer approximation becomes nondegenerate.

Next we analyze the motion of the nuclei. There are three linearly independent normal modes for the small oscillations of the internal coordinates. Of course, *in absence of a Jahn-Teller effect*, the equilateral configuration is the equilibrium one. One mode is totalsymmetric, and cannot split the electronic levels. The remaining modes are degenerate, having in fact E symmetry, and couple to the electronic doublet, originating in fact the dynamical Jahn-Teller effect. The notation $E \otimes \varepsilon$ means indeed that an E vibrational mode is coupled to an E electronic state: conventionally, one uses upper case letters as symmetry labels for the vibrational states, and lower case ones for the electronic states.

The adiabatic electronic ground state follows the nuclear motion. For a cyclic pseudorotation, shown in Fig. 2.3, the Hamiltonian is periodical, but the electronic wavefunction is *antiperiodical*. The total wavefunction in the Born-Oppenheimer approximation factors into the electronic one times the nuclear one. Given that the total wavefunction must be single-valued, even the nuclear wavefunction must be quantized using antiperiodical boundary conditions, and this affects the pseudorotation spectrum in a measurable way.

This feature has to do with the peculiar shape of the Born-Oppenheimer surface, shown in Fig. 2.4. If we adopt a two-dimensional Cartesian normal coordinate $\boldsymbol{\xi} = (\xi_1, \xi_2)$, the ionic displacements are parametrized as:

$$
\begin{aligned}
x_A &= \xi_1 & y_A &= \xi_2 \\
x_B &= -\tfrac{1}{2}\xi_1 + \tfrac{\sqrt{3}}{2}\xi_2 & y_B &= -\tfrac{\sqrt{3}}{2}\xi_1 - \tfrac{1}{2}\xi_2 \\
x_C &= -\tfrac{1}{2}\xi_1 - \tfrac{\sqrt{3}}{2}\xi_2 & y_C &= \tfrac{\sqrt{3}}{2}\xi_1 - \tfrac{1}{2}\xi_2
\end{aligned}
$$
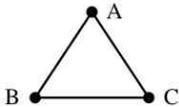
Figure 2.2: A homonuclear trimer in its equilateral configuration.

$$|B\rangle - |C\rangle \qquad |B\rangle - |A\rangle \qquad |C\rangle - |A\rangle \qquad |C\rangle - |B\rangle$$
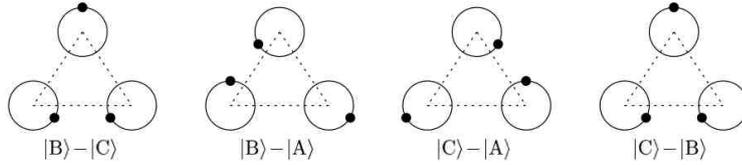
Figure 2.3: A schematic representation of a (counterclockwise) pseudorotation, where subsequent snapshots differ by $2\pi/3$. The corresponding electronic ground states in the tight-binding approximation, are also shown.

The meaning of this coordinate choice is transparent with reference to Fig 2.3: when the atom A is displaced by $\boldsymbol{\xi}$, the displacements of B and C are of equal magnitude $|\boldsymbol{\xi}|$, but pointing in directions rotated by $-2\pi/3$ and $-4\pi/3$, respectively. If we neglect Jahn-Teller coupling beyond linear order, no potential energy is associated to a motion at constant $|\boldsymbol{\xi}|$, which is indeed a free pseudorotation (or a "rotation wave"), also schematized in the succession of snapshots in Fig. 2.3.

In absence of Jahn-Teller coupling, the surface would simply be a parabola, everywhere doubly degenerate. The linear Jahn-Teller splitting is function of $|\boldsymbol{\xi}|$, hence to linear order the electronic eigenvalues are:

$$E_\pm(\boldsymbol{\xi}) \propto |\boldsymbol{\xi}|^2 \pm \text{const } |\boldsymbol{\xi}|. \tag{2.2}$$

This double-valued function is displayed in Fig. 2.4 and has a conical intersection at the origin. The double cone is also called a diabolo (after a spinning toy of the same shape), so the degeneracy points are also called "diabolical". The lowest sheet $E_-(\boldsymbol{\xi})$ has a circular valley of radius $\xi_{\min}$, where a classical particle travels freely (if nonlinear Jahn-Teller coupling is neglected). Nothing exotic happens if the nuclear motion can be considered as classic; but when we quantize the nuclear degrees of freedom, *antiperiodical* boundary conditions have to be imposed for the cyclic motion, as said above.

A simple approximation for the rotovibrational levels is thus:

$$E(u, j) = \left(u + \frac{1}{2}\right)\omega_0 + Aj^2, \tag{2.3}$$
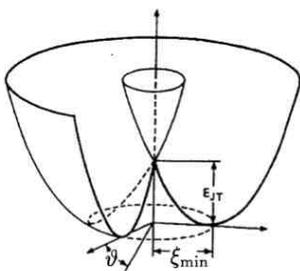
Figure 2.4: The Born-Oppenheimer surface of the Jahn-Teller split doublet: a double-valued function with a conical intersection (a.k.a. diabolical point). The potential minimum is a circle of radius $\xi_{\min}$ centered at the degeneracy point.

14

corresponding to an oscillation of frequency ω_0 and quantum number u, and a two-dimensional internal rotation with rotor constant A. The antiperiodical boundary conditions imply half-odd-integer values for the quantum number j; notice that the ground state ($j = \pm 1/2$) is twofold degenerate.

The pseudorotation term in the spectrum can be compared to Eq. (1.20); the moment of inertia in the prefactor becomes obviously a nuclear rotor, but the spectrum is the same if we identify the inaccessible flux ϕ with *half* a flux quantum ϕ_0 (a.k.a. π flux).

There is no magnetic field in this problem; the flux is purely topological and can be regarded as an obstruction: the nuclear path cannot be contracted without crossing a degeneracy point. It is remarkable that the topological nature of this problem was clearly stated as early as 1963—much earlier than topology became fashionable in electronic structure—by Herzberg and Longuet-Higgins [35], who say verbatim: *"…a conically self-intersecting potential surface has a different topological character from a pair of distinct surfaces which happen to meet at a point. Indeed, if an electronic wave function changes sign when we move round a closed loop in configuration space, we can conclude that somewhere inside the loop there must be a singular point at which the wave function is degenerate"*.

In modern jargon, we would say that the cases $\phi = 0$ and $\phi = \phi_0/2$ are *topologically distinct*; owing T-invariance, other flux values are ruled out. We anticipate that the present problem is revisited in order to introduce the $\mathbb{Z}_2$ topological invariant in Sect. 4.2.3.

The present paradigm also illustrates the robustness of topological properties against smooth deformations. For instance, here we have addressed the ultrasimple tight-binding model, but the ground wavefunction can be "continuously deformed" to the exact correlated wavefunction: topology-wise, the two wavefunctions are essentially the same object, insofar as the conical intersection is present. Notice also that at the conical intersection the Born-Oppenheimer approximation breaks down.

One could also address more general closed paths, according to their winding number round the obstruction. Only paths having the same winding number can be continuously deformed into each other: they are "homotopic".

2.3 Quantization of the surface charge

The pioneering selfconsistent calculations of the electronic structure of surfaces, performed at IBM (Yorktown Heights) and at Bell Labs in the mid 1970s, pointed out the occurrence of quantization of charge at insulating surfaces. After an early paper by V. Heine in 1966 [38], the theorem made its appearance in a 1974 paper by Appelbaum and Hamann [39]. Other papers addressed the issue in the 1970s [40, 41],

but the topological explanation came much later; it will be discussed below, Sect. 5.6.1.

The quantization of surface charge may appear counterintuitive, if one sticks at the idea that a solid is an array of classical charges (ions), as many people still do. Possibly because of its counterintuitive content, this important theorem is surprisingly ignored even by well known specialists in surface physics. The extreme of such ignorance occurs in an invited review paper—which I abstain from quoting— about polar surfaces. The theorem is even more ignored in quantum chemistry, where it addresses end charges in linear polymers.

Electrons are quantum particles, and classical ideas may prove wrong. Solids are *not* assemblies of ions; they are assemblies of atoms, having ionic character only because neighboring atoms have a different electronegativity [42]. At the surface, one has to look at what happens to the bonds.

A simple statement of the theorem is the following. If the bulk of the crystal is centrosymmetric, and if the surface is insulating, then the charge per surface cell may only be an integer or half integer; the surface charge can be nonquantized only if the bulk is noncentrosymmetric, or if the surface is metallic. The theorem is dealt with in detail in Secs. 5.6.1 and 5.7.

Quite often, the actual quantized value is zero because of energy considerations; therefore even polar surfaces are (counterintuitively) neutral under the above two essential hypotheses, which I stress again: the bulk is centrosymmetric, and the surface is insulating. The microscopic mechanism can be understood as an intrinsic surface-state neutralization [42]; however, topology guarantees quantization *independently* of microscopic details.

We nowadays regard bulk-surface correspondence in many phenomena as one of the hallmarks of geometry and topology in condensed matter physics. In modern jargon, I would say that the surface charge of insulators is "topologically protected". More about the bulk-boundary correspondence will be said in Sec. 8.3.4.

2.4 Topological soliton charge in polyacetylene

We consider here trans-poliacetylene (Fig. 2.5): a planar stereoregular polymer, and we focus on its π electrons only. There are two π electrons per monomer; we neutralize them with one positive charge per carbon atom, this making each
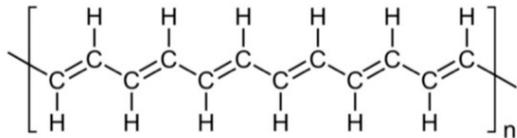
Figure 2.5: Alternant trans-polyacetilene: n monomers are shown. We consider here only π electrons, whose orbitals have a nodal plane in the plane of the molecule.

16
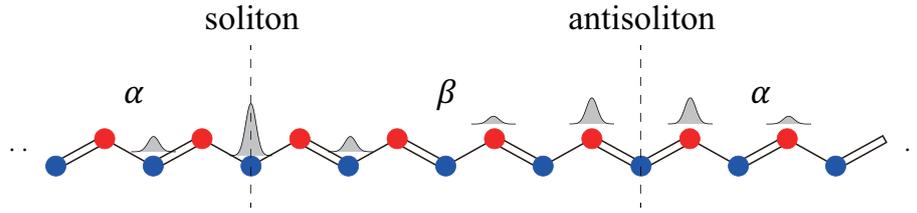
Figure 2.6: Alternation defects in polyacetylene: either two contiguous single bonds (soliton), or two contiguous double bonds (antisoliton).

monomer (a.k.a. one-dimensional crystal cell) neutral.

In the infinite-chain limit the pristine system is insulating (insofar as it is alternant). Let us consider alternation defects, such as those sketched in Fig. 2.6, and called solitons/antisolitons. Such defects are charged; in the limit of an isolated soliton in an infinite chain this charge is topological and equals $\pm e$ if we consider singlet ground states only (with doubly occupied orbitals and zero spin density). This outstanding and very early discovery was made by Su, Schrieffer, and Heeger in 1979 [135].

It became clear much later (1990s) that the reasons behind the topological quantization of the soliton charge are basically the same as for the quantization of the boundary charge in $1d$, or equivalently of the $1d$ polarization, discussed below in Secs. 5.6.1 and 5.7.

The theory of polarization has an equivalent formulation in terms of Wannier functions (Sec. 5.3.3). The quantization of both the end charge in a linear polymer and of the soliton charge in polyacetylene have a simple explanation in terms of Wannier-function counting [126].

2.5 Integer quantum Hall effect

2.5.1 Classical theory (Drude-Zener)

We consider any $2d$ system, in the setup shown in Fig. 2.7. If dissipation is accounted for by a single relaxation time τ, the Newton equation of motion for a single carrier of mass m and charge $-e$, is

$$m\left(\frac{d\mathbf{v}}{dt} + \frac{1}{\tau}\mathbf{v}\right) = -e\left(\mathbf{E} + \frac{1}{c}\mathbf{v} \times \mathbf{B}\right). \tag{2.4}$$

Setting $d\mathbf{v}/dt = 0$ we get the steady-state solution:

$$\mathbf{v} = -\frac{e\tau}{m}\left(\mathbf{E} + \frac{1}{c}\mathbf{v} \times \mathbf{B}\right). \tag{2.5}$$
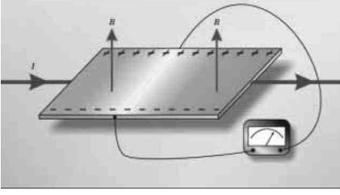
Figure 2.7: Hall effect in a $2d$ system. The **E** field is applied along x, while the **B** field is along z. The system is shorted in the y direction; the current **j** has both longitudinal (x) and transverse, a.k.a. Hall (y) components.

In terms of the cyclotron frequency $\omega_c = \frac{eB}{mc}$ the solution with $E_y = 0$ is

$$
\begin{aligned}
v_x &= -\frac{e\tau}{m}E_x - \omega_c\tau v_y \\
v_y &= \omega_c\tau v_x.
\end{aligned}
\tag{2.6}
$$

If n is the carrier density, the current is $\mathbf{j} = -ne\,\mathbf{v}$:

$$
\begin{aligned}
j_x &= \frac{ne^2\tau}{m}E_x - \omega_c\tau j_y \\
j_y &= \omega_c\tau j_x.
\end{aligned}
\tag{2.7}
$$

in zero **B** field we retrieve the standard Drude (diagonal) conductivity:

$$
j_x = \sigma_0 E_x, \qquad \sigma_0 = \frac{ne^2\tau}{m},
\tag{2.8}
$$

while for $\mathbf{B} \neq 0$ the conductivity tensor is:

$$
\begin{aligned}
j_x &= \frac{\sigma_0}{1 + (\omega_c\tau)^2}E_x = \sigma_{xx}E_x \\
j_y &= \frac{\omega_c\tau\sigma_0}{1 + (\omega_c\tau)^2}E_x = \sigma_{yx}E_x.
\end{aligned}
\tag{2.9}
$$

Inversion of the conductivity tensor

$$
\rho_{xx} = \frac{\sigma_{xx}}{\sigma_{xx}^2 + \sigma_{yx}^2} \qquad \rho_{xy} = \frac{\sigma_{yx}}{\sigma_{xx}^2 + \sigma_{yx}^2}
\tag{2.10}
$$

provides a remarkably simple expression for the longitudinal and transverse resistivity

$$
\rho_{xx} = 1/\sigma_0 = \frac{m}{ne^2\tau}, \quad \rho_{xy} = \frac{m\omega_c}{ne^2} = \frac{1}{nec}B.
\tag{2.11}
$$

The Hall resistivity is therefore linear in B and independent of both mass and relaxation time; more accurately, since we may consider even carriers of *positive* charge e ("holes"), its sign does depend on the carrier charge. Notice also that in the nondissipative regime ($\tau \gg 1/\omega_c$) both ρ_{xx} and σ_{xx} vanish.

18

If we write n as N/A (number of carriers per unit area), then

$$\rho_{xy} = \frac{1}{nec}B = \frac{\phi}{Nec}, \qquad (2.12)$$

where $\phi = AB$ is the magnetic flux through area A. Although we are still at a purely classical level, it is instructive to multiply and divide ρ_{xy} by h. We may thus identically write

$$\rho_{xy} = \frac{1}{\nu}\frac{h}{e^2}, \quad \nu = \frac{N\phi_0}{\phi}. \qquad (2.13)$$

Here $\phi_0 = hc/e$ is the flux quantum, as defined above. The dimensionless quantity ν, called the filling factor, equals the ratio between the number of electrons N and the number of flux quanta ϕ/ϕ_0. Eq. (2.13) expresses the transverse resistivity in terms of the natural resistance unit h/e^2. Since 1990 this is a new metrology standard, accurate to more than nine figures: 1 klitzing = h/e^2 = 25812.807557(18) ohm.

In $2d$ resistance and resistivity have the same dimensions, and coincide in the transverse case. We write therefore the Hall resistance as

$$V_y/I_x = R_{\mathrm{H}} \equiv R_{xy} = \frac{1}{\nu}\frac{h}{e^2}. \qquad (2.14)$$

Upon obvious dimensionality arguments, even in the quantum case the Hall resistance can be written in this way; but then the concentration- and B-dependence of ν are expected to be very different from the simple monothonical form of Eq. (2.13).

One further important relationship is worth deriving. We start casting the transverse conductivity tensor as $\sigma^{(-)}_{\alpha\beta} = \sigma_{\mathrm{H}}\,\epsilon_{\alpha\beta}$, where $\epsilon_{\alpha\beta}$ is the antisymmetric tensor. The Hall cutrrent reads

$$j_\alpha = \sigma_{\mathrm{H}}\,\epsilon_{\alpha\beta}E_\beta, \qquad (2.15)$$

where sum on the repeated indices is understood. Then continuity equation yields

$$-e\frac{\partial n}{\partial t} = -\frac{\partial j_\alpha}{\partial r_\alpha} = -\sigma_{\mathrm{H}}\,\epsilon_{\alpha\beta}\frac{\partial E_\beta}{\partial r_\alpha} = \frac{1}{c}\frac{\partial B}{\partial t}, \qquad (2.16)$$

where the last equality owes to Maxwell's equations. Hence

$$\frac{\partial n}{\partial B} = -\frac{\sigma_{\mathrm{H}}}{ec}. \qquad (2.17)$$

This is clearly consistent with Eq. (2.11), obtained within the classical Drude-Zener theory: in fact $\sigma_{\mathrm{H}} = -1/\rho_{xy} = -nec/B$. But here we have derived Eq. (2.17) in a very general way, and it therefore holds even in the quantum Hall regime, where σ_{H} is B-independent: see Fig. 2.9. In the quantum Hall context Eq. (2.17) goes under the name of Streda formula [43]; it is derived in an alternative way below, Eq. (4.66).
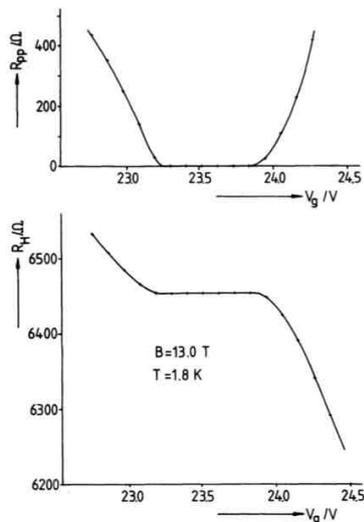
Figure 2.8: The original figure from von Klitzing et al. Ref. [44]. The gate voltage V_g was supposed to control the carrier density. Instead, the Hall resistance is quantized and insensitive to V_g over a large interval; over the same interval, the longitudinal resistance vanishes.

2.5.2 Landau levels

In quantum mechanics, the Schrödinger equation for an electron in $2d$ subject to a perpendicular **B** field (and in a flat potential) can be exactly dealt with, both in the Landau gauge and in the central gauge. The spectrum is discrete $\varepsilon_n = (n + \frac{1}{2})\hbar\omega_c$. We define the magnetic length as $\ell = (\hbar c/eB)^{1/2}$; it diverges in the zero-field limit, and is of the order of 100 angstrom in a typical quantum Hall experiment. In the Landau gauge ($A_x = By, A_y = 0$) the eigenfunctions with energy n are

$$\psi_{nk}(x, y) \propto e^{ikx}\chi_n(y - \ell^2 k), \tag{2.18}$$

where $\chi_n(y)$ are harmonic oscillator eigenfunctions with frequency ω_c. Each LL is infinitely degenerate (one eigenfunction for each k). For a system of area A, the number of states in each level is $\mathcal{N} = A/(2\pi\ell^2)$; this has a simple form in terms of the magnetic flux ϕ through area A: $\mathcal{N} = \phi/\phi_0$.

If we now consider N noninteracting electrons, the lowest LL is completely filled when $N = \mathcal{N}$; more generally, one expects a periodicity in the filling factor $\nu = N/\mathcal{N} = N\phi/\phi_0$, whenever ν crosses an integer value, in most physical properties.

2.5.3 The experiment

The Hall resistance of a noninteracting $2d$ electron gas has been computed quantum-mechanically by Ando in 1974 [45]. The result, when expressed as in Eq. (2.14) showed indeed oscillations in ν with integer period. The experiment, performed by von Klitzing and collaborators in 1980 [44], provided qualitatively different and very surprising results, shown in Fig. 2.8. The discovery of the quantum Hall effect

triggered a revolution with far reaching consequences in electronic structure theory at large; Klaus von Klitzing was awarded the Nobel prize in 1985.

In the original experiment, the $2d$ electrons were confined by a MOSFET (metal-oxide-semi-conductor field-effect transistor); later, higher mobilities were obtained at semiconductor heterojunctions. Fig. 2.8 shows a very robust plateau, where $R_{xx} = 0$ and $R_{xy} = 6453.3 \pm 0.1$, corresponding to the filling factor $\nu = 4$. The accuracy in the quantized R_{xy} value is clearly far beyond the experimental control of the carrier concentration and of the **B** field uniformity over the sample. A novel, qualitatively different, state of matter was discovered. In modern jargon, the plateaus are "topologically protected".

A modern realization of the integer quantum Hall effect is shown in Fig. 2.9, where ρ_{xx} and ρ_{xy} are plotted as a function of the magnetic field. The plateau quantization is accurate to nine figures. The $2d$ electron gas is typically confined at a GaAs/GaAlAs heterojunction. The $\nu = 1$ value is achieved above $\simeq 10$ tesla; at low field (high ν) the system becomes dissipative ($\rho_{xx} > 0$), while the classical linear behavior of ρ_{xy} is recovered; the slope depends on electron concentration n.

2.5.4 Early theoretical interpretation

The breakthrough paper, by Laughlin, appeared as early as 1981 [46]. This is a remarkably concise paper (two pages) which, in retrospect, is based on topological arguments. One key ingredient of the theory is *disorder*: in fact, the quantum Hall effect becomes less spectacular for very "clean" samples, while some "dirtyness"
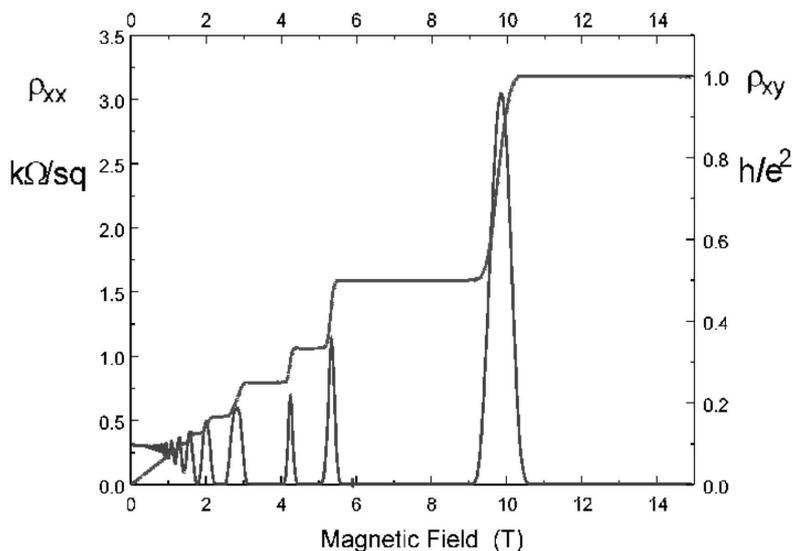
Figure 2.9: A modern realization of the integer quantum Hall effect

21

enhances the effect.

Laughlin devised a Gedankenexperiment based on the setup shown in Fig. 2.10. The corresponding $2d$ Schrödinger Hamiltonian in the Landau gauge is

$$H(\varphi) = \frac{1}{2m}\left[\left(p_x + \frac{e}{c}A_x\right)^2 + p_y^2\right] + eEy + \mathscr{V}(x,y), \qquad (2.19)$$

where E is the field across the ribbon, and $\mathscr{V}$ is an arbitrary substrate potential. The addition of a constant vector potential along x, $A_x \to A_x + \Delta A$, in Eq. (2.19) corresponds to threading a flux $\varphi = L\,\Delta A$ through the cylinder; we use the symbol φ, not to be confused with the real magnetic flux ϕ normal to the surface.

Similarly to what discussed in Sect. 1.6, the eigenvalues acquire a φ dependence. According to Eq. (1.21), if $\varepsilon_n(\varphi)$ is the n-th eigenvalue the current transported by the corresponding eigenstate is $I_x = -c\,\partial\varepsilon_n(\varphi)/\partial\varphi$. For an independent-particle system with N carriers the current is thus

$$I_x = -c\frac{\partial U(\varphi)}{\partial\varphi}, \qquad (2.20)$$

where $U(\varphi)$ is the total energy of the system. Implicitly, we are assuming a dissipationless system.

The expression in Eq. (2.20) for the current is remarkably simple, general, and robust: it does not depend on the substrate potential $\mathscr{V}(x,y)$, nor the number N of carriers, and not even on their mass m. But for a disordered potential the eigenstates come in two kinds: localized and extended. The latter ones are phase-coherent round the loop, while the former are exponentially localized for $L \to \infty$. The localized states are insensitive to the flux insertion (like the OBCs eigenstates in Sect. 1.6), and the whole current is carried by the delocalized ones. Therefore Eq. (2.20) provides a nonzero result insofar at least one of the occupied eigenstates in the disordered sample is extended, i.e. phase-coherent round the loop; besides this, the number and nature of the current-carrying states is irrelevant. It is therefore crucial to address the nature of the single-particle eigenstates in a quantum Hall sample.
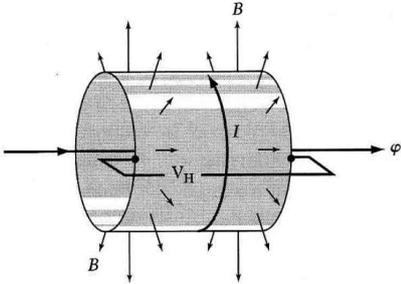
Figure 2.10: Geometry for Laughlin's Gedankenexperiment. A $2d$ channel is bent into a loop of circumference L, and a magnetic **B** field of constant magnitude pierces the cylinder normal to the surface. A current $I \equiv I_x$ circles the loop; $V_H \equiv V_y$ is the Hall voltage. The loop may be threaded by the inaccessible flux φ.
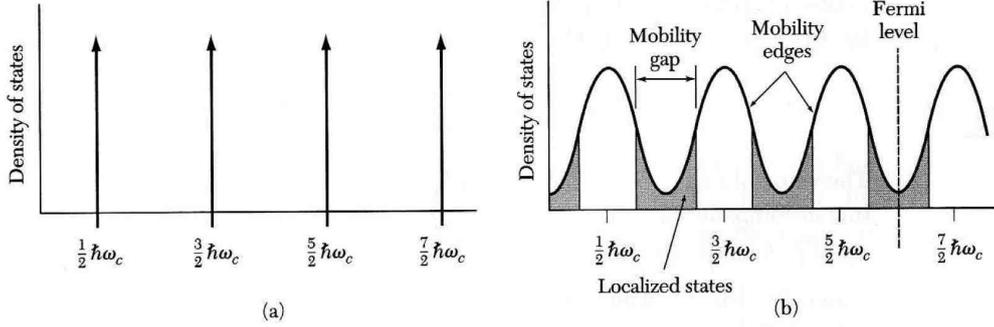
Figure 2.11: The density of states for a $2d$ system of noninteracting carriers. (a) Clean system, with zero substrate potential. (b) Actual sample, in presence of substrate disorder and impurities.

For a clean sample (flat substrate) the LLs are sharp, all eigenstates are extended (in the Landau gauge), and the density of states is a series of delta functions, shown in Fig. 2.11 (a); the weight of each delta is ϕ/ϕ_0. In presence of disorder, the deltas broaden into alternating bands of localized and extended states, as sketched in Fig. 2.11 (b).

The electron fluid is in the quantum Hall regime whenever the Fermi level falls in a region of localized states. Therefore $\sigma_{xx} = 0$ (the fluid is a "quantum Hall insulator"), and $\rho_{xx} = 0$ (transport is dissipationless).

We now imagine to adiabatically increase the vector potential by an amount $\Delta A = \phi_0/L$, where ϕ_0 is a flux quantum: all of the current-carrying states are mapped back into themselves, while the localized ones are unaffected. Hence the ground state has the same energy; nonetheless Eq. (2.20) implies $U(\varphi+\phi_0)-U(\varphi) \simeq -\phi_0 I_x/c \neq 0$. This is only possible if an integer number of electrons is transferred from one cylinder edge to the other, each of them contributing the energy eV_y. If we call $-\nu$ such integer number, the relationship is then

$$\phi_0 I_x/c = \nu e V_y; \qquad R_{\mathrm{H}} = V_y/I_x = \frac{\phi_0}{\nu c e} = \frac{1}{\nu}\frac{h}{e^2}. \qquad (2.21)$$

The flux φ acts therefore as a charge pump; the pump cycle is one flux quantum ϕ_0.

Ideally, the sample ground state can be continuously "deformed" from dirty to clean. Insofar as the Fermi level stays is in a region of nonconducting states, the (topological) integer ν cannot change, even if the number of current carrying states does obviously change. The identification of ν with the number of filled LLs comes from the clean-sample limit, which is exactly soluble. Setting $\mathscr{V}(x,y) = 0$ the eigenfunctions of Eq. (2.19) are

$$\psi_{nk}(x,y) = \mathrm{e}^{ikx}\chi_n(y-y_0), \quad y_0 = \ell^2 k - \frac{cE}{\omega_c B}. \qquad (2.22)$$

23

For a finite L, the allowed k's are integer multiples of $2\pi/L$ and the corresponding centers y_0 are spaced by $2\pi\ell^2/L = L\phi_0/\phi$. Threading a flux φ shifts y_0 linearly in φ; when φ equals one flux quantum each eigenfunction goes over to the next. Therefore one carrier is shifted for each n; the integer index ν measures therefore the number of occupied LLs. Similar arguments can be reformulated in different gauges and in different geometries [47, 48].

Chapter 3

Berryology

3.1 Distance and connection

Berryology is used here as synonymous for geometry in nonrelativistic quantum mechanics. Topological (i.e. quantized) quantities are defined via geometrical quantities, analogously to what we did above in Sect. 1.2. But many important quantities (notably the Berry phase) are merely geometrical [25].

The founding concept of geometry is distance. Let $|\Psi_1\rangle$ and $|\Psi_2\rangle$ be two quantum states in the same Hilbert space: we adopt for their distance an appropriately modified form of the Bures distance [49]:

$$D_{12}^2 = -\ln |\langle \Psi_1 | \Psi_2 \rangle|^2. \tag{3.1}$$

This distance vanishes when the states $|\Psi_1\rangle$ and $|\Psi_2\rangle$ coincide, while it diverges when the states $|\Psi_1\rangle$ and $|\Psi_2\rangle$ are orthogonal. Actually, Eq. (3.1) defines a pseudodistance, since it violates one of the axioms in calculus textbooks. Such violation does not make any harm in the present context.

The states $|\Psi_1\rangle$ and $|\Psi_2\rangle$ are defined up to an arbitrary phase factor: fixing this factor amounts to a gauge choice. Eq. (3.1) is clearly gauge-invariant. The pseudodistance in Eq. (3.1) can equivalently be rewritten as

$$D_{12}^2 = -\ln\langle \Psi_1 | \Psi_2 \rangle - \ln\langle \Psi_2 | \Psi_1 \rangle, \tag{3.2}$$

where the two terms are not separately gauge-invariant. While the distance is obviously real, each of the two terms in Eq. (3.2) is in general a complex number. If we write

$$\langle \Psi_1 | \Psi_2 \rangle = |\langle \Psi_1 | \Psi_2 \rangle| e^{i\varphi_{21}}, \tag{3.3}$$

then the imaginary part of each of the two terms in Eq. (3.2) assumes a transparent meaning:

$$-\text{Im} \ln \langle \Psi_1 | \Psi_2 \rangle = \varphi_{12}, \qquad \varphi_{21} = -\varphi_{12}. \tag{3.4}$$

Besides distance, an additional geometrical concept is therefore needed: the connection, which fixes the relative phases between two states in the Hilbert space.

The connection is arbitrary and cannot have any physical meaning by itself. Nonetheless, after the 1984 groundbreaking paper by Michael Berry [33], several physical observables are expressed in terms of the connection and related quantities.

3.2 Geometry in a parameter space

Let us assume that a generic time-independent quantum Hamiltonian has a parametric dependence. The Schrödinger equation is

$$H(\boldsymbol{\xi})|\Psi(\boldsymbol{\xi})\rangle = E(\boldsymbol{\xi})|\Psi(\boldsymbol{\xi})\rangle, \tag{3.5}$$

where the d-dimensional real parameter $\boldsymbol{\xi}$ is defined in a suitable domain of $\mathbb{R}^d$: a $2d$ $\boldsymbol{\xi}$ has been chosen for display in Fig. 3.1. In most of this Chapter we discuss the most general case, and therefore we do not specify which quantum system is described by this Hamiltonian, nor what the physical meaning of the parameter $\boldsymbol{\xi}$ is.

In the subsequent Chapters $|\Psi(\boldsymbol{\xi})\rangle$ will be identified with either a single-particle wavefunction (a.k.a. orbital) or a many-electron wavefunction. As for the parameter $\boldsymbol{\xi}$, it may be a nuclear coordinate, a phase angle, a magnetic flux, a Bloch vector, a momentum, and more: it could therefore have various dimensions. Sometimes, the parameter $\boldsymbol{\xi}$ is called the "slow variable", while the electronic coordinates are the "fast variable". In the final part of this Chapter, from Sect. 3.9.1 onwards, we address the special case where the parameter $\boldsymbol{\xi}$ is identified with a Bloch vector $\mathbf{k}$.

The state vectors $|\Psi(\boldsymbol{\xi})\rangle$ are all supposed to be normalized and to reside in the same Hilbert space: this amounts to saying that the wavefunctions are supposed to obey $\boldsymbol{\xi}$-independent boundary conditions. We focus on the ground state $|\Psi_0(\boldsymbol{\xi})\rangle$, and we assume it to be nondegenerate for $\boldsymbol{\xi}$ in some domain of $\mathbb{R}^d$.

We have referred to choosing the phase of $|\Psi_0(\boldsymbol{\xi})\rangle$ as to the choice of the gauge: the semantic is a bit ambiguous. In presence of magnetic fields, we may change the magnetic gauge: this changes the Hamiltonian and the eigenfunctions. Once the magnetic gauge—hence the Hamiltonian—is fixed, we still remain with the phase arbitrariness referred to above. All measurable quantities (e.g. expectation values) are obviously gauge-invariant (in both senses), but the reverse is also true: all gauge-invariant properties are—at least in principle—measurable. This is the outstanding message of the famous Berry's paper [33].

We have defined above the distance and the phase difference between the ground eigenstates at two different $\boldsymbol{\xi}$ points:

$$D^2_{\boldsymbol{\xi}_1\boldsymbol{\xi}_2} = -\ln|\langle\Psi_0(\boldsymbol{\xi}_1)|\Psi_0(\boldsymbol{\xi}_2)\rangle|^2. \tag{3.6}$$

$$\Delta\varphi_{12} = -\operatorname{Im}\ln\langle\Psi_0(\boldsymbol{\xi}_1)|\Psi_0(\boldsymbol{\xi}_2)\rangle\ . \tag{3.7}$$

For any given choice of the two states, Eq. (3.7) provides a $\Delta\varphi_{12}$ which is unique modulo 2π, except in the very special case that the states are orthogonal.

3.3 Berry phase

We have already observed that the phase difference $\Delta\varphi_{12}$ between any two states is gauge-dependent and cannot have any physical meaning by itself. Matters are quite different when we consider the *total* phase difference along a closed path which joins several points in a given order, as shown in Fig. 3.2:

$$
\begin{aligned}
\gamma &= \Delta\varphi_{12} + \Delta\varphi_{23} + \Delta\varphi_{34} + \Delta\varphi_{41}\\
&= -\operatorname{Im}\ln\langle\Psi_0(\boldsymbol{\xi}_1)|\Psi_0(\boldsymbol{\xi}_2)\rangle\langle\Psi_0(\boldsymbol{\xi}_2)|\Psi_0(\boldsymbol{\xi}_3)\rangle\ \times\\
&\quad\times\ \langle\Psi_0(\boldsymbol{\xi}_3)|\Psi_0(\boldsymbol{\xi}_4)\rangle\langle\Psi_0(\boldsymbol{\xi}_4)|\Psi_0(\boldsymbol{\xi}_1)\rangle.
\end{aligned}
\tag{3.8}
$$

It is now clear that all the gauge-arbitrary phases cancel in pairs, such as to make the overall phase γ a gauge–invariant quantity. The above simple–minded algebra leads to a result of overwhelming physical importance: in fact, a gauge–invariant quantity is potentially a physical observable. We stress once more this fact, which is indeed the revolutionary message of Berry's celebrated paper, appeared in 1984 [33, 50].

Next we consider a smooth closed curve C in the parameter domain, such as in Fig. 3.3, and we discretize it with a set of points on it. Using Eq. (3.7), we write the phase difference between any two contiguous points as

$$\mathrm{e}^{-i\Delta\varphi} = \frac{\langle\Psi_0(\boldsymbol{\xi})|\Psi_0(\boldsymbol{\xi}+\Delta\boldsymbol{\xi})\rangle}{|\langle\Psi_0(\boldsymbol{\xi})|\Psi_0(\boldsymbol{\xi}+\Delta\boldsymbol{\xi})\rangle|}. \tag{3.9}$$

If we further assume that the gauge is so chosen that the phase varies in a *differentiable* way along the path, then from Eq. (3.9) we get to leading order in $\Delta\boldsymbol{\xi}$:

$$-i\Delta\varphi \simeq \langle\Psi_0(\boldsymbol{\xi})|\nabla_{\boldsymbol{\xi}}\Psi_0(\boldsymbol{\xi})\rangle\cdot\Delta\boldsymbol{\xi}. \tag{3.10}$$

$\bullet\ |\psi(\boldsymbol{\xi}_3)\rangle$

$\bullet\ |\psi(\boldsymbol{\xi}_2)\rangle$

$\bullet\ |\psi(\boldsymbol{\xi}_1)\rangle$ $\bullet\ |\psi(\boldsymbol{\xi}_4)\rangle$
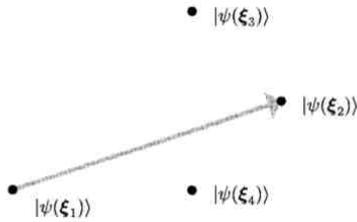
Figure 3.1: State vectors in the two-dimensional $\boldsymbol{\xi}$-space. The phase difference between two of them is defined as $\Delta\varphi_{12} = -\operatorname{Im}\ln\langle\Psi_0(\boldsymbol{\xi}_1)|\Psi_0(\boldsymbol{\xi}_2)\rangle$, and their pseudodistance as $D_{12}^2 = -\ln|\langle\Psi_0(\boldsymbol{\xi}_1)|\Psi_0(\boldsymbol{\xi}_2)\rangle|^2$.
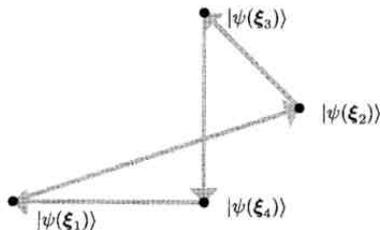
Figure 3.2: A closed path joining four states in $\boldsymbol{\xi}$-space.

In the limiting case of a set of points which becomes dense on the continuous path, the total phase difference γ converges to a circuit integral:

$$\gamma = \sum_{s=1}^{M} \Delta\varphi_{s,s+1} \longrightarrow \oint_C \boldsymbol{\mathcal{A}}(\boldsymbol{\xi}) \cdot d\boldsymbol{\xi}, \tag{3.11}$$

where $\boldsymbol{\mathcal{A}}(\boldsymbol{\xi})$ is called the Berry connection:

$$\boldsymbol{\mathcal{A}}(\boldsymbol{\xi}) = i \langle \Psi_0(\boldsymbol{\xi}) | \nabla_{\boldsymbol{\xi}} \Psi_0(\boldsymbol{\xi}) \rangle. \tag{3.12}$$

Since the state vectors are assumed to be normalized at any $\boldsymbol{\xi}$, the connection is *real*; we can therefore equivalently write

$$\boldsymbol{\mathcal{A}}(\boldsymbol{\xi}) = -\mathrm{Im} \, \langle \Psi_0(\boldsymbol{\xi}) | \nabla_{\boldsymbol{\xi}} \Psi_0(\boldsymbol{\xi}) \rangle. \tag{3.13}$$

A number of manifestations of the Berry phase occurring in molecular and condensed matter phenomena will be discussed in the present Notes. They are obviously discussed in many reviews papers; here we quote Refs. [50, 51, 52, 3, 13]. Let me also quote the very recent book by David Vanderbilt [24].

At this point it is also worth emphasizing that in computational physics there are no derivatives. The ground state $|\Psi_0(\boldsymbol{\xi})\rangle$ is generally found by diagonalizing a matrix on a finite set of $\boldsymbol{\xi}$ points, and the phase (i.e. the gauge) is chosen by the
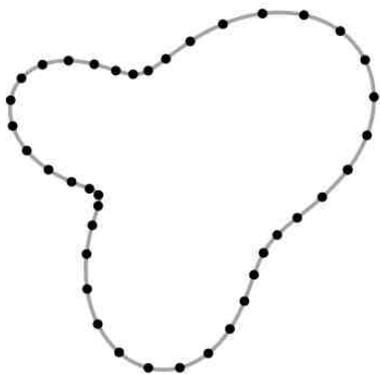
Figure 3.3: A smooth closed curve C in $\boldsymbol{\xi}$-space, and its discretization.

28

diagonalization routine; the phase is therefore nonsmooth and possibly even random. The discrete form in Eqs. (3.7) and (3.11) at finite M is the one universally used in numerical work; it is unaffected by any erratic phase factor. More will be said about discretizing the Berry phase in Sect. 3.9.4.

3.4 Connection and curvature

The loop integral of the Berry connection (i.e. the Berry phase γ) is non trivial in two cases: either the curl of $\boldsymbol{\mathcal{A}}(\boldsymbol{\xi})$ is nonzero, or the curl is zero but the curve C is not in a simply connected domain. In the former case, we can invoke Stokes theorem; the formulation is very simple when $\boldsymbol{\xi}$ is a $3d$ parameter. If C is the boundary of a surface Σ (i.e. $C \equiv \partial\Sigma$), and the curl of $\boldsymbol{\mathcal{A}}(\boldsymbol{\xi})$ is regular on Σ, then Stokes' theorem in $3d$ (Green's theorem in $2d$) reads

$$\gamma = \oint_{\partial\Sigma} \boldsymbol{\mathcal{A}}(\boldsymbol{\xi}) \cdot d\boldsymbol{\xi} = \int_{\Sigma} \boldsymbol{\Omega}(\boldsymbol{\xi}) \cdot \mathbf{n}\, d\sigma, \tag{3.14}$$

where $\boldsymbol{\Omega}$ is the Berry curvature, defined as

$$\begin{aligned} \boldsymbol{\Omega}(\boldsymbol{\xi}) &= \nabla_{\boldsymbol{\xi}} \times \boldsymbol{\mathcal{A}}(\boldsymbol{\xi}) = -\mathrm{Im}\, \langle\nabla_{\boldsymbol{\xi}}\Psi_0(\boldsymbol{\xi})| \times |\nabla_{\boldsymbol{\xi}}\Psi_0(\boldsymbol{\xi})\rangle \\ &= i\langle\nabla_{\boldsymbol{\xi}}\Psi_0(\boldsymbol{\xi})| \times |\nabla_{\boldsymbol{\xi}}\Psi_0(\boldsymbol{\xi})\rangle, \end{aligned} \tag{3.15}$$

with the usual meaning of the cross product between three-component bra and ket states. Equation (3.14) may be spelled out by saying that the curvature is the Berry phase per unit area of Σ.

For $d \neq 3$ the Berry curvature is conveniently written as the $d \times d$ antisymmetric matrix

$$\boldsymbol{\Omega}_{\alpha\beta}(\boldsymbol{\xi}) = -2\,\mathrm{Im}\, \langle\partial_\alpha\Psi_0(\boldsymbol{\xi})|\partial_\beta\Psi_0(\boldsymbol{\xi})\rangle; \tag{3.16}$$

Greek subscripts are Cartesian coordinates throughout, and $\partial_\alpha = \partial/\partial\xi_\alpha$. The Stokes theorem can still be applied, generalizing Eq. (3.14) to

$$\gamma = \frac{1}{2}\int_{\Sigma} d\xi^\alpha \wedge d\xi^\beta\, \boldsymbol{\Omega}_{\alpha\beta}(\boldsymbol{\xi}). \tag{3.17}$$

More will be said below about this kind of notation and its convenience.

The Berry connection is also known as "gauge potential", and the Berry curvature as "gauge field" [52]. It is worth pointing out that the former is gauge-dependent, while the latter is gauge-invariant and therefore corresponds in general to a measurable quantity, even before any integration. The two quantities play (in $\boldsymbol{\xi}$-space) a similar role as the vector potential and the magnetic field in elementary magnetostatics: $\mathbf{A}(\mathbf{r})$ is gauge-dependent, nonmeasurable; $\mathbf{B}(\mathbf{r}) = \nabla_{\mathbf{r}} \times \mathbf{A}(\mathbf{r})$ is gauge-invariant, measurable.

The Berry phase γ, defined as the integral over a closed curve C of the connection, is gauge invariant only modulo 2π. This indeterminacy is resolved by Eqs. (3.14) and (3.17) whenever the curve C is the boundary $\partial\Sigma$ of a surface Σ where the curvature is regular. In fact, the curvature is gauge-invariant and has no modulo 2π indeterminacy.

3.5 Chern number

The rhs of Eqs. (3.14) and (3.17) is the flux of the Berry curvature on the surface Σ; such flux remains meaningful even on a closed surface (e.g. a sphere or a torus), in which case $\partial\Sigma$ is the empty set. The key result is that such an integral is quantized. Here we limit ourselves to $3d$, where we identify Σ with the sphere S^2 (Fig. 3.4):

$$\frac{1}{2\pi} \int_{S^2} \mathbf{\Omega}(\boldsymbol{\xi}) \cdot \mathbf{n} \, d\sigma = C_1; \tag{3.18}$$

C_1 is an integer $\in \mathbb{Z}$, called Chern number of the first class.

The proof is based on a similar algebra as for Dirac's theory of the magnetic monopole [51, 53]. The theorem goes sometimes under the name of Gauss-Bonnet-Chern theorem; the analogy with Eq. (1.4) is perspicuous. A specific example is dealt with in detail in Sec. 4.1.

The curvature is regular (and divergence-free) on the closed surface S^2; the lhs of Eq. (3.18) is the flux of $\mathbf{\Omega}(\boldsymbol{\xi})$ across S^2. The integrand $\mathbf{\Omega}(\boldsymbol{\xi})$ is the curl of the connection $\mathcal{A}(\boldsymbol{\xi})$; the latter in general *cannot* be defined as a single-valued function globally on S^2, but only on *patches* of it [51, 53]. To fix the ideas, suppose that $\mathbf{\Omega}(\boldsymbol{\xi})$ is singular at $\boldsymbol{\xi} = 0$, and that S^2 is the spherical surface centered at the origin (Fig. 3.4). We cut this surface at the equator $\xi_z = 0$ and we consider the flux across the two open surfaces:

$$\int_{S^2} \mathbf{\Omega}(\boldsymbol{\xi}) \cdot \mathbf{n} \, d\sigma = \int_{S_+} \mathbf{\Omega}(\boldsymbol{\xi}) \cdot \mathbf{n} \, d\sigma + \int_{S_-} \mathbf{\Omega}(\boldsymbol{\xi}) \cdot \mathbf{n} \, d\sigma. \tag{3.19}$$

We notice that $\partial S_+ = \partial S_- = C$, but the surface normals $\mathbf{n}$ have opposite orientations. From Stokes theorem, Eq. (3.14), we get:

$$\int_{S_\pm} \mathbf{\Omega}(\boldsymbol{\xi}) \cdot \mathbf{n} \, d\sigma = \pm \oint_C \mathcal{A}_\pm(\boldsymbol{\xi}) \cdot d\boldsymbol{\xi} \tag{3.20}$$
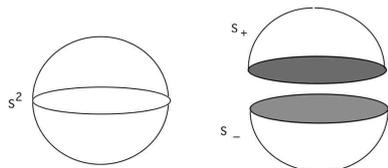
Figure 3.4: A sphere cut at the equator in two hemispheres.

$$\int_{S^2} \boldsymbol{\Omega}(\boldsymbol{\xi}) \cdot \mathbf{n} \, d\sigma = \oint_C \boldsymbol{\mathcal{A}}_+(\boldsymbol{\xi}) \cdot d\boldsymbol{\xi} - \oint_C \boldsymbol{\mathcal{A}}_-(\boldsymbol{\xi}) \cdot d\boldsymbol{\xi} \,. \tag{3.21}$$

The two upper and lower Berry connections $\boldsymbol{\mathcal{A}}_\pm(\boldsymbol{\xi})$ may only differ by a gauge transformation; the rhs of Eq. (3.21) is the difference of two Berry phases on the same path and is necessarily a multiple of 2π. This concludes the proof of Eq. (3.18). We emphasize that the Chern number is a robust topological invariant of the wavefunction, and is at the origin of observable effects.

3.6   Metric

It is expedient to define the ground-state projector (a.k.a. density matrix) and its complement, i.e.

$$\hat{P}(\boldsymbol{\xi}) = |\Psi_0(\boldsymbol{\xi})\rangle\langle\Psi_0(\boldsymbol{\xi})|; \quad \hat{Q}(\boldsymbol{\xi}) = \hat{1} - \hat{P}(\boldsymbol{\xi}). \tag{3.22}$$

Both $\hat{P}(\boldsymbol{\xi})$ and $\hat{Q}(\boldsymbol{\xi})$ are gauge-invariant (for a fixed Hamiltonian).

Starting from Eq. (3.6), the infinitesimal distance is

$$D^2_{\boldsymbol{\xi},\boldsymbol{\xi}+d\boldsymbol{\xi}} = \sum_{\alpha,\beta=1}^d g_{\alpha\beta}(\boldsymbol{\xi}) d\xi_\alpha d\xi_\beta, \tag{3.23}$$

where the metric tensor is easily shown to be

$$\begin{aligned}
g_{\alpha\beta}(\boldsymbol{\xi}) &= \text{Re}\, \langle\partial_\alpha\Psi_0(\boldsymbol{\xi})|\partial_\beta\Psi_0(\boldsymbol{\xi})\rangle \\
&- \langle\partial_\alpha\Psi_0(\boldsymbol{\xi})|\Psi_0(\boldsymbol{\xi})\rangle\langle\Psi_0(\boldsymbol{\xi})|\partial_\beta\Psi_0(\boldsymbol{\xi})\rangle \\
&= \text{Re}\, \langle\partial_\alpha\Psi_0(\boldsymbol{\xi})|\hat{Q}(\boldsymbol{\xi})|\partial_\beta\Psi_0(\boldsymbol{\xi})\rangle;
\end{aligned} \tag{3.24}$$

the projector $\hat{Q}(\boldsymbol{\xi})$ is the same as defined in Eq. (3.22). This quantum metric tensor was first proposed by Provost and Vallee in 1980 [54].

At this point we may compare Eq. (3.24) to Eq. (3.16), noticing that the insertion of $\hat{Q}(\boldsymbol{\xi})$ is irrelevant in the latter, i.e.

$$\boldsymbol{\Omega}_{\alpha\beta}(\boldsymbol{\xi}) = -2\,\text{Im}\, \langle\partial_\alpha\Psi_0(\boldsymbol{\xi})|\hat{Q}(\boldsymbol{\xi})|\partial_\beta\Psi_0(\boldsymbol{\xi})\rangle. \tag{3.25}$$

It is therefore clear that $g_{\alpha\beta}$ and $\boldsymbol{\Omega}_{\alpha\beta}$ are, apart for a trivial -2 factor, the real (symmetric) and the imaginary (antisymmetric) parts of the same tensor, which we are going to call $\mathcal{F}_{\alpha\beta}$ in the following:

$$\mathcal{F}_{\alpha\beta}(\boldsymbol{\xi}) = \langle\partial_\alpha\Psi_0(\boldsymbol{\xi})|\hat{Q}(\boldsymbol{\xi})|\partial_\beta\Psi_0(\boldsymbol{\xi})\rangle. \tag{3.26}$$

The metric-curvature tensor $\mathcal{F}_{\alpha\beta}$ is gauge-invariant. A compact equivalent expression is

$$\mathcal{F}_{\alpha\beta}(\boldsymbol{\xi}) = \text{Tr}\, \{\partial_\alpha\hat{P}(\boldsymbol{\xi})\hat{Q}(\boldsymbol{\xi})\partial_\beta\hat{P}(\boldsymbol{\xi})\}, \tag{3.27}$$

manifestly gauge-invariant and Hermitian.

3.7 Parallel-transport gauge and sum over states

As said above, the main ingredients in Berryology are $\boldsymbol{\xi}$-derivatives $|\partial_{\boldsymbol{\xi}}\Psi_0(\boldsymbol{\xi})\rangle$ of the ground state. To lowest order we have

$$
|\Psi_0(\boldsymbol{\xi}+\Delta\boldsymbol{\xi})\rangle \;\simeq\; \mathrm{e}^{-i\varphi(\boldsymbol{\xi})}\bigg(\,|\Psi_0(\boldsymbol{\xi})\rangle \tag{3.28}
$$
$$
+\; \sum_{n>0}|\Psi_n(\boldsymbol{\xi})\rangle\frac{\langle\Psi_n(\boldsymbol{\xi})|\,\Delta\hat{H}\,|\Psi_0(\boldsymbol{\xi})\rangle}{E_0(\boldsymbol{\xi})-E_n(\boldsymbol{\xi})}\bigg),
$$

where $\Delta\hat{H} = \hat{H}(\boldsymbol{\xi}+\Delta\boldsymbol{\xi}) - \hat{H}(\boldsymbol{\xi})$. The overall phase factor is generally omitted in most textbooks: this corresponds to a specific gauge, called the parallel-transport gauge [3]. The name indicates that the first-order change in $|\Psi_0(\boldsymbol{\xi})\rangle$ is chosen as orthogonal to $|\Psi_0(\boldsymbol{\xi})\rangle$. Within this gauge the Berry connection, Eq. (3.12), is everywhere vanishing. The parallel-transport gauge can be enforced locally at any $\boldsymbol{\xi}$ point, but it cannot be enforced globally along a closed path C if the Berry phase γ for that path is nonzero.

Given that the phase relationship at different $\boldsymbol{\xi}$'s is arbitrary, the phase function $\varphi(\boldsymbol{\xi})$ in Eq. (3.28) allows for gauge freedom. Assuming that $\varphi(\boldsymbol{\xi})$ is a differentiable function, the most general Berry connection, Eq. (3.12), is

$$
\mathcal{A}_\alpha(\boldsymbol{\xi}) = i\langle\Psi_0(\boldsymbol{\xi})|\partial_\alpha\Psi_0(\boldsymbol{\xi})\rangle = \partial_\alpha\varphi(\boldsymbol{\xi}). \tag{3.29}
$$

Indeed, this shows once more that the (arbitrary) connection fixes the phase relationship between infinitesimally close state vectors.

Since our metric-curvature tensor $\mathcal{F}_{\alpha\beta}(\boldsymbol{\xi})$ is a gauge-invariant quantity, we may safely evaluate it in any gauge, including the parallel-transport gauge. The result is

$$
\mathcal{F}_{\alpha\beta}(\boldsymbol{\xi}) \tag{3.30}
$$
$$
= \sideset{}{'}\sum_{n\neq0} \frac{\langle\Psi_0(\boldsymbol{\xi})|\partial_\alpha H(\boldsymbol{\xi})|\Psi_n(\boldsymbol{\xi})\rangle\langle\Psi_n(\boldsymbol{\xi})|\partial_\beta H(\boldsymbol{\xi})|\Psi_0(\boldsymbol{\xi})\rangle}{[E_0(\boldsymbol{\xi})-E_n(\boldsymbol{\xi})]^2}.
$$

This expression shows explicitly that both the curvature and the metric are ill defined and singular wherever the ground state is degenerate with the first excited state. Indeed, this is the main reason why the domain may happen not to be simply connected.

3.8 Time-reversal and inversion symmetries

According to Eq. (3.27) the ground-state projector uniquely determines the curvature

$$
\boldsymbol{\Omega}_{\alpha\beta}(\boldsymbol{\xi}) = -2\,\mathrm{Im}\,\mathrm{Tr}\,\{\partial_\alpha\hat{P}(\boldsymbol{\xi})\hat{Q}(\boldsymbol{\xi})\partial_\beta\hat{P}(\boldsymbol{\xi})\}. \tag{3.31}
$$

It is therefore expedient to analyze the symmetries of the ground-state projector $\hat{P}(\boldsymbol{\xi})$, which coincide with the symmetries of the Hamiltonian; these in turn depend on the nature of the parameter $\boldsymbol{\xi}$. We only address spinless electrons (no spin-orbit coupling).

When $\boldsymbol{\xi}$ is even under time reversal (like e.g. a nuclear coordinate), then T-invariance implies that both $H(\boldsymbol{\xi})$ and $\hat{P}(\boldsymbol{\xi})$ are real for any $\boldsymbol{\xi}$, and Eq. (3.31) warrants that the curvature is everywhere vanishing. The Berry phase γ can be nonzero (modulo 2π) only if the curve C loops around a singularity or, more generally, it does not lie in a simply connected domain; the only allowed values are $\gamma = 0 \bmod 2\pi$ or $\gamma = \pi \bmod 2\pi$. Some consequences of this feature are outlined in Sect. 3.10.5.

When instead $\boldsymbol{\xi}$ is odd under time-reversal (like e.g. a momentum) then T-symmetry requires $\hat{P}(-\boldsymbol{\xi}) = \hat{P}^*(\boldsymbol{\xi})$, therefore

$$\boldsymbol{\Omega}_{\alpha\beta}(-\boldsymbol{\xi}) = -\boldsymbol{\Omega}_{\alpha\beta}(\boldsymbol{\xi}). \tag{3.32}$$

The Berry phase along an inversion-symmetric path vanishes; the Chern number vanishes as well.

Next we switch to inversion symmetry. When evaluating any $\boldsymbol{\xi}$-dependent matrix-element (or trace), inversion of the coordinates at fixed $\boldsymbol{\xi}$ is equivalent to keeping the coordinates fixed and inverting $\boldsymbol{\xi}$. This statement holds whether $\boldsymbol{\xi}$ is a coordinate or a momentum; both are in fact odd under inversion. Therefore inversion symmetry implies $\hat{P}(-\boldsymbol{\xi}) = \hat{P}(\boldsymbol{\xi})$, and

$$\boldsymbol{\Omega}_{\alpha\beta}(-\boldsymbol{\xi}) = \boldsymbol{\Omega}_{\alpha\beta}(\boldsymbol{\xi}). \tag{3.33}$$

If both time-reversal and inversion symmetry are present, then the Berry curvature is everywhere vanishing. The Berry phase can be only zero or π; the latter case requires a domain which is not simply connected, as above.

Crucial to the above arguments is the fact that the double derivative appearing in Eq. (3.31) are even under either time-reversal or inversion. Summarizing the symmetry results, for the case where $\boldsymbol{\xi}$ is a momentum: a non vanishing Chern number can only occur in absence of T-symmetry, but may occur even in inversion-symmetric cases.

3.9   Bloch geometry

3.9.1   Bloch orbitals

We have remained very general so far. The case where the parameter $\boldsymbol{\xi}$ coincides with the Bloch vector $\mathbf{k}$ bears a particular relevance in the context of the present

Notes. In the framework of first-principle calculations for crystalline systems, the Bloch states $|\psi_{j\mathbf{k}}\rangle$ are the KS orbitals. The domain where the $\mathbf{k}$ parameter varies (the reciprocal cell or, equivalently, the Brillouin zone, BZ) has the geometry of a torus in $1d$, $2d$, $3d$.

The Bloch orbital of the j-th band is $|\psi_{j\mathbf{k}}\rangle = \mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}}|u_{j\mathbf{k}}\rangle$, where $|u_{j\mathbf{k}}\rangle$ are cell-periodical, and are eigenfunctions of the Hamiltonian $H_{\mathbf{k}} = \mathrm{e}^{-i\mathbf{k}\cdot\mathbf{r}}H\mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}}$. The normalization of the $|\psi_{j\mathbf{k}}\rangle$ is not uniform across the literature; we adopt here what has become the current standard, i.e. they are normalized in the crystal cell. In other papers, including the classical review of Ref. [1], they are instead normalized like plane waves. The relative phases at different $\mathbf{k}$'s are arbitrary. Whenever possible, it is customary to enforce the so-called periodic gauge $|\psi_{j\mathbf{k}+\mathbf{G}}\rangle = |\psi_{j\mathbf{k}}\rangle$, which implies

$$|u_{j\mathbf{k}+\mathbf{G}}\rangle = \mathrm{e}^{-i\mathbf{G}\cdot\mathbf{r}}|u_{j\mathbf{k}}\rangle, \tag{3.34}$$

where (here and throughout) $\mathbf{G}$ is a reciprocal vector. We stress, however, that in topologically nontrivial crystals it is generally *impossible* to adopt a periodic gauge.

If we set the Born-von-Kàrmàn periodicity over a "supercell" (multiple of the elementary cell), then the Bloch vectors are discrete. At the independent-particle level the ground state is a Slater determinant of doubly occupied Bloch orbitals $|\psi_{j\mathbf{k}}\rangle$, from the whole BZ in insulators, or from a portion of it (the so-called Fermi volume) in metals. When spin-orbit interaction is neglected we may deal with "spinless electrons". For a singlet ground state trivial factors of two have to be restored afterwards; even when the ground state is not a singlet (e.g. in magnetic or antiferromagnetic materials), the relevant spinful quantities are easily built out of the spinless ones.

The Slater determinant is invariant by unitary mixing of the orbitals between themselves: they can be mixed for different $\mathbf{k}$'s (only in insulators), thus yielding e.g. the Wannier functions, discussed in Sect. 5.3.3. The $|\psi_{j\mathbf{k}}\rangle$—and equivalently the—$|u_{j\mathbf{k}}\rangle$ can also be unitarily mixed for different j's at fixed $\mathbf{k}$, i.e. for an insulator with n occupied bands:

$$|u_{j\mathbf{k}}\rangle \quad\rightarrow\quad |\tilde{u}_{j\mathbf{k}}\rangle = \sum_{j'=1}^{n} U_{jj'}(\mathbf{k})\,|u_{j'\mathbf{k}}\rangle. \tag{3.35}$$

All physical properties must be invariant under such generalized gauge transformation; this is particularly relevant in case of band crossing, where the band index j becomes ambiguous. The mixed orbitals are no longer Hamiltonian eigenstates; any gauge where instead the $|u_{j\mathbf{k}}\rangle$ are eigenstates will be called "Hamiltonian gauge".

3.9.2 Connection, curvature, and metric

While the $|\psi_{j\mathbf{k}}\rangle$ at different $\mathbf{k}$'s are mutually orthogonal, the $|u_{j\mathbf{k}}\rangle$ live instead in the same Hilbert space (they are all cell-periodical): we consider therefore the geometrical properties of the $|u_{j\mathbf{k}}\rangle$ states. We assume that, when taking the limit from a discrete Born-von-Kàrmàn $\mathbf{k}$-set to a continuous variable, the $|u_{j\mathbf{k}}\rangle$ states become differentiable functions of $\mathbf{k}$. The physical meaning of all the mathematical quantities introduced next will be discussed in the following Chapters, and not anticipated in the present one.

The Berry connection, in an insulator with n occupied states, is defined as

$$\boldsymbol{\mathcal{A}}(\mathbf{k}) = i \sum_{j=1}^{n} \langle u_{j\mathbf{k}} | \nabla_{\mathbf{k}} u_{j\mathbf{k}} \rangle; \tag{3.36}$$

the interesting closed paths C on the torus are lines across the reciprocal cell, from one face to the opposite one, whose end points are indeed equivalent. When addressing the Berry phase, there is a minor difference with respect to what has been defined so far, because of Eq. (3.34). This modified geometric phase is sometimes called a "Zak phase" [3, 55]; here we will call it simply Berry phase.

The Berry geometric phase is, according to the previous section,

$$\gamma = \int_{C} \boldsymbol{\mathcal{A}}(\mathbf{k}) \cdot d\mathbf{k}, \tag{3.37}$$

and depends on the choice of the origin in the crystal cell. For centrosymmetric crystals, if the origin is at a center of inversion symmetry, the only allowed values are $\gamma = 0$ and $\gamma = \pi$ (modulo 2π). The Berry curvature is easily found as

$$\Omega_{\alpha\beta}(\mathbf{k}) = i \sum_{j=1}^{n} (\langle \partial_\alpha u_{j\mathbf{k}} | \partial_\beta u_{j\mathbf{k}} \rangle - \langle \partial_\beta u_{j\mathbf{k}} | \partial_\alpha u_{j\mathbf{k}} \rangle), \tag{3.38}$$

or, in vector notation

$$\boldsymbol{\Omega}(\mathbf{k}) = i \sum_{j=1}^{n} \langle \nabla u_{j\mathbf{k}} | \times | \nabla u_{j\mathbf{k}} \rangle. \tag{3.39}$$

In order to get the metric we need to define the gauge-invariant distance between two sets of n occupied orbitals at $\mathbf{k}$ and $\mathbf{k} + \boldsymbol{\kappa}$. This is done by by means of the overlap matrix (also called connection matrix)

$$S_{jj'}(\mathbf{k}, \mathbf{k} + \boldsymbol{\kappa}) = \langle u_{j\mathbf{k}} | u_{j'\mathbf{k}+\boldsymbol{\kappa}} \rangle, \tag{3.40}$$

which generalizes Eq. (3.6) as

$$D_{\mathbf{k},\mathbf{k}+\boldsymbol{\kappa}}^2 = -\ln \det S^\dagger(\mathbf{k}, \mathbf{k} + \boldsymbol{\kappa}) S(\mathbf{k}, \mathbf{k} + \boldsymbol{\kappa}). \tag{3.41}$$

The matrix $S^\dagger S$ is Hermitian and its determinant is positive and gauge-invariant (in the above sense); the infinitesimal distance and the Bloch metric are

$$D^2_{\mathbf{k},\mathbf{k}+d\mathbf{k}} = \sum_{\alpha,\beta=1}^{d} g_{\alpha\beta}(\mathbf{k})\, dk_\alpha dk_\beta \qquad (3.42)$$

$$g_{\alpha\beta}(\mathbf{k}) = \mathrm{Re}\ \sum_{j=1}^{n} \langle \partial_\alpha u_{j\mathbf{k}} | \partial_\beta u_{j\mathbf{k}} \rangle - \sum_{jj'=1}^{n} \langle \partial_\alpha u_{j'\mathbf{k}} | u_{j\mathbf{k}} \rangle \langle u_{j\mathbf{k}} | \partial_\beta u_{j'\mathbf{k}} \rangle. \qquad (3.43)$$

The Bloch metric made its first appearance in electronic structure in 1997, in the famous Marzari-Vanderbilt paper about the maximally localized Wannier functions [56, 19]. We notice that the second term in Eq. (3.43) is real, ergo the metric and the curvature can be cast as the real and imaginary (times -2) parts, respectively, of a gauge-invariant metric-curvature tensor

$$\mathcal{F}_{\alpha\beta}(\mathbf{k}) = \sum_{j=1}^{n} \langle \partial_\alpha u_{j\mathbf{k}} | \partial_\beta u_{j\mathbf{k}} \rangle - \sum_{jj'=1}^{n} \langle \partial_\alpha u_{j\mathbf{k}} | u_{j\mathbf{k}} \rangle \langle u_{j\mathbf{k}} | \partial_\beta u_{j\mathbf{k}} \rangle.. \qquad (3.44)$$

3.9.3 Bloch projector

We define the ground-state Bloch projector, for an insulator with n occupied states, as

$$\mathcal{P}_{\mathbf{k}} = \sum_{j=1}^{n} |u_{j\mathbf{k}}\rangle\langle u_{j\mathbf{k}}|. \qquad (3.45)$$

This projector is gauge invariant and furthermore it determines the $|u_{j\mathbf{k}}\rangle$ apart from an irrelevant gauge transformation. Therefore all gauge invariant properties, like e.g. the metric-curvature tensor, can in principle be expressed in terms of $\mathcal{P}_{\mathbf{k}}$.

We consider the derivative

$$\partial_{\mathbf{k}}\mathcal{P}_{\mathbf{k}} = \sum_{j=1}^{n_b} |u_{j\mathbf{k}}\rangle\langle \partial_{\mathbf{k}} u_{j\mathbf{k}}| + \sum_{j=1}^{n_b} |\partial_{\mathbf{k}} u_{j\mathbf{k}}\rangle\langle u_{j\mathbf{k}}|. \qquad (3.46)$$

While $\partial_{\mathbf{k}}\mathcal{P}_{\mathbf{k}}$ is Hermitian and gauge-invariant, each of the two terms in the r.h.s. is neither Hermitian nor gauge-invariant. In fact, the trace of each of them (times $\pm i$) is nothing else than the Berry connection $\mathcal{A}(\mathbf{k})$. A somewhat lengthy calculation proves that the metric and the curvature can be expressed in trace form as

$$g_{\alpha\beta}(\mathbf{k}) = \mathrm{Re}\ \mathrm{Tr}\ \{\mathcal{P}_{\mathbf{k}}\, \partial_\alpha\mathcal{P}_{\mathbf{k}}\partial_\beta\mathcal{P}_{\mathbf{k}}\}. \qquad (3.47)$$

$$\Omega_{\alpha\beta}(\mathbf{k}) = i\,\mathrm{Tr}\ \{\mathcal{P}_{\mathbf{k}}\, [\partial_\alpha\mathcal{P}_{\mathbf{k}}, \partial_\beta\mathcal{P}_{\mathbf{k}}]\}. \qquad (3.48)$$

Both the metric and the curvature have the dimension of a squared length, ergo their BZ integral is dimensionless in $2d$. Since the BZ is a torus (a closed surface, a.k.a. compact orientable $2d$ manifold in mathematical jargon), the Gauss-Bonnet-Chern theorem guarantees that

$$\frac{1}{2\pi} \int_{BZ} d\mathbf{k} \, \Omega_{xy}(\mathbf{k}) = C_1, \qquad d = 2, \tag{3.49}$$

where $C_1 \in \mathbb{Z}$ is the Chern number in the context of electronic structure. It can be nonzero only in absence of T-symmetry.

The Chern number made its first appearance in electronic structure in 1982, in the famous TKNN (Thouless, Kohmoto, Nightingale, and den Nijs) paper about the quantum Hall effect [57] (Sec. 4.2.4); shortly afterwards a Chern number in the (k, t) variables was used by Thouless to demonstrate the quantization of particle transport [58] (Sect. 4.5). These (and other) discoveries were awarded with the 2016 Nobel prize [59].

3.9.4 Discretization in computer implementations

We have already said that there are no derivatives in computational physics. The Bloch vectors are obtained by matrix diagonalization on a discrete mesh of $\mathbf{k}$-points and the gauge is therefore erratic. The discretization of a Chern number exploits Stokes' theorem and is discussed below in Secs. 4.1.4 and 8.3.4. The discretization of the metric has been performed in various ways; my favorite option would be to exploit Eq. (3.41), where $\boldsymbol{\kappa}$ is the difference between any pair of mesh points. $D^2_{\mathbf{k},\mathbf{k}+\boldsymbol{\kappa}}$ is a scalar function, manifestly gauge invariant, defined on the $\boldsymbol{\kappa}$ mesh. It is immediate to evaluate, for any given $\mathbf{k}$, its discretized second order term in $\boldsymbol{\kappa}$.

The Berry phase requires an ad-hoc procedure, which generalizes the single-state discretization of Eqs. (3.7) and (3.11). The Berry connection can be clearly written as

$$\boldsymbol{\mathcal{A}}(\mathbf{k}) = i \left. \frac{\partial}{\partial \boldsymbol{\kappa}} \ln \det S(\mathbf{k}, \mathbf{k} + \boldsymbol{\kappa}) \right|_{\boldsymbol{\kappa}=0} = - \left. \frac{\partial}{\partial \boldsymbol{\kappa}} \text{Im} \ln \det S(\mathbf{k}, \mathbf{k} + \boldsymbol{\kappa}) \right|_{\boldsymbol{\kappa}=0}. \tag{3.50}$$

Since the Berry phase in any dimension d is a line integral, we provide the formula in $1d$ only; the d-dimensional case needs more complex notations, but is conceptually trivial. The BZ integral can be equivalently performed on the reciprocal cell; if a is the $1d$ lattice constant we have

$$\gamma = \int_0^{\frac{2\pi}{a}} d\varphi, \qquad d\varphi = \mathcal{A}(k) dk. \tag{3.51}$$

If we discretize with M equally spaced points, exploiting Eq. (3.50) we get:

$$k_s = \frac{2\pi}{Ma} s, \qquad s = 0, 1, \ldots M - 1, \tag{3.52}$$

$$\gamma = \int_0^{\frac{2\pi}{a}} d\varphi \simeq \sum_{s=1}^{M} \Delta\varphi_{s,s+1}, \qquad \Delta\varphi_{s,s+1} = -\text{Im} \ln \det S(k_s, k_{s+1}). \qquad (3.53)$$

The S matrix elements have to be evaluated enforcing the periodic gauge and diagonalizing the Hamiltonian on M points (not $M+1$), ergo

$$S_{jj'}(k_s, k_{s+1}) = \langle u_{jk_s} | u_{k_{j's+1}} \rangle, \quad S_{jj'}(k_{M-1}, k_M) = \langle u_{jk_{M-1}} | e^{-i\frac{2\pi}{a}x} | u_{k_0} \rangle; \qquad (3.54)$$

the discretized Berry phase then becomes

$$\gamma \simeq -\text{Im} \sum_{s=1}^{M} \ln \det S(k_s, k_{s+1}) = -\text{Im} \ln \prod_{s=1}^{M} \det S(k_s, k_{s+1}). \qquad (3.55)$$

The last expression on the r.h.s. is the one universally implemented in the computer codes; it is clearly unaffected by any erratic phase factor and/or band ordering resulting. The eigenvectors provided by the diagonalization routine can be used as they are.

For the sake of completeness, we also mention that at the time of writing (2018) the discretization of the $3d$ BZ integral of the Chern-Simons 3-form, Eq. (3.71) below, still poses a formidable challenge [60, 61].

3.10 NonAbelian geometry

3.10.1 Generalities

What we have called "Bloch geometry" in insulators is the geometry of many states considered altogether, which determine in a gauge-invariant way the ground-state properties. These could be e.g. the Berry phase along a line between two points (separated by $\mathbf{G}$), or the BZ integral of the metric-curvature tensor.

We have somewhat hidden so far the nonAbelian features, summing explicitly over the indices. We are going to recast some of these results (and to present some additional ones) using the powerful notation of modern differential geometry in a nonAbelian framework: exterior differentiation and exterior algebra (also called Grassmann algebra).

So far, we have written the connection and the curvature as vector fields; in order to proceed, it proves better to switch to the language of differential forms. To this aim, we define the nonAbelian connection as the 1-form

$$A = i\langle u_{j\mathbf{k}} | \partial_{\mathbf{k}} u_{j'\mathbf{k}} \rangle \cdot d\mathbf{k} \equiv i\langle u_{j\mathbf{k}} | \partial_\alpha u_{j'\mathbf{k}} \rangle dk^\alpha, \qquad (3.56)$$

where we sum over repeated indices. Notice that the differential is on the right, i.e. the symbol A includes the $d\mathbf{k}$, a 2-form includes two differentials, and so on;

the coefficients of the forms are matrices over the band indices. In mathematics the form A is by definition a nonAbelian Chern-Simons 1-form; in particle physics it is also known as a nonAbelian gauge potential. In $1d$ the Berry phase is written as

$$\gamma = \int_{\text{BZ}} \text{Tr}\{A\}, \qquad d = 1. \tag{3.57}$$

In order to go to higher dimensions we need two tools: the wedge product (a.k.a. exterior product), and the exterior differentiation.

3.10.2 Exterior product and differentiation

The wedge product generalizes the vector product, and is associative. If both ω and η are 1-forms, then $\omega \wedge \eta$ is a 2-form whose coefficients are antisymmetric. In our case the coefficients of both ω and η are matrices, and the coefficients of $\omega \wedge \eta$ are their commutator; clearly $\text{Tr}\{\omega \wedge \eta\} = 0$. If one of the two forms is a function (i.e. a 0-form), then the wedge product is equivalent to scalar multiplication. In general, if ω is a p-form and η is a q-form, their wedge product is a $(p+q)$-form and

$$\omega \wedge \eta = (-1)^{pq}\eta \wedge \omega. \tag{3.58}$$

The second tool needed is the exterior differentiation, indicated simply with d. When ω is a 0-form (i.e. a scalar), then $d\omega$ is just the ordinary differential of ω; for any form $d(d\omega) = 0$, i.e. $d^2 = 0$. If ω is a 1-form, then

$$\omega = f_\alpha dk^\alpha, \qquad d\omega = \frac{\partial f_\alpha}{\partial k_\beta} dk^\alpha \wedge dk^\beta, \tag{3.59}$$

i.e. instead of antisymmetrizing the derivatives (as one does in vector notation) here we antisymmetrize the differentials. In general, if ω is a p-form then $d\omega$ is a $(p+1)$-form which generalizes the curl:

$$\omega = f_J dk^J, \qquad d\omega = \frac{\partial f_J}{\partial \alpha_j} dk^{\alpha_j} \wedge dk^J, \tag{3.60}$$

where $J = (\alpha_1, \alpha_2, \ldots \alpha_p)$ is a multi-index and f_J are functions (the coefficients of the p-form).

These compact notations were introduced about 1920-30, mostly by Élie Cartan, who proved a generalized Stokes'theorem: if Σ is a $p+1$-dimensional domain and $\partial\Sigma$ its p-dimensional boundary, then

$$\int_{\partial\Sigma} \omega = \int_\Sigma d\omega. \tag{3.61}$$

Special cases are: Green's theorem (ω is a 1-form in 2-space), divergence theorem (ω is a 2-form in 3-space), and Stokes' theorem (ω is a 1-form in 3-space).

3.10.3 NonAbelian connection and curvature

Because of norm conservation, the coefficients of our A are Hermitian. A caveat is in order here: in the mathematical and particle-physics literature (and in Wikipedia as well) the 1-form A does not include the i factor, and is therefore antiHermitian. Here I present the formulæ as they appear in the electronic-structure literature (the choice of $\pm i$ is non uniform, though).

Equipped with the above rules, we come back to our nonAbelian connection A. The unitary transformation U acts on A as

$$A \;\rightarrow\; A^U = U^{-1}AU + i\,U^{-1}dU = U^\dagger AU + i\,U^\dagger dU; \qquad (3.62)$$

the scalar (0-form) U is the nonAbelian gauge transformation (an element of the gauge group). The connection is not gauge-covariant, and therefore does not correspond to a measurable quantity. The 2-form

$$F = dA - iA \wedge A \qquad (3.63)$$

is the nonAbelian curvature (gauge-field strength in particle physics), and is Hermitian. F is gauge-covariant and corresponds therefore to a measurable quantity :

$$
\begin{aligned}
F \;\rightarrow\; F^U &= dA^U - i\,A^U \wedge A^U \\
&= d(U^\dagger AU + iU^\dagger dU) - i\,(U^\dagger AU + iU^\dagger dU) \wedge (U^\dagger AU + iU^\dagger dU) \\
&= U^{-1}(dA - iA \wedge A)U = U^\dagger FU.
\end{aligned}
\qquad (3.64)
$$

The trace of F is gauge-invariant, and in fact it is one half of the Berry curvature, Eq. (3.55), in differential form:

$$\mathrm{Tr}\{F\} = \mathrm{Tr}\{dA - iA \wedge A\} = \Omega_{\alpha\beta}\,dk^\alpha dk^\beta = \frac{1}{2}\Omega_{\alpha\beta}\,dk^\alpha \wedge dk^\beta, \qquad (3.65)$$

where $A \wedge A$ does not contribute to the trace. The first Chern number obtains integrating the 2-form on the $2d$ BZ (a compact orientable manifold, in mathematical jargon):

$$C_1 = \frac{1}{2\pi}\int_{\mathrm{BZ}} \mathrm{Tr}\;\{F\} = \frac{1}{2\pi}\int_{\mathrm{BZ}} \mathrm{Tr}\;\{dA - iA \wedge A\}, \qquad d = 2. \qquad (3.66)$$

The metric $D^2 = g_{\alpha\beta}dk^\alpha dk^\beta$ is a symmetric tensor, therefore is incompatible with the exterior algebra; furthermore it cannot be expressed in terms of A solely. However we may define the metric-curvature nonAbelian 2-form as

$$\tilde{F} = \eta - A \otimes A \qquad (3.67)$$

where the matrix coefficients of η are $\langle \partial_\alpha u_j | \partial_\beta u_{j'} \rangle$. Then the Hermitian part of $\tilde{F}$ is the nonAbelian metric, and its antiHermitian part, times $2i$, coincides with the nonAbelian curvature. F. The trace of $\tilde{F}$ over the band indices yields the ordinary metric-curvature tensor, Eq. (3.44).

3.10.4 Chern-Simons 3-form

The gauge-invariant 2-form $\mathrm{Tr}\{F\}$ is called the first Chern form; its BZ integral yields the first Chern number C_1, Eq. (3.66). The key reason (not explained here) why C_1 is an integer is that $d\,\mathrm{Tr}\{F\} = 0$; furthermore any 2-form which obeys such condition can be expressed locally as the exterior derivative of a 1-form, function of A, the Chern-Simons (CS in the following) 1-form. In this case the function is particularly simple: $\mathrm{Tr}\{F\} = d\,\mathrm{Tr}\{A\}$. In plain language, the curvature is the curl of the connection. Notice also that, when $C_1 \neq 0$, the 1-form cannot be defined globally: there is an "obstruction". The 2-form, instead, is defined globally: see Sect. 3.5 and the toy-model example in Sect. 4.1.

What has been discussed so far is the simplest example of the relationship between CS forms in odd dimension and Chern forms in even dimension. In fact, a similar relationship exists between CS 3-forms and Chern 4-forms. The next interesting form is therefore the gauge-covariant 4-form $F \wedge F$, and its gauge-invariant trace: the latter is the second Chern form. Even $d\,\mathrm{Tr}\{F \wedge F\} = 0$, and therefore—because of the first theorem invoked above—the integral of $\mathrm{Tr}\{F \wedge F\}$ over a 4-dimensional "Brillouin zone" is an integer: the second Chern number $C_2 \in \mathbb{Z}$, whose expression is

$$C_2 = -\frac{1}{8\pi^2} \int_{\mathrm{BZ}} \mathrm{Tr}\,\{F \wedge F\}, \qquad d = 4. \qquad (3.68)$$

According to the second theorem invoked above (called the Poincaré lemma) the condition $d\mathrm{Tr}\,\{F \wedge F\} = 0$ also implies that $\mathrm{Tr}\,\{F \wedge F\}$ can be expressed locally as the exterior derivative of a 3-form: this is in fact the CS 3-form, by now somewhat famous in electronic structure theory. The relationship between the Chern 4-form and the CS 3-form is

$$\mathrm{Tr}\,\{F \wedge F\} = d\,\mathrm{Tr}\,\left\{A \wedge dA - \frac{2i}{3} A \wedge A \wedge A\right\}, \qquad (3.69)$$

and can be verified by a somewhat lengthy calculation. In order to emphasize the analogies with the first Chern number and with the Berry connection (i.e. the CS 1-form), we rewrite the previous equations as

$$C_2 = \frac{1}{2\pi} \left(-\frac{1}{4\pi} \int_{\mathrm{BZ}} \mathrm{Tr}\,\{F \wedge F\} \right), \qquad (3.70)$$

$$\begin{aligned} \omega_{\mathrm{CS}} &= -\frac{1}{4\pi} \mathrm{Tr}\left\{A \wedge dA - \frac{2i}{3} A \wedge A \wedge A\right\} \\ &= -\frac{1}{4\pi} \varepsilon_{\alpha\beta\gamma} \left(A_\alpha \partial_\beta A_\gamma - \frac{2i}{3} A_\alpha A_\beta A_\gamma \right) dk^\alpha dk^\beta dk^\gamma \qquad (3.71) \end{aligned}$$

The 1-form $\mathrm{Tr}\{A\}$ and the 3-form ω_{CS} share several common features: in particular, they are gauge-dependent but their integrals over the BZ (1- or 3-dimensional, respectively) are dimensionless and gauge-invariant modulo 2π. Both are therefore angles:

$$\gamma = \int_{\mathrm{BZ}} \mathrm{Tr}\{A\}, \quad d = 1; \qquad \theta = \int_{\mathrm{BZ}} \omega_{\mathrm{CS}}, \quad d = 3. \tag{3.72}$$

The experimental relevance of θ in electronic structure has been pointed out by Qi, Hughes, and Zhang in 2008 [62, 63, 15], and will be discussed below in Chapter 6.

3.10.5 $\mathbb{Z}_2$ topological invariants

We have already observed that in presence of inversion symmetry the allowed values for the $1d$ Berry phase are either $\gamma = 0 \bmod 2\pi$ or $\gamma = \pi \bmod 2\pi$. Analogously, in presence of either inversion symmetry or T-symmetry the allowed values for the CS angle are either $\theta_{\mathrm{CS}} = 0 \bmod 2\pi$ or $\theta_{\mathrm{CS}} = \pi \bmod 2\pi$. Each of the two value-invariants clearly has a one-to-one mapping to the additive group of the integers modulo two: the $\mathbb{Z}_2$ group. Furthermore, this $\mathbb{Z}_2$ classification is topological: we may therefore classify the materials as either $\mathbb{Z}_2$-even or $\mathbb{Z}_2$-odd. In modern jargon, the $\mathbb{Z}_2$ index is "protected" by some kind of symmetry. Notice that, instead, the $\mathbb{Z}$ invariants (i.e. the Chern numbers) do not need any symmetry; quite on the contrary, C_1 can be nontrivial only when T-symmetry is *absent.*

The macroscopic polarization of a centrosymmetric polymer (a $1d$ "crystal") is a very simple case of $\mathbb{Z}_2$ topological classification occurring in nature; it is protected by inversion symmetry and T-symmetry together [64]. A different $\mathbb{Z}_2$ invariant occurs in molecular physics, where the role of the **k** parameter is assumed by a nuclear coordinate (Sect. 4.2.3). The effect is due to a conical intersection [65], which Berry calls "diabolical point" [25]; in this case only T-symmetry is required.

The CS angle θ_{CS} can be used in principle to classify 3-dimensional insulators according to their $\mathbb{Z}_2$ topological index in presence of the above (and even other) symmetries: this is discussed below in Sec. 6.1.1. Since spin-orbit interaction is essential in this context, the connection A is defined in terms of spinorbitals (not spinless orbitals).

Chapter 4

Manifestations of the Berry phase

4.1 A toy-model Hamiltonian

Our toy model here is a two-level spinless Hamiltonian, of the form

$$
\begin{aligned}
H(\boldsymbol{\xi}) &= \boldsymbol{\xi} \cdot \vec{\sigma} \\
&= \xi \left(\sin \vartheta \cos \varphi \, \sigma_x + \sin \vartheta \sin \varphi \, \sigma_y + \cos \vartheta \, \sigma_z \right),
\end{aligned}
\tag{4.1}
$$

where σ_α are the three Pauli matrices. The spectrum is non degenerate for $\boldsymbol{\xi} \neq 0$, and the lowest eigenvalue is $-\xi$. Upon symmetry arguments, we can already guess the curvature to be isotropic.

4.1.1 Connection and curvature

The lowest eigenvector is

$$
|\psi(\vartheta, \varphi)\rangle = \begin{pmatrix} \sin \frac{\vartheta}{2} \mathrm{e}^{-i\varphi} \\ -\cos \frac{\vartheta}{2} \end{pmatrix}.
$$

This corresponds to a specific gauge choice; the eigenvector can be multiplied by an overall (ϑ, φ)-dependent phase factor. The Berry connection and curvature are

$$
\begin{aligned}
\mathcal{A}_\vartheta &= i\langle \psi | \partial_\vartheta \psi \rangle = 0 \\
\mathcal{A}_\varphi &= i\langle \psi | \partial_\varphi \psi \rangle = \sin^2 \frac{\vartheta}{2} \\
\Omega &= \partial_\vartheta \mathcal{A}_\varphi - \partial_\varphi \mathcal{A}_\vartheta = \frac{1}{2} \sin \vartheta.
\end{aligned}
\tag{4.2}
$$

The curvature is gauge-invariant, while the connection is gauge-dependent. Within our gauge choice the connection displays a vortex at the south pole ($\vartheta = \pi$); other gauges yield the singularity at a different point, but a singularity is unavoidable. It

is impossible to find a gauge which is smooth on the whole closed surface, and a nonsingular connection; the singularity—often called "obstruction"—can be moved but not removed. The algebra is the same as for Dirac's theory of the magnetic monopole [51, 53]: the degeneracy at the origin is the monopole, and the singularity is the "Dirac string".

4.1.2 Chern number

The domain of the parameters (ϑ, φ) is a rectangle, which indeed has the topology of a torus: a closed surface. Integrating the Berry curvature therein we get

$$\frac{1}{2\pi} \int_{S^2} d\vartheta d\varphi \, \frac{1}{2} \sin \vartheta = 1, \tag{4.3}$$

i.e. the Chern number of the lowest eigenstate in this problem. This integer measures the strength of the singularity (magnetic monopole), which resides in a site *inaccessible* to the quantum system. The highest eigenstate has $C_1 = -1$, since the total Chern number is zero.

This simple example illustrates well the meaning of a topological invariant of the quantum mechanical ground state. The Chern number is in fact very robust under continuous deformations of the surface and of the Hamiltonian: its value is always one insofar as one (and only one) degeneracy point is included in the closed surface.

4.1.3 Berry phase

Suppose now we evaluate the Berry phase over any closed curve C on the sphere (Fig. 4.1)

$$\gamma = \oint_C \boldsymbol{\mathcal{A}}(\boldsymbol{\xi}) \cdot d\boldsymbol{\xi}. \tag{4.4}$$

Owing to Stokes theorem, the Berry phase for this toy model problem clearly equals the solid angle spanned by the curve, divided by 2. The inherent 4π arbitrariness in the solid angle leads to the well known 2π arbitrariness in the Berry phase: for instance at the equator $\gamma = \pi$ modulo 2π. If we cut the sphere in two hemispheres, as in Sec. 4.1.2 (see Fig. 3.4), the difference of the boundary Berry phases equals 0 modulo 2π, i.e. 2π times an integer, as it must be. But in order to tell *which* integer (the actual Chern number) the two connections are useless: one has to integrate the curvature, as in Eq. (4.3). Despite this feature, in numerical work Chern numbers are typically computed via Berry phases, as described in the next Section.
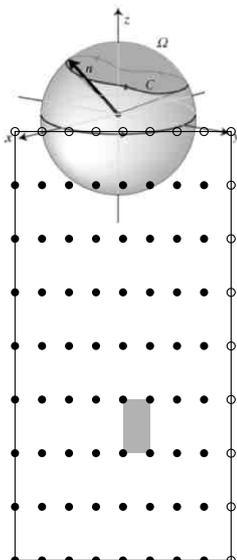
Figure 4.1: A closed curve C on the surface of the sphere, and the solid angle spanned by it.
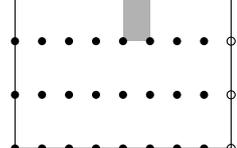
Figure 4.2: Discretization of the domain $[0, \pi] \times [0, 2\pi]$; the Hamiltonian is *not* diagonalized at the empty circles, only at the black ones, thus enforcing toroidal topology.

4.1.4 Computing a Chern number

This exactly soluble example also provides the occasion for illustrating the standard computational approach to Chern numbers. Suppose we discretize the (ϑ, φ) domain with a rectangular mesh, and that we diagonalize the Hamiltonian at the points of the mesh. The gauge at any point is chosen by the diagonalization routine and is thus erratic; we only enforce the toroidal topology by requiring that the phases at the opposite edges of the rectangle are the same: Fig. 4.2.

Then for each small rectangle we compute the discrete Berry phase as in Eq. (3.8), i.e.

$$
\begin{aligned}
\gamma \;=\; & -\,\text{Im} \ln \langle \psi(\vartheta, \varphi)|\psi(\vartheta + \Delta\vartheta, \varphi)\rangle \langle \psi(\vartheta + \Delta\vartheta, \varphi)|\psi(\vartheta + \Delta\vartheta, \varphi + \Delta\varphi)\rangle \\
& \times\;\; \langle \psi(\vartheta + \Delta\vartheta, \varphi + \Delta\varphi)|\psi(\vartheta, \varphi + \Delta\varphi)\rangle \langle \psi(\vartheta, \varphi + \Delta\varphi)|\psi(\vartheta, \varphi)\rangle. \quad (4.5)
\end{aligned}
$$

In this simple, analytically soluble, case we know the exact value: the Berry phase on the contour of each rectangle is $\gamma = \frac{1}{2}[\cos(\vartheta) - \cos(\vartheta + \Delta\vartheta)]\,\Delta\varphi$ modulo 2π. The four-point Berry phase, Eq. (4.5), provides an approximated value for this. The Chern number is the integral over the domain, and is therefore equal to the sum of all the phases computed as in Eq. (4.5) and covering the whole domain.

How do we then get rid of the modulo 2π indeterminacy in Eq. (4.5)? The size of $\Delta\vartheta\Delta\varphi$ is small, and each contribution γ to the sum is also small (proportional to $\Delta\vartheta\Delta\varphi$), although Eq. (4.5) is in principle arbitrary modulo 2π. It should be now pretty clear that the right solution is in choosing the "Im log" branch with values in $[-\pi, \pi]$.

As said above, each four-point Berry phase provides only an approximation to the corresponding analytical integration; it is thus natural to guess that a discrete

computation of the Chern number yields approximately an integer, coinciding with the exact value only in the limit of a dense mesh. Instead, this is not the case; even a coarse mesh provides exactly an integer. Indeed, the role of the mesh is to select *which* integer: for our toy model either $C_1 = 0$ (if the singularity is outside the surface), or $C_1 \pm 1$ (for the lower and higher state, if the singularity is inside the surface). This selection is a virtue of the branch prescription discussed above for the multivalued function Im log.

In order to show that the discretized Chern number is exactly integer we notice that when we sum over the small rectangles the contributions from their inner sides cancel in pairs. Therefore the total phase must be equal (mod 2π) to the discrete phase computed on the outer contour of the whole domain. We already observed that the states at the opposite edges must be the same, with the same phases (see Fig. 4.2), hence the sum on the edges yields an overlap product real and positive. This proves that the total (discrete) phase is *exactly* zero mod 2π.

Last but not least: where is the obstruction? In the continuous formulation, any gauge choice yields a singularity at a single (ϑ, φ) point. In the discrete formulation, there is no way to locate the singularity: in some sense, the singularity is everywhere since the gauge is in principle erratic at all mesh points. While in mathematicians' topology the invariant is robust for *continuous* deformations of the Hamiltonian and/or of the surface, our computational topology is robust even for *discontinuous* deformations and, furthermore, even for a very coarse mesh. Four points on the sphere, which define a tetrahedron, would be enough. I regard all this as a manifestation of the "unreasonable effectiveness' of topology.
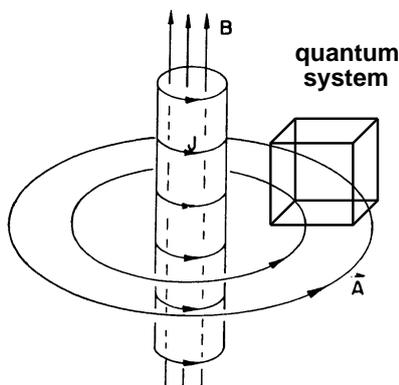
Figure 4.3: A particle in a box, transported round a solenoid

4.2 Early discoveries reinterpreted

4.2.1 Aharonov-Bohm effect

Here we reformulate the Aharonov-Bohm effect as a special case of a Berry phase. Suppose we have an electron in a box (infinite potential well) centered at the origin. We take the ground wavefunction as real, and we write it as $\chi(\mathbf{r})$. The time-independent Schrödinger equation is:

$$\left[\frac{p^2}{2m} + V(\mathbf{r})\right]\chi(\mathbf{r}) = E\chi(\mathbf{r}). \tag{4.6}$$

Displacing the center of the box at position $\mathbf{R}$ changes the Hamiltonian to

$$H(\mathbf{R}) = \frac{p^2}{2m} + V(\mathbf{r} - \mathbf{R}) \ : \tag{4.7}$$

we will identify the $\boldsymbol{\xi}$ parameter with the box position $\mathbf{R}$. Because of translational invariance, the $\mathbf{R}$-dependence of the state vectors is

$$\langle\mathbf{r}|\psi(\mathbf{R})\rangle = \chi(\mathbf{r} - \mathbf{R}), \tag{4.8}$$

while the eigenvalue is $\mathbf{R}$-independent.

Suppose now that a magnetic field is switched on somewhere in space. Then the Hamiltonian becomes

$$H(\mathbf{R}) = \frac{1}{2m}\left[\mathbf{p} + \frac{e}{c}\mathbf{A}(\mathbf{r})\right]^2 + V(\mathbf{r} - \mathbf{R}) \ , \tag{4.9}$$

where $\mathbf{A}$ is the vector potential and e is the electron charge. It can be easily verified that a solution of the Schrödinger equation can be formally written in the form:

$$\langle\mathbf{r}|\psi(\mathbf{R})\rangle = \exp\left(-\frac{ie}{\hbar c}\int_{\mathbf{R}}^{\mathbf{r}}\mathbf{A}(\mathbf{r}') \cdot d\mathbf{r}'\right)\chi(\mathbf{r} - \mathbf{R}). \tag{4.10}$$

But such a solution in general is *not* a single-valued function of $\mathbf{r}$, since the phase factor depends on the path. Therefore we restrict ourselves to a less general case, where the magnetic field is generated by a solenoid: the $\mathbf{B}$ field is nonzero only within a given cylinder, and we don't allow our box to overlap this cylinder by suitably restricting the domain of $\mathbf{R}$. This situation is sketched in Fig. 4.3. With such a choice the wavefunction, Eq. (4.10), is a single valued function of $\mathbf{r}$ for any fixed $\mathbf{R}$, and is therefore an honest ground wavefunction. As for the dependence on $\mathbf{R}$, Eq. (4.10) only guarantees *local* single-valuedness, since the domain is not simply connected: when the system is transported on a closed path winding once round the

solenoid, the electron wavefunction picks up a Berry's phase. This phase difference can be actually detected in interference experiments.

The Berry connection of the problem is

$$\mathcal{A}(\mathbf{R}) = i\langle\psi(\mathbf{R})|\nabla_\mathbf{R}\psi(\mathbf{R})\rangle = i\langle\chi(\mathbf{R})|\nabla_\mathbf{R}\chi(\mathbf{R})\rangle - \frac{e}{\hbar c}\mathbf{A}(\mathbf{R}), \qquad (4.11)$$

where the first term vanishes: $\chi(\mathbf{r})$ is real. Therefore in the present case the Berry connection is proportional to the ordinary vector potential. A gauge transformation in the quantum-mechanical sense also coincides with an electromagnetic gauge transformation, which changes $\mathbf{A}$ while leaving $\mathbf{B}$ invariant. In fact, in this example $\mathbf{B}$ is essentially the Berry curvature. The Berry phase is

$$\gamma = -\frac{e}{\hbar c}\oint_C \mathbf{A}(\mathbf{R})\cdot d\mathbf{R} = -\frac{2\pi}{\phi_0}\oint_C \mathbf{A}(\mathbf{R})\cdot d\mathbf{R}, \qquad (4.12)$$

where ϕ_0 is the flux quantum. Therefore γ measures the flux of the magnetic field across the *interior* of the solenoid, a space region *not accessed* by the quantum system: above, we have called it "inaccessible flux". Only the fractional part of the flux has physical meaning.

4.2.2 Molecular Aharonov-Bohm effect

Here we identify the "slow coordinate" $\boldsymbol{\xi}$ with a d-dimensional nuclear coordinate, and the state vector $|\Psi(\boldsymbol{\xi})\rangle$ with the electronic ground-state wavefunction in the Born-Oppenheimer approximation.

We start from the complete Hamiltonian $\mathcal{H}$ of an isolated molecular system, and we explicitly separate the nuclear kinetic energy:

$$\mathcal{H}(\boldsymbol{\xi},[\mathbf{x}]) = \frac{1}{2}\sum_{\alpha,\beta=1}^{d}\mathfrak{M}_{\alpha\beta}^{-1}\,p_\alpha p_\beta + H(\boldsymbol{\xi},[\mathbf{x}]), \qquad (4.13)$$

where $[\mathbf{x}]$ indicates the electronic degrees of freedom collectively, $p_\alpha = -i\hbar\,\partial/\partial\xi_\alpha$ is the canonical momentum conjugated to ξ_α, and the inverse mass matrix $\mathfrak{M}^{-1}$ in general may be a function of $\boldsymbol{\xi}$, but not of the momenta.

The Born-Oppenheimer approximation starts by writing the eigenfunctions of Eq. (4.13) in the Schrödinger representation as the product $\langle[\mathbf{x}]|\Psi(\boldsymbol{\xi})\rangle\,\Phi(\boldsymbol{\xi})$. Our aim is obtaining an effective Schrödinger equation for the nuclear wavefunction $\Phi(\boldsymbol{\xi})$, where the electronic degrees of freedom have been integrated out. We start considering the effect of the canonical nuclear momentum $\mathbf{p}$ on the product ansatz:

$$\mathbf{p}\,|\Psi(\boldsymbol{\xi})\rangle\,\Phi(\boldsymbol{\xi}) = -i\hbar\,|\Psi(\boldsymbol{\xi})\rangle\,\nabla_{\boldsymbol{\xi}}\Phi(\boldsymbol{\xi}) - i\hbar\,|\nabla_{\boldsymbol{\xi}}\Psi(\boldsymbol{\xi})\rangle\,\Phi(\boldsymbol{\xi}). \qquad (4.14)$$

We then multiply by the electronic eigenbra $\langle \Psi(\boldsymbol{\xi})|$ on the left, thus integrating over the electronic degrees of freedom. We get the effective nuclear momentum $\boldsymbol{\pi}$ acting on Φ as:

$$\boldsymbol{\pi}\, \Phi(\boldsymbol{\xi}) = [\, \mathbf{p} - i\hbar\langle\Psi(\boldsymbol{\xi})|\nabla_{\boldsymbol{\xi}}\Psi(\boldsymbol{\xi})\rangle\,]\,\Phi(\boldsymbol{\xi}) = [\,\mathbf{p} - \hbar\boldsymbol{\mathcal{A}}(\xi)\,]\,\Phi(\boldsymbol{\xi}), \qquad (4.15)$$

where we easily recognize the Berry connection. The momentum $\boldsymbol{\pi}$ is the kinematical (also called covariant, or mechanical) momentum, to be distinguished from $\mathbf{p} = -i\hbar\nabla_{\boldsymbol{\xi}}$, which is the canonical momentum.

Whenever the time scales of nuclear and electronic motions are well separated the coupling between different electronic states can be neglected, and the adiabatic approximations allows to treat the slow variable $\boldsymbol{\xi}$ in $H(\boldsymbol{\xi}, [\mathbf{x}])$ as a classical parameter. The electronic eigenvalue $E(\boldsymbol{\xi})$ of a given state (e.g. the ground state) plays therefore the role of a (scalar) potential for nuclear motion, whose effective Hamiltonian acting on $\Phi(\boldsymbol{\xi})$ is then:

$$\mathsf{H}_{\text{eff}} = \frac{1}{2}\sum_{\alpha,\beta=1}^{d} \mathfrak{M}_{\alpha\beta}^{-1}\,\pi_\alpha\pi_\beta + E(\boldsymbol{\xi}). \qquad (4.16)$$

In the molecular physics literature the extra term in Eq. (4.15) is seldom mentioned, and $\boldsymbol{\pi}$ is identified with $\mathbf{p}$. The reason is that for a T-invariant Hamiltonian, and in absence of spin-orbit interaction, the wave function can always be taken as real. This corresponds to the parallel transport gauge, and the Berry connection vanishes at all $\boldsymbol{\xi}$; the tradeoff is that—in some cases—the electronic wave function is *not* single valued along a closed path: see Fig. 2.3. The alternative approach, due to Mead and Truhlar [36, 65], is to choose a different gauge, where the electronic wave function is single valued and complex. The Berry phase is gauge invariant (modulo 2π); the values allowed by T-symmetry are 0 and π; the two cases are experimentally distinguishable, as previously shown in Sec. 2.2.

We stress that, whenever the ionic motion is purely classical and governed by Newton's equation, the vector-potential-like term in Eq. (4.15) is irrelevant: the corresponding curvature (magnetic-field-like) is in fact identically vanishing along the nuclear trajectory on the Born-Oppenheimer surface. We anticipate that the case where a genuine magnetic field is present—and the Hamiltonian is no longer T-invariant—is *qualitatively* different in this respect, see Sec. 4.3 below.

4.2.3 The $\mathbb{Z}_2$ invariant in molecular physics

The topological nature of the molecular Aharonov-Bohm effect has already been pointed out in Sec. 2.2, and briefly outlined in modern topological language in Sect. 3.10.5; here we give more details. The ground state is degenerate with the

first excited state at the conical intersection ("diabolical point" according to Berry [25]); therefore according to Eq. (3.30) one would expect the curvature to have a singularity. This does not happen here: thanks to T-symmetry the curvature is everywhere zero, including at the conical intersection.

We focus on the simplest possible case, dealt with in Sec. 2.2: a two-level system with (or without) a conical intersection. In absence of spin-orbit interaction, a gauge which respects T-symmetry requires the electronic wave function to be everywhere real; this gauge coincides with the parallel-transport gauge. In the nontrivial case ($\mathbb{Z}_2$odd, i.e. $\gamma = \pi$ mod 2π) this connection cannot be smooth and single-valued on a closed path which winds around the intersection; however, it is a simple exercise to devise a smooth gauge throughout the path: it is enough to adopt complex electronic wavefunctions.

At variance with the curvature, the connection cannot be smooth everywhere and has a singularity at the intersection, in any gauge: in modern jargon this is generally called an "obstruction". In this case, the singularity is traced back to the breakdown of Born-Oppenheimer approximation at the intersection.

The topological invariant is easily retrieved from the discrete approach: only three Hamiltonian diagonalizations on the path are needed. With reference to Fig. 2.3 we define

$$|\psi_1\rangle = \frac{1}{\sqrt{2}}(|B\rangle - |C\rangle), \ |\psi_2\rangle = \frac{1}{\sqrt{2}}(|B\rangle - |A\rangle), \ |\psi_3\rangle = \frac{1}{\sqrt{2}}(|C\rangle - |A\rangle). \quad (4.17)$$

Independently of the chosen gauge, whether if it respects T-symmetry or otherwise, we always have

$$\langle\psi_1|\psi_2\rangle\langle\psi_2|\psi_3\rangle\langle\psi_3|\psi_1\rangle = -\frac{1}{8}, \quad (4.18)$$

i.e. $\gamma = \pi$ mod 2π: the ground state is $\mathbb{Z}_2$ odd.

The $\mathbb{Z}_2$ topological invariant is extremely robust against smooth deformations of the Hamiltonian. For instance, here we have addressed the ultrasimple tight-binding model, but the ground wavefunction can be "continuously deformed" to the exact correlated wavefunction: topology-wise, the two wavefunctions are essentially the same object: they can be regarded as analogous to the proverbial doughnut and coffee cup, Figs. 1.1 and 1.3. The $\mathbb{Z}_2$ invariant is "protected" by T-symmetry: in fact, breaking this symmetry opens a gap at the intersection.

4.2.4 Integer quantum Hall effect (TKNN invariant)

The famous TKNN (Thouless, Kohmoto, Nightingale, and den Nijs) paper appeared in 1982 [57] and marks the very first occurrence of a topological invariant (the first Chern number) in electronic structure. David Thouless was awarded 1/2 of the 2016 Nobel prize; among his outstanding contributions to condensed matter physics, the
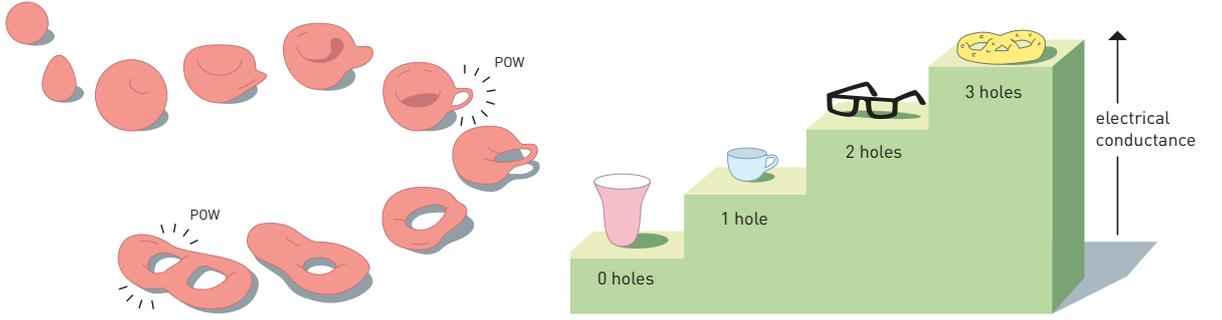
Fig 3. Topology. This branch of mathematics is interested in properties that change step-wise, like the number of holes in the above objects. Topology was the key to the Nobel Laureates' discoveries, and it explains why electrical conductivity inside thin layers changes in integer steps.

Figure 4.4: Cartoon posted in 2016 at `www.nobelprize.org` in the Section "Popular information".

TKNN paper plays a predominant role. The Nobel Prize foundations published in 2016 the cartoon reproduced here in Fig. 4.4.

The outline provided here is inspired by Kohmoto [66]; in the quantum Hall context, first Chern number is synonymous of TKNN invariant. We consider a two-dimensional independent-electron system in a lattice-periodical potential, and subject to a perpendicular **B** field. The Hamiltonian is *not* translationally invariant, but one can address the magnetic translation group. We choose a large enough "supercell", such that the magnetic flux is commensurate (i.e. an integer number of flux quanta ϕ_0 thread the supercell): in this case a continuous **k** vector can be defined in the magnetic Brillouin zone. It must be stressed that **k** here plays the same role as a Bloch vector, although it is *not* a Bloch vector in the conventional sense.

As in Sect. 3.9.1, we define $|\psi_{j\mathbf{k}}\rangle = \mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}}|u_{j\mathbf{k}}\rangle$ and $H_{\mathbf{k}} = \mathrm{e}^{-i\mathbf{k}\cdot\mathbf{r}}H\mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}}$; the latter takes here the form

$$H_{\mathbf{k}} = \frac{1}{2m}\left[\mathbf{p} + \hbar\mathbf{k} + \frac{e}{c}\mathbf{A}(\mathbf{r})\right]^2 + \mathcal{V}(\mathbf{r}), \tag{4.19}$$

where $\mathcal{V}$ is the substrate potential. The velocity can be expressed as

$$\mathbf{v} = \frac{1}{\hbar}\nabla_{\mathbf{k}}H_{\mathbf{k}}, \tag{4.20}$$

a formula often recurring in the present Notes in various forms, see e.g. Eqs. (1.18) and (2.20).

The Kubo formula for transverse conductivity is [67]

$$\sigma_{xy} = 2\hbar e^2 \,\mathrm{Im}\, \sum_{jj'} \frac{1}{(2\pi)^2} \int_{\epsilon_{j\mathbf{k}} < \epsilon_{\mathrm{F}} < \epsilon_{j'\mathbf{k}}} d\mathbf{k}\, \frac{\langle u_{j\mathbf{k}}|v_x|u_{j'\mathbf{k}}\rangle\langle u_{j'\mathbf{k}}|v_y|u_{j\mathbf{k}}\rangle}{(\epsilon_{j\mathbf{k}} - \epsilon_{j'\mathbf{k}})^2} \tag{4.21}$$

51

$$= 2\frac{e^2}{\hbar} \operatorname{Im} \sum_{jj'} \frac{1}{(2\pi)^2} \int_{\epsilon_{j\mathbf{k}}<\epsilon_{\mathrm{F}}<\epsilon_{j'\mathbf{k}}} d\mathbf{k} \, \frac{\langle u_{j\mathbf{k}}|\partial_x H_{\mathbf{k}}|u_{j'\mathbf{k}}\rangle\langle u_{j'\mathbf{k}}|\partial_y H_{\mathbf{k}}|u_{j\mathbf{k}}\rangle}{(\epsilon_{j\mathbf{k}} - \epsilon_{j'\mathbf{k}})^2}.$$

If we now consider the case where the Fermi level lies in a gap, with n filled bands (Landau levels in a flat potential), Eq. (4.21) becomes the BZ integral

$$\sigma_{xy} = 2\frac{e^2}{\hbar} \operatorname{Im} \frac{1}{(2\pi)^2} \int_{\mathrm{BZ}} d\mathbf{k} \sum_{j=1}^{n} \sum_{j'=n+1}^{\infty} \frac{\langle u_{j\mathbf{k}}|\partial_x H_{\mathbf{k}}|u_{j'\mathbf{k}}\rangle\langle u_{j'\mathbf{k}}|\partial_y H_{\mathbf{k}}|u_{j\mathbf{k}}\rangle}{(\epsilon_{j\mathbf{k}} - \epsilon_{j'\mathbf{k}})^2}. \qquad (4.22)$$

The integrand is just a simple generalization of the sum-over-states formula of Eq. (3.30). Using the same arguments as in Ch. 3 it is rather straightforward to arrive at the identity

$$\operatorname{Im} \sum_{j=1}^{n} \sum_{j'=n+1}^{\infty} \frac{\langle u_{j\mathbf{k}}|\partial_x H_{\mathbf{k}}|u_{j'\mathbf{k}}\rangle\langle u_{j'\mathbf{k}}|\partial_y H_{\mathbf{k}}|u_{j\mathbf{k}}\rangle}{(\epsilon_{j\mathbf{k}} - \epsilon_{j'\mathbf{k}})^2}$$

$$= \operatorname{Im} \sum_{j=1}^{n} \langle \partial_x u_{j\mathbf{k}}|\partial_y u_{j\mathbf{k}}\rangle = -\frac{1}{2}\mathbf{\Omega}_{xy}(\mathbf{k}), \qquad (4.23)$$

where the many-band Berry curvature, Eq. (3.38), appears. Since the BZ is a torus, the BZ integral of the curvature equals 2π times an integer, the (first) Chern number C_1. The milestone TKNN discovery is that Hall conductivity is a Chern number when expressed in klitzing^{-1}:

$$\sigma_{xy} = -\frac{e^2}{h}C_1. \qquad (4.24)$$

Notice that the sign choices are not uniform across the literature.

Conductivity is a property of the *excitations* of the system, as it is perspicuous in the Kubo formula above. The Chern number, instead, is a *ground state* property. The identity relating them belongs to the general class of fluctuation-dissipation theorems, although this looks like an oxymoron, the Hall conductivity being here dissipationless. The interpretation of the Chern number as a ground-state quantum fluctuation will be elaborated somewhere else in the present Notes.

The topological nature of the observable explains its extreme robustness under variations of magnetic field, carrier density, substrate disorder, and more. The topological invariant C is identified with the filling ν using the same arguments as in Sect. 2.5.4; the integer can only be varied by crosssing a conducting state. It is expedient to change the sign, thus identifying $\nu = -C$. Since in the quantum-Hall regime the conductivity tensor is off-diagonal, the Hall resistance (or resistivity) is

$$\rho_{xy} = \frac{1}{\nu}\frac{h}{e^2}. \qquad (4.25)$$

52

One final observation is in order: we have not counted a factor of two for double spin occupation, for a good reason: at the high **B** fields needed to observe quantization the Zeeman separation between the two spin states is much larger than the Landau level separation. Therefore the electrons entering the quantum Hall effect are spin polarized (or "spinless"). We quote here the fact that in graphene instead the experimental conditions for the quantum Hall effect are different, and the factor of two has to be restored.

4.2.5 Classical limit of TKNN

By virtue of topology, in the quantized (high field) regime the Hall resistivity ρ_{xy} is insensitive to carrier density n: this was indeed the revolutionary finding of von Klitzing and collaborators (see Fig. 2.8 and its caption). In the classical (low field) regime, instead, ρ_{xy} is linear in $1/n$. We reproduce here the Drude-Zener formula of Eq. (2.11):

$$\rho_{xx} = \frac{m}{ne^2\tau}, \qquad \rho_{xy} = \frac{B}{nec} = \frac{m\omega_{\mathrm{c}}}{ne^2}, \tag{4.26}$$

where $\omega_{\mathrm{c}} = \frac{eB}{mc}$.

As discussed in Sec. 2.5.2 the degeneracy of each Landau level in a flat potential is ϕ/ϕ_0, where ϕ is the magnetic flux through the sample. The density of states is

$$\mathcal{D}(\epsilon) = \frac{\phi}{\phi_0} \sum_n \delta[\,\epsilon - (n+1/2)\hbar\omega_{\mathrm{c}}\,], \tag{4.27}$$

shown in Fig. 2.11(a). When a disordered substrate is present $\mathcal{D}(\epsilon)$ becomes instead of the kind shown in Fig. 2.11(b). The density of states is conserved in average: whenever the chemical potential μ is much larger than $\hbar\omega_{\mathrm{c}}$ the number of occupied states will be the same in the clean and in the disordered case. The density in the $\mu \gg \hbar\omega_{\mathrm{c}}$ limit is therefore

$$n = \frac{2}{A} \int_0^\mu d\epsilon \, \mathcal{D}(\epsilon) \simeq \frac{B}{\phi_0} \frac{2\mu}{\hbar\omega_{\mathrm{c}}} = \frac{B}{\phi_0}\nu. \tag{4.28}$$

Notice that ν includes now the spin factor, since the Zeeman splitting vanishes in the $B \to 0$ limit. Expressing now Eq. (4.25) in terms of n we have

$$\rho_{xy} = \frac{B}{n\phi_0} \frac{h}{e^2} = \frac{B}{nec}, \tag{4.29}$$

which is indeed the classical formula of Eq. (4.26).

At this point we are ready to address the classical limit in the experimental data of Fig. 2.9, reproduced here as Fig. 4.5. In this sample at the conditions of
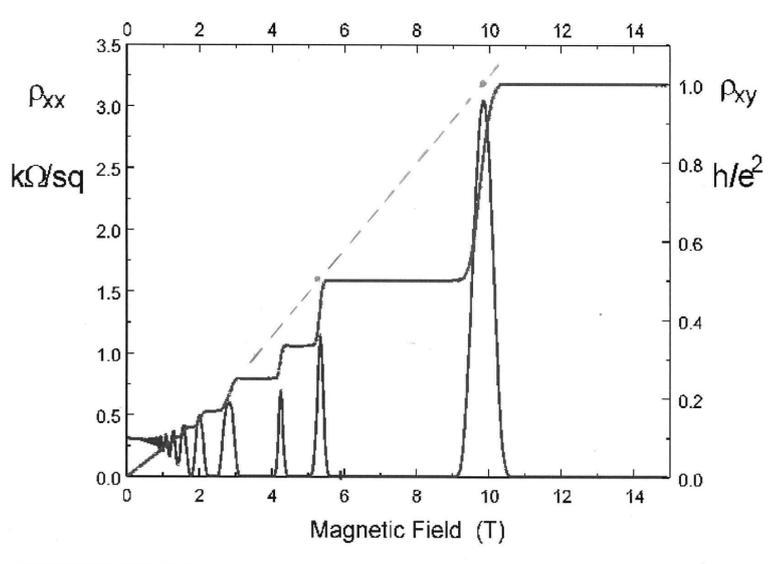
Figure 4.5: A modern realization of the integer quantum Hall effect. The dashed line is the corresponding Drude-Zener transverse conductivity.

the actual experiment the classical-quantum transition occurs somewhere between 1 and 2 tesla: below this B value the system becomes dissipative ($\rho_{xx} \neq 0$), and the plateaus in ρ_{xy} disappear. At $B \simeq 1\mathrm{T}$ we have $\rho_{xy} \simeq \rho_{xx}$; the Drude-Zener formulæ of Eq. (4.26) suggest therefore that the effective relaxation time for this sample is $\tau \simeq 1/\omega_c$, evaluated at $B \simeq 1-2$ tesla.

In Fig. 4.5 I have indicated (dashed line) the classical Drude-Zener transverse conductivity for this sample: it is linear in B and coincides with the exact conductivity for integer values of the filling. It is also realized that the slope of the dashed line is *larger* than the experimental one in the low-B regime. The reason is that the effective carrier density in the high-B regime is one-half the one in the low-B regime: electrons are spin-polarized in the former case, and unpolarized in the latter.

4.3 Adiabatic approximation in a magnetic field

The general problem of the nuclear motion—both classical and quantum—in presence of an external magnetic field has been first solved in 1988 by Schmelcher, Cederbaum, and Meyer [68]. It is remarkable that such a fundamental problem was solved so late, and that even today the relevant literature is ignored by textbooks and little cited. The solution is a manifestation of geometrical effects in electronic wavefunctions [69], which appears in a spectacular way even when the nuclear motion

is addressed at the classical level.

When a genuine magnetic field, generated by some external source, acts on the molecular system, the Hamiltonian of Eq. (4.13) is modified by the addition of a vector potential term in the kinetic energies of both the nuclei and the electrons. Proceeding as in the zero field case, one writes an ansatz wavefunction and arrives at the effective Hamiltonian for the ionic motion, Eq. (4.16), where an extra term must be added to the kinematical momentum $\boldsymbol{\pi}$ of Eq. (4.15). There are thus *two* vector potentials in the effective nuclear Hamiltonian: a geometric one, and a genuinely magnetic one.

However, with respect to the zero-field case, there is a qualitative difference whose importance is overwhelming. Since the electronic Hamiltonian is no longer invariant under T-symmetry, the electronic wavefunction is necessarily complex, and the curvature is in general nonzero. No singularity is needed to produce geometrical effects on the nuclear motion; the Berry phase will be in general nonzero on any path in the space of nuclear coordinates.

Suppose we are interested into the nuclear motion at the purely classical level. The Hamiltonian of Eq. (4.16)—whose kinematical momentum $\boldsymbol{\pi}$ includes now the two different vector potentials—yields the Hamilton equations of motion, which can be transformed into the Newton equations of motion: within the latter, the effects of the vector potentials appear in terms of fields, in the form of Lorentz forces. The curl of the magnetic vector potential obviously yields the magnetic field due to the external source; the curl of the geometric vector potential (Berry curvature) yields an additional "magnetic-like" field which is *nonzero* even on the classical trajectory of the nuclei. We stress that this is at variance with the zero-field case, where the Berry phase had no effect on the ionic motion at the classical level, and could only be detected when quantizing the ionic degrees of freedom.

Within a naïve Born–Oppenheimer approximation—where Berry phases are neglected—the magnetic field acts on the nuclei as it they were "naked" charges: a proper treatment must instead account for electronic screening: this is provided by the geometric vector potential. Surprisingly, there are very few calculations of the effect: it is pretty clear, however, that the geometric term is no small correction.

4.3.1 Hydrogen atom in a constant B field

We consider the very pedagogical example of a hydrogen atom, hence we identify the electronic degrees of freedom [$\mathbf{x}$], used in Sect. 4.2.2, with a single coordinate $\mathbf{r}$, and the parameter $\boldsymbol{\xi}$ with the nuclear coordinate $\mathbf{R}$. If the atom is subject only to a magnetic field—and neglecting irrelevant spin-dependent terms—the complete (nonrelativistic) Hamiltonian $\mathcal{H}$ and the electronic Hamiltonian H are

$$\mathcal{H}(\mathbf{R}, \mathbf{r}) = \frac{1}{2M} \left[\mathbf{p} - \frac{e}{c} \mathbf{A}(\mathbf{R}) \right]^2 + H(\mathbf{R}, \mathbf{r}); \qquad (4.30)$$

$$H(\mathbf{R}, \mathbf{r}) = \frac{1}{2m} \left[-i\hbar\nabla_{\mathbf{r}} + \frac{e}{c}\mathbf{A}(\mathbf{r}) \right]^2 - \frac{e^2}{|\mathbf{r} - \mathbf{R}|}. \tag{4.31}$$

As explained above, the nuclear kinematical momentum of Eq. (4.15) becomes

$$\boldsymbol{\pi} = \mathbf{p} - \hbar\boldsymbol{\mathcal{A}}(\mathbf{R}) - \frac{e}{c}\mathbf{A}(\mathbf{R}) \tag{4.32}$$

The case of a constant $\mathbf{B}$ field can be dealt with analytically. We choose the central gauge $\mathbf{A}(\mathbf{r}) = \frac{1}{2}\mathbf{B} \times \mathbf{r}$. If $\phi(\mathbf{r})$ is the exact ground eigenfunction when the proton sits at $\mathbf{R} = 0$, the eigenfunction at a generic $\mathbf{R}$ is:

$$\langle \mathbf{r}|\psi(\mathbf{R})\rangle = e^{-\frac{ie}{2\hbar c}\mathbf{r}\cdot\mathbf{B}\times\mathbf{R}}\phi(\mathbf{r} - \mathbf{R}), \tag{4.33}$$

with an $\mathbf{R}$-independent eigenvalue. The Berry connection is clearly

$$\begin{aligned}
\boldsymbol{\mathcal{A}}(\mathbf{R}) &= i\langle\psi(\mathbf{R})|\nabla_{\mathbf{R}}\psi(\mathbf{R})\rangle = -\frac{e}{2\hbar c}\langle\psi(\mathbf{R})|\mathbf{B}\times\mathbf{r}|\psi(\mathbf{R})\rangle = -\frac{e}{2\hbar c}\mathbf{B}\times\mathbf{R} \\
&= -\frac{e}{\hbar c}\mathbf{A}(\mathbf{R}), \tag{4.34}
\end{aligned}$$

since the $\mathbf{R}$-derivative of $\phi(\mathbf{r} - \mathbf{R})$ does not contribute. Replacing Eq. (4.34) into Eq. (4.32) we find $\boldsymbol{\pi} = \mathbf{p}$, as it must be: the nucleus travels at constant speed, and is not deflected by a Lorentz force.

Remarkably, the "magnetic-like" field due to the Berry phase is—in this simple example—exactly opposite to the external magnetic field, thus providing the complete screening which is physically expected.

4.3.2   A molecule in a constant B field

Quite similarly to the atomic case considered above, even when a neutral molecule travels at constant speed the cancellation is complete [69]; rotational and vibrational motions are less trivial. A paradigmatic example of rotovibrations is considered by Ceresoli et al. in Ref. [70]: a H_2 molecule with a simple analytical wavefunction.

We consider here in more detail the simpler case of pure rotations of a rigid molecule along the $\mathbf{B}$ axis, following Ref. [69, 71]. We will define below a screening factor σ to measure the amount of cancellation; $\sigma = -1$ (complete screening) is what indeed occurs for rigid translational motions; the other limit $\sigma = 0$ would correspond to bare nuclei (or ion cores, if we deal with valence electrons only).

Consider first the motion of a (nuclear) charge eZ on a circular path. In one cycle the bare magnetic phase is $\gamma_B = 2\pi Z\phi/\phi_0$, where ϕ is the magnetic flux through the path and $\phi_0 = hc/e$ is the flux quantum. For the rotation of a rigid molecule the bare magnetic phase is then

$$\gamma_B = B\frac{e}{c\hbar}\pi\sum_\ell Z_\ell R_\ell^2, \tag{4.35}$$

where R_ℓ is the radius of the nuclear orbit. Now we consider the many-body ground state as a function of B and the rotational angle θ: the Berry phase is

$$\gamma = i \int_0^{2\pi} d\theta \, \langle \Psi_0 | \partial_\theta \Psi_0 \rangle, \tag{4.36}$$

and is linear in B (we assume that $|\Psi_0\rangle$ is the singlet ground state of a T-invariant Hamiltonian). Both phases in Eqs. (4.35) and (4.36) do not depend on the electromagnetic gauge, but they depend on where the rotation axis is set; we assume that the molecule is rotated around its center of mass. Eqs. (4.35) and (4.36) have been computed for several molecules by Ceresoli and Tosatti [71] in order to evaluate their rotational g factor, which is the dimensionless ratio between the total molecular magnetic moment and the mechanical momentum (expressed with respect of the center of mass); in terms of the screening factor $\sigma = \gamma/\gamma_B$ one has

$$g = (\sigma + 1) \frac{\sum_\ell Z_\ell R_\ell^2}{\sum_\ell M_\ell R_\ell^2}, \tag{4.37}$$

where the mass is expressed in proton units. The rotational g factor is experimentally accessible via molecular beam resonance or microwave spectroscopy techniques.

We introduce the Berry curvature at $B = 0$ in the B, θ variables:

$$\Omega_{B\theta} = i(\langle \partial_B \Psi_0 | \partial_\theta \Psi_0 \rangle - \langle \partial_\theta \Psi_0 | \partial_B \Psi_0 \rangle). \tag{4.38}$$

It is further expedient to adopt the central gauge, and to set the center of the gauge at the rotation axis: in this case the curvature is θ-independent, and to linear order in B the Berry phase is $\gamma = 2\pi B \Omega_{B\theta}$. This Berry phase has an interesting relationship to the linear magnetizability of the molecule $\chi_{zz} = \partial m_z / \partial B$ (we set $\mathbf{B}$ along the z-axis). If we adopt the central gauge $\mathbf{A}(\mathbf{r}) = \frac{1}{2}\mathbf{B} \times \mathbf{r}$ the electronic Hamiltonian is

$$\hat{H} = \hat{H}_0 + \frac{eB}{2mc}\hat{L}_z + \frac{e^2 B^2}{8mc^2}\sum_i (x_i^2 + y_i^2), \tag{4.39}$$

where $\hat{L}_z$ is the many-body angular-momentum operator. The magnetizability χ_{zz} is by definition (minus) the second derivative of the ground-state energy with respect to B: it is comprised of two terms, deriving from the first (paramagnetic) and second (diamagnetic) B-dependent terms in Eq. (4.39). We consider in the following the paramagnetic response only $\chi_{zz}^{(\mathrm{p})}$:

$$\chi_{zz}^{(\mathrm{p})} = \frac{e^2}{2m^2 c^2} \sum_{n \neq 0} \frac{\langle \Psi_0 | \hat{L}_z | \Psi_n \rangle \langle \Psi_0 | \hat{L}_z | \Psi_n \rangle}{E_n - E_0} \tag{4.40}$$

The Hamiltonian of Eq. (4.39) clearly depends on the gauge origin. Magnetic properties of molecules are notoriously plagued by gauge-invariance problems, when

evaluated with variational wavefunctions and incomplete basis sets. We do not discuss the issue here, assuming we are working on a complete basis; however, in order to compare to the Berry curvature $\Omega_{B\theta}$, we set the gauge origin at the center of mass. The curvature, according to Eq. (3.31), can be rewritten as:

$$
\begin{aligned}
\Omega_{B\theta} &= -2\operatorname{Im}\sum_{n\neq 0}\frac{\langle\Psi_0|\partial_B\hat{H}|\Psi_n\rangle\langle\Psi_0|\partial_\theta\hat{H}|\Psi_n\rangle}{(E_n-E_0)^2}\\
&= -\frac{e}{mc}\operatorname{Im}\sum_{n\neq 0}\frac{\langle\Psi_0|\hat{L}_z|\Psi_n\rangle\langle\Psi_0|\partial_\theta\hat{H}|\Psi_n\rangle}{(E_n-E_0)^2}.
\end{aligned}
\tag{4.41}
$$

Next we wish to evaluate $\partial_\theta\hat{H}=\partial_\theta\hat{H}_0$, which is (minus) the torque operator [72]. The Hamiltonian transforms under rotation by an angle θ as:

$$
\hat{H}_0 \quad\to\quad \mathrm{e}^{i\frac{\theta}{\hbar}\hat{L}_z}\hat{H}_0\mathrm{e}^{-i\frac{\theta}{\hbar}\hat{L}_z},
\tag{4.42}
$$

$$
\partial_\theta\hat{H}_0=\frac{i}{\hbar}[\hat{L}_z,\hat{H}_0],\qquad \langle\Psi_0|\,\partial_\theta\hat{H}_0\,|\Psi_n\rangle=\frac{i}{\hbar}(E_n-E_0)\langle\Psi_0|\,\hat{L}_z\,|\Psi_n\rangle
\tag{4.43}
$$

$$
\Omega_{B\theta}=\frac{e}{\hbar mc}\sum_{n\neq 0}\frac{\langle\Psi_0|\hat{L}_z|\Psi_n\rangle\langle\Psi_0|\hat{L}_z|\Psi_n\rangle}{E_n-E_0}.
\tag{4.44}
$$

We thus arrive at the important relationship

$$
\chi_{zz}^{(\mathrm{p})}=\frac{\hbar e}{2mc}\Omega_{B\theta}=\frac{\hbar e}{2mc}\frac{1}{2\pi}\frac{\partial\gamma}{\partial B}.
\tag{4.45}
$$

4.4 Semiclassical transport

The semiclassical theory of Bloch electron dynamics plays a fundamental role in the physics of metals and semiconductors, and is a typical textbook topic [73].

At the classical level, particle transport is considered to take place in phase space, whose points are labeled by $\mathbf{r}$ and $\mathbf{p}$, and is described by the distribution function $f(\mathbf{r},\mathbf{p})$ which satisfy the Boltzmann equations. Electrons in crystalline materials may fit into this framework if we identify the particles a wave packets constructed from the Bloch orbitals. The semiclassical theory addresses therefore the motion of a wave packet built as a superposition of Bloch states from the n-th band

$$
|W\rangle=\int_{\mathrm{BZ}}d\mathbf{k}\,a(\mathbf{k},t)|\psi_{n\mathbf{k}}\rangle,
\tag{4.46}
$$

where the envelope function is well localized in $\mathbf{k}$-space. Because of this, it is delocalized in $\mathbf{r}$ space; we assume, however, that its center of mass is well defined. Owing to this, we may define the wave vector $\mathbf{k}$ and the center $\mathbf{r}$ of the wave packet as

$$
\mathbf{k}=\int_{\mathrm{BZ}}d\mathbf{k}'\,\mathbf{k}'\,|a(\mathbf{k}',t)|^2;\qquad \mathbf{r}=\langle W|\mathbf{r}|W\rangle.
\tag{4.47}
$$

4.4.1 Textbook equations of motion

In absence of collisions, the equations of motion (EoM) reported in textbooks and routinely used in device engineering are

$$\dot{\mathbf{r}} = \frac{1}{\hbar}\frac{\partial \epsilon_{\mathbf{k}}}{\partial \mathbf{k}}$$
$$\hbar\dot{\mathbf{k}} = -e\left(\mathbf{E} + \frac{1}{c}\dot{\mathbf{r}} \times \mathbf{B}\right), \tag{4.48}$$

where $\epsilon_{\mathbf{k}}$ is the band structure of the relevant band, and $\mathbf{E}$ and $\mathbf{B}$ are the perturbing fields, assumed weak and slowly varying in space and time.

As emphasized in Ref. [73], the derivation of Eq. (4.48), despite the formal simplicity of the result, is "a formidable task". The early derivations date from the 1930s; the problem was reconsidered several times in the literature, by Slater [74], Luttinger [75], and Zak [76] among others.

4.4.2 Modern equations of motion

The modern analysis of semiclassical transport owes to a couple of papers by Q. Niu and coworkers [77, 78] (see also Ref. [13]). They start observing that a wave packet generally possesses two kinds of motion: the center-of-mass motion and the self-rotation around its center. Owing to the latter, the wave packet is endowed with an orbital magnetic moment, whose expression is

$$\mathbf{m}(\mathbf{k}) = -\frac{ie}{2\hbar c}\langle \nabla_{\mathbf{k}} u_{\mathbf{k}}| \times (H_{\mathbf{k}} - \epsilon_{\mathbf{k}}) |\nabla_{\mathbf{k}} u_{\mathbf{k}}\rangle. \tag{4.49}$$

Because of this, the band structure acquires a $\mathbf{B}$-dependent term

$$\epsilon_{\mathbf{k}} \to \tilde{\epsilon}_{\mathbf{k}} = \epsilon_{\mathbf{k}} - \mathbf{m}(\mathbf{k}) \cdot \mathbf{B}; \tag{4.50}$$

Furthermore, the canonical momentum $\hbar\mathbf{k}$ has to be replaced with the kinetic momentum, which includes the (geometrical) vector potential. In the Newton-like EoM, its contribution is reminiscent of a Lorentz force in reciprocal space. Eq. (4.48) must then be replaced with

$$\dot{\mathbf{r}} = \frac{1}{\hbar}\frac{\partial \tilde{\epsilon}_{\mathbf{k}}}{\partial \mathbf{k}} - \dot{\mathbf{k}} \times \boldsymbol{\Omega}(\mathbf{k})$$
$$\hbar\dot{\mathbf{k}} = -e\left(\mathbf{E} + \frac{1}{c}\dot{\mathbf{r}} \times \mathbf{B}\right), \tag{4.51}$$

where $\boldsymbol{\Omega}(\mathbf{k})$ is the Berry curvature of the relevant band, having the dimensions of a squared length. Notice that the curvature is nonzero even in presence of T-symmetry, provided that the crystal is noncentrosymmetric. The term $\dot{\mathbf{k}} \times \boldsymbol{\Omega}(\mathbf{k})$ goes under the name of "anomalous velocity".

4.4.3 Equations of motion in symplectic form

A more accurate analysis performed in 2003 by Panati, Spohn, and Teufel [79] proves that Eq. (4.51) is not yet the ultimate form for the semiclassical EoM; it is only correct if $\mathbf{B}$ is constant in space. Furthermore, the equations which supersede Eq. (4.51) are more symmetrical and elegant.

One starts defining

$$\mathscr{H}(\mathbf{r}, \mathbf{k}) = \epsilon_{\mathbf{k}} - e\phi(\mathbf{r}) - \mathbf{m}(\mathbf{k}) \cdot \mathbf{B}(\mathbf{r}), \qquad (4.52)$$

where $\phi(\mathbf{r})$ is the potential of the electric field $\mathbf{E}(\mathbf{r})$, and $\mathbf{m}(\mathbf{k})$ is the wave packet's moment, Eq. (4.49). Then the EoM are

$$\begin{aligned}
\dot{\mathbf{r}} &= \frac{1}{\hbar} \frac{\partial \mathscr{H}(\mathbf{r}, \mathbf{k})}{\partial \mathbf{k}} - \dot{\mathbf{k}} \times \mathbf{\Omega}(\mathbf{k}) \\
\dot{\mathbf{k}} &= -\frac{1}{\hbar} \frac{\partial \mathscr{H}(\mathbf{r}, \mathbf{k})}{\partial \mathbf{r}} - \frac{e}{\hbar c} \dot{\mathbf{r}} \times \mathbf{B}(\mathbf{r}), .
\end{aligned} \qquad (4.53)$$

In order to realize the relationship of Eq. (4.53) to the standard symplectic form of Hamilton's equations, it is expedient to adopt atomic units, and furthermore to redefine $\mathbf{B}$ by including the $1/c$ factor: we get

$$\begin{aligned}
\dot{\mathbf{r}} &= \nabla_{\mathbf{p}}\mathscr{H}(\mathbf{r}, \mathbf{p}) - \dot{\mathbf{p}} \times \mathbf{\Omega}(\mathbf{p}) \\
\dot{\mathbf{p}} &= -\nabla_{\mathbf{r}}\mathscr{H}(\mathbf{r}, \mathbf{p}) - \dot{\mathbf{r}} \times \mathbf{B}(\mathbf{r}).
\end{aligned} \qquad (4.54)$$

Notice that the translational momentum is coupled to $\mathbf{B}$ via the explicit Lorentz force $\dot{\mathbf{r}} \times \mathbf{B}(\mathbf{r})$, unlike in the standard Hamiltonian formulation where such coupling would appear via a vector-potential-dependent $\mathscr{H}$. It is easy to prove that Eq. (4.54) conserves the energy:

$$\frac{d\mathscr{H}}{dt} = \nabla_{\mathbf{r}}\mathscr{H} \cdot \dot{\mathbf{r}} + \nabla_{\mathbf{p}}\mathscr{H} \cdot \dot{\mathbf{p}} = 0 \qquad (4.55)$$

In the simple case where $\mathbf{\Omega}$ and $\mathbf{B}$ are identically vanishing, $\mathscr{H}$ becomes the standard classical Hamiltonian and Eq. (4.54) are the corresponding Hamilton's equations. If $\mathbb{I}$ is the 3×3 identity, the standard 6×6 symplectic matrix is

$$\Theta = \begin{pmatrix} 0 & -\mathbb{I} \\ \mathbb{I} & 0 \end{pmatrix}, \qquad (4.56)$$
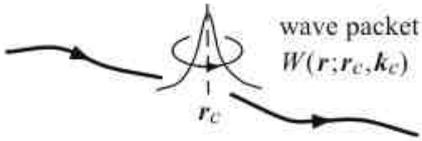
Figure 4.6: A wave packet carries quasi momentum $\hbar\mathbf{k}$, but also an orbital moment and a corresponding magnetic moment $\mathbf{m}(\mathbf{k})$

and the symplectic form of Eq. (4.54) is

$$\Theta \begin{pmatrix} \dot{\mathbf{r}} \\ \dot{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \nabla_{\mathbf{r}} \mathscr{H} \\ \nabla_{\mathbf{p}} \mathscr{H} \end{pmatrix}. \tag{4.57}$$

When either $\mathbf{B}$ or $\boldsymbol{\Omega}$ (or both) are nonzero, one recasts identically Eq. (4.54) in the similar form

$$\tilde{\Theta} \begin{pmatrix} \dot{\mathbf{r}} \\ \dot{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \nabla_{\mathbf{r}} \mathscr{H} \\ \nabla_{\mathbf{p}} \mathscr{H} \end{pmatrix}, \tag{4.58}$$

where $\tilde{\Theta}$ is a deformed symplectic matrix defined as

$$\tilde{\Theta} = \begin{pmatrix} \vec{B}(\mathbf{r}) & -\mathbb{I} \\ \mathbb{I} & \vec{\Omega}(\mathbf{p}) \end{pmatrix}, \tag{4.59}$$

and the 3×3 antisymmetric matrices $\vec{B}(\mathbf{r})$ and $\vec{\Omega}(\mathbf{p})$ are

$$\vec{B} = \begin{pmatrix} 0 & -B_z & B_y \\ B_z & 0 & -B_x \\ -B_y & B_x & 0 \end{pmatrix}, \quad \vec{\Omega} = \begin{pmatrix} 0 & -\Omega_z & \Omega_y \\ \Omega_z & 0 & -\Omega_x \\ -\Omega_y & \Omega_x & 0 \end{pmatrix}. \tag{4.60}$$

Finally, we notice that if we set to zero both the Berry curvature $\vec{\Omega}$ and the moment $\mathbf{m}(\mathbf{k})$ in Eq. (4.52), then Eq. (4.58) provides an alternative form for Hamilton's equations, where the $\mathbf{B}$ field appears in the symplectic matrix and no vector potential appears in the Hamiltonian $\mathscr{H}$. This form is well known in classical mechanics, and goes under the name of gauge-invariant Hamiltonian formulation.

4.4.4 Geometrical correction to the density of states

In a remarkable 2005 paper by Xiao, Shi, and Niu [80] it was pointed out that Eq. (4.51), in presence of a nonzero $\mathbf{B}$ field, violates Liouville's theorem. This means that the volume element $\Delta V = \Delta \mathbf{r} \Delta \mathbf{p}$ changes in time during the evolution of the system; it is possible, however to remedy this shortcoming in an elegant way.

We are not following the original paper; instead, we start from the more elegant symplectic formulation above. The time evolution of the volume element is

$$\begin{aligned} \frac{1}{\Delta V} \frac{\partial V}{\partial t} &= \nabla_{\mathbf{r}} \cdot \dot{\mathbf{r}} + \nabla_{\mathbf{p}} \cdot \dot{\mathbf{p}} \\ &= \nabla_{\mathbf{r}} \cdot [\nabla_{\mathbf{p}} \mathscr{H} - \dot{\mathbf{p}} \times \boldsymbol{\Omega}(\mathbf{p})] + \nabla_{\mathbf{p}} \cdot [-\nabla_{\mathbf{r}} \mathscr{H} - \dot{\mathbf{r}} \times \mathbf{B}(\mathbf{r})] \\ &= -\nabla_{\mathbf{r}} \cdot [\dot{\mathbf{p}} \times \boldsymbol{\Omega}(\mathbf{p})] - \nabla_{\mathbf{p}} \cdot [\dot{\mathbf{r}} \times \mathbf{B}(\mathbf{r})]. \end{aligned} \tag{4.61}$$

After a somewhat lengthy calculation, one arrives at a very simple result, which—most notably—is a total derivative

$$\frac{1}{\Delta V} \frac{\partial V}{\partial t} = -\frac{d}{dt} \ln [1 + \mathbf{B}(\mathbf{r}) \cdot \boldsymbol{\Omega}(\mathbf{p})], \tag{4.62}$$

61

and is clearly nonzero whenever $\mathbf{B}$ and $\mathbf{\Omega}$ are non constant in phase space. It is worth mentioning that Eq. (4.62) bears a simple relationship to the deformed symplectic matrix, Eq. (4.59): in fact

$$1 + \mathbf{B}(\mathbf{r}) \cdot \mathbf{\Omega}(\mathbf{p}) = \sqrt{\det \tilde{\Theta}(\mathbf{r}, \mathbf{p})}. \qquad (4.63)$$

In order to arrive at Eqs. (4.62) and (4.63) two special features of the matrix $\tilde{\Theta}(\mathbf{r}, \mathbf{p})$ seem to be essential: its elements are either functions of $\mathbf{r}$ or $\mathbf{p}$, but not of both; the divergences of $\mathbf{B}(\mathbf{r})$ and $\mathbf{\Omega}(\mathbf{p})$ are both zero.

Within ordinary statistical mechanics, the density of states in phase space is $1/h^d$ (in dimension d). The number of states in volume $\Delta V = \Delta \mathbf{r} \Delta \mathbf{p}$ is $\Delta V / h^3$, and—owing to Liouville's theorem—this number remains constant during time evolution. Here instead, such number remains constant only if we appropriately modify the density of states. For the sake of clarity we restore Gauss units; the modification needed is

$$\frac{1}{h^d} \longrightarrow \frac{1}{h^d}\Big[1 + \frac{2\pi}{\phi_0} \mathbf{B}(\mathbf{r}) \cdot \mathbf{\Omega}(\mathbf{p}/\hbar)\Big], \qquad (4.64)$$

where ϕ_0 is the flux quantum.

While Xiao, Shi, and Niu [80] made a fundamental discovery in physics, the mathematical formalism was apparently well known within advanced analytical mechanics. It was in fact later pointed out that—within a generalized formulation of Hamiltonian dynamics—Eq. (4.51) are indeed "Hamiltonian", and a generalized form of Liouville's theorem holds [81]. I would regard this as a matter of semantics; nonetheless the proof provided by such mathematical-physics tools is more general and more elegant than the one arrived at in Ref. [80].

4.4.5 Outstanding consequences of the modified density of states

After Ref. [80] appeared, it become immediately clear that the geometrical correction to the density of states, Eq. (4.64), goes well beyond the scope of semiclassical approximation. Statistical mechanics paraphernalia, such as partition functions and the like, must adopt Eq. (4.64).

As a consequence of the modified density of states, the Fermi volume of a metal changes when a macroscopic $\mathbf{B}$ field is switched on at constant electron density. If instead we keep the chemical potential μ constant, then the electron density n depends on $\mathbf{B}$. At zero temperature (for each spin channel)

$$
\begin{aligned}
n &= \frac{1}{(2\pi)^d} \int_{\text{BZ}} d\mathbf{k} \left(1 + \frac{2\pi}{\phi_0} \mathbf{B} \cdot \mathbf{\Omega}(\mathbf{k})\right) \vartheta(\mu - \epsilon_{\mathbf{k}}) \\
\left(\frac{\partial n}{\partial \mathbf{B}}\right)_\mu &= \frac{1}{(2\pi)^{d-1}\phi_0} \int_{\text{BZ}} d\mathbf{k}\, \mathbf{\Omega}(\mathbf{k})\, \vartheta(\mu - \epsilon_{\mathbf{k}})
\end{aligned}
\qquad (4.65)
$$

The latter assumes a perspicuous meaning in $2d$, when μ is in a gap:

$$\left(\frac{\partial n}{\partial B}\right)_{\mu} = \frac{1}{2\pi\phi_0}\int_{\text{BZ}} d\mathbf{k}\,\mathbf{\Omega}(\mathbf{k}) = \frac{C_1}{\phi_0} = -\frac{1}{ec}\sigma_{xy}. \tag{4.66}$$

For a quantum Hall system, this goes under the name of Streda formula [43], and had been first derived in 1982 in a very different way. We have given a very elementary proof above, Eq. (2.17).

4.5 Quantum transport

4.5.1 Transport by a single state

We are going to study here the current induced by an adiabatic change of the potential, or more generally of the Hamiltonian, in the single-particle case. We indicate as $\psi_n(t)$ the adiabatic instantaneous eigenstates, and with $\psi(t)$ the time evolution of the ground state. In order to get rid of the dynamical phase, it is better to deal with density matrices

$$\rho(t) = |\psi(t)\rangle\langle\psi(t)| = |\psi_0(t)\rangle\langle\psi_0(t)| + \Delta\rho(t). \tag{4.67}$$

The velocity of this state is

$$\begin{aligned}
\mathbf{v}(t) &= \text{Tr}\,\{\rho(t)\,\mathbf{v}\} \tag{4.68}\\
&= \langle\psi_0(t)|\,\mathbf{v}\,|\psi_0(t)\rangle + \sum_n \langle\psi_0(t)|\,\Delta\rho(t)\,|\psi_n(t)\rangle\langle\psi_n(t)|\,\mathbf{v}\,|\psi_0(t)\rangle.
\end{aligned}$$

The first term on the rhs is zero in the special case where—at all times—the Hamiltonian $H(t)$ is T-invariant and the state is nondegenerate.

We write the exact density matrix of the time-evolved ground state as

$$\hat{\rho}(t) = |\Psi_0(t)\rangle\langle\Psi_0(t)| + \Delta\hat{\rho}(t). \tag{4.69}$$

Since the adiabatic term commutes with $\hat{H}(t)$, the equation of motion is

$$[\,\hat{H}(t), \Delta\hat{\rho}(t)\,] = i\hbar\frac{d}{dt}\hat{\rho}(t) \simeq i\hbar\frac{d}{dt}|\Psi_0(t)\rangle\langle\Psi_0(t)|, \tag{4.70}$$

where we have neglected terms of higher order in the adiabaticity parameter. Taking then the matrix elements between $\langle\Psi_0(t)|$ and $|\Psi_n(t)\rangle$ one gets

$$(E_0 - E_n)\langle\Psi_0(t)|\Delta\hat{\rho}(t)|\Psi_n(t)\rangle = i\hbar\langle\dot{\Psi}_0(t)|\Psi_n(t)\rangle; \tag{4.71}$$

63

the $n = 0$ term vanishes because of norm conservation. Notice also that, to this order, $\langle \Psi_m(t)|\Delta\hat{\rho}(t)|\Psi_n(t)\rangle = 0$ when both $m \neq 0$ and $n \neq 0$; therefore

$$\Delta\hat{\rho}(t) = i\hbar \sum_{n \neq 0} |\Psi_0(t)\rangle \frac{\langle \dot{\Psi}_0(t)|\Psi_n(t)\rangle}{E_0 - E_n} \langle \Psi_n(t)| + \text{H.c.,} \tag{4.72}$$

where H.c. stays for Hermitian conjugate. Replacement into Eq. (4.68) yields

$$\begin{aligned}
\mathbf{v}(t) &= \langle \psi_0(t)| \, \mathbf{v} \, |\psi_0(t)\rangle \\
&\quad + i\hbar \sum_{n \neq 0} \left[\frac{\langle \dot{\psi}_0(t)|\psi_n(t)\rangle \langle \psi_n(t)| \, \mathbf{v} \, |\psi_0(t)\rangle}{E_0 - E_n} - \text{c.c.} \right].
\end{aligned} \tag{4.73}$$

4.5.2 Current carried by filled bands

We now exploit the previous result for a system of noninteracting electrons in the case where the Hamiltonian $H(t)$ is lattice periodical and the ground state is insulating; this means that the gap remains finite at all t. It will be enough to consider the simple case of just one filled band, with band index zero; the current carried by each Bloch orbital $|\psi_{0\mathbf{k}}\rangle = \mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}}|u_{0\mathbf{k}}\rangle$ is

$$\begin{aligned}
\mathbf{v}_{\mathbf{k}}(t) &= \langle \psi_{0\mathbf{k}}| \, \mathbf{v} \, |\psi_{0\mathbf{k}}\rangle + i\hbar \sum_{j \neq 0} \left[\frac{\langle \dot{\psi}_{0\mathbf{k}}|\psi_{j\mathbf{k}}\rangle \langle \psi_{j\mathbf{k}}| \, \mathbf{v} \, |\psi_{0\mathbf{k}}\rangle}{\epsilon_{0\mathbf{k}} - \epsilon_{j\mathbf{k}}} - \text{c.c.} \right] \\
&= \langle u_{0\mathbf{k}}| \, \mathbf{v} \, |u_{0\mathbf{k}}\rangle + i\hbar \sum_{j \neq 0} \left[\frac{\langle \dot{u}_{0\mathbf{k}}|u_{j\mathbf{k}}\rangle \langle u_{j\mathbf{k}}| \, \mathbf{v} \, |u_{0\mathbf{k}}\rangle}{\epsilon_{0\mathbf{k}} - \epsilon_{j\mathbf{k}}} - \text{c.c.} \right],
\end{aligned} \tag{4.74}$$

where the t dependence of the rhs is now implicit. We then adopt the usual formula for the velocity, Eqs. (1.18) and (5.16), and the analogue of Eq. (3.28):

$$\mathbf{v} = \frac{1}{\hbar} \nabla_{\mathbf{k}} H_{\mathbf{k}}, \quad |\nabla_{\mathbf{k}} u_{0\mathbf{k}}\rangle = \frac{1}{\hbar} \sum_{j \neq 0} |u_{j\mathbf{k}}\rangle \frac{\langle u_{j\mathbf{k}}| \, \mathbf{v} \, |u_{0\mathbf{k}}\rangle}{\epsilon_{0\mathbf{k}} - \epsilon_{j\mathbf{k}}} \tag{4.75}$$

$$v_{\mathbf{k}}(t) = \frac{1}{\hbar} \nabla_{\mathbf{k}} \epsilon_{0\mathbf{k}} + i \left(\langle \dot{u}_{0\mathbf{k}}|\nabla_{\mathbf{k}} u_{0\mathbf{k}}\rangle - \langle \nabla_{\mathbf{k}} u_{0\mathbf{k}}|\dot{u}_{0\mathbf{k}}\rangle \right). \tag{4.76}$$

The first term on the rhs integrates to zero over the BZ, while the second is clearly a Berry curvature component in the four-dimensional $\mathbf{k}, t$ domain. The current density carried by a filled band in dimension d is

$$\mathbf{j}(t) = -\frac{ie}{(2\pi)^d} \int_{\text{BZ}} d\mathbf{k} \, \left(\langle \dot{u}_{0\mathbf{k}}|\nabla_{\mathbf{k}} u_{0\mathbf{k}}\rangle - \langle \nabla_{\mathbf{k}} u_{0\mathbf{k}}|\dot{u}_{0\mathbf{k}}\rangle \right); \tag{4.77}$$

the Bloch states are normalized to one in the crystal cell (as everywhere in the present Notes).

4.5.3 Quantization of charge transport

Let us consider the special case of a simple cubic crystal with lattice constant a. The transported charge in time T in the z direction across one cell is

$$
\begin{aligned}
Q &= \int_0^T dt\, I_z(t) = a^2 \int_0^T dt\, j_z(t) = -\frac{ea^2}{(2\pi)^2} \int_{-\pi/a}^{\pi/a} dk_x \int_{-\pi/a}^{\pi/a} dk_y \\
&\times \frac{i}{2\pi} \int_0^T dt \int_{-\pi/a}^{\pi/a} dk_z\, (\,\langle \dot{u}_{0\mathbf{k}} | \partial_z u_{0\mathbf{k}} \rangle - \langle \partial_z u_{0\mathbf{k}} | \dot{u}_{0\mathbf{k}} \rangle\,).
\end{aligned} \tag{4.78}
$$

If the time evolution of the Hamiltonian is cyclic $H(T) = H(0)$, then the second line in Eq. (4.78) is clearly a Chern number C (in the k_z, t variables) and is integer. Notice also that C is dimensionless, and therefore does not depend on how fast the Hamiltonian varies with time; ideally, the adiabatic regime means $T \to \infty$.

We arrive therefore at the outstanding result

$$
Q = \int_0^T dt\, I_z(t) = e \times \text{ integer} \tag{4.79}
$$

first proved by Thouless in 1983 [58]; it holds of course for any dimension d. Let me restate the theorem: if the Hamiltonian is changed adiabatically in such a way that it returns to its starting value in time T, the transported charge in an infinite periodic system is quantized provided that the system remains insulating at all times. A cycle pumps an integer number of elementary charges across the system.

Thouless quantization of charge transport [58], discussed above, also has profound relationships to several other topics: to Faraday laws of electrolysis (Sect. 4.6.1); to the topological explanation of the quantization of surface charge (discussed in Sect. 5.6.1); and to the modern theory of polarization (discussed in Sect. 5.3).

4.6 Charge transport in ionic liquids

Among the examples which realize a "Thouless pump", the original paper suggests a sliding charge-density wave. A more outstanding manifestation of quantized charge transport was pointed out shortly afterwards by Pendry and Hodges [82]: Faradays' laws of of electrolysis. The mass/charge transfer ratio shows that charge is always transported in units of e per ion, to the extent that electrolytic cells are used as standards of current. If a given ion sits at one electrode at $t = 0$, and if it drifts to the other one at $t = T$, the Hamiltonian can be considered as cyclic, whence charge quantization follows. However, at intermediate t values the charge "belonging" to a given ion is definitely non quantized, and arbitrarily defined: for a review of the possible definitions, see Refs. [83, 84].

4.6.1 Faraday law and oxidation numbers

Faraday's first law of electrolysis addresses insulating liquids in electrolytic cells, and can be recast in modern terms as follows. When a macroscopic number N of nuclei of a given chemical species passes from one electrode to the other, the transported electrical charge is an integer multiple of N times e. The law is additive: it concerns each different chemical species in the cell independently. It addresses ionic conduction in electrolytes, including molten salts; it does not apply to metallic conduction, which is unrelated to the motion of nuclei.

Faraday's discovery dates from 1832, decades before the existence of atoms became accepted, and even before Mendeleev's birthdate. Remarkably, Faraday wrote: "Although we know nothing of what an atom is, yet we cannot resist forming some idea of a small particle...". What he actually measured were charge-to-mass ratios; when scaled to the actual atomic masses, Faraday's "equivalent numbers" can be regarded as the archetypical definition of oxidation numbers in insulating liquids. We also stress that such definition is not static: Faraday's numbers are—within a modern quantum-mechanical view—dynamical properties in the adiabatic limit.

In modern chemistry, the oxidation number (also known as oxidation state) of an atom in a molecule or in a solid is an integer determined by an agreed set of rules,[85] and is often related in a postdictive [86]—*not* predictive—way to some measurable properties, thus facilitating the interpretation of several experimental observations [85, 86, 87, 88, 89]. In simple cases—when no d shells are involved—the oxidation numbers are simply determined by the octet rule: atoms are assigned an octet in order of decreasing electronegativity until all valence electrons are distributed [85], except for hydrogen which is obviously assigned a pair. The rule is unambiguous, given that the ordering of the (simple) elements as provided by different ionicity scales is the same. Oxidation numbers are formal quantitites not related to the wavefunction or to any quantum-mechanical operator.

Nowadays we know "what an atom is", but we also know that solids and liquids are not assemblies if ions; they are assemblies of atoms, having ionic character only because neighboring atoms have different electronegativity [42]. Literally dozens of electronegativity scales and/or definitions of atomic charges have been proposed in the literature [83], none of them yielding integer values.

The reason why integer oxidation numbers in electrolytes are measurable quantities is deeply rooted in topology and in Thouless' 1983 paper [58]. While this paper only addresses solid-state issues, the relevance of his result for understanding quantized charge transport in ionic liquids was emphasized shortly after in a very little quoted—and presumably also little known—paper by Pendry and Hodges [82]. Since then, a few authors have built upon their result and endorsed the topological nature of oxidation numbers in solids and liquids [87, 86, 89, 90]. The main message from topology is that the ionic charges behave as integer only when adiabatically

transported in a macroscopic system; no trace of quantization may appear in a "snapshot" of either a solid or a liquid at a given time. The quantized charge transport may occur either in a real experiment (as in Faraday's) or in a "gedanken" experiment, nowadays called a "Thouless pump" [13].

Charge transport in condensed matter has a close relationship to the theory of polarization (discussed in Sect. 5.3). In fact the current density (due to electrons and nuclei) which flows in a macroscopic insulator while the nuclei are adiabatically displaced equals by definition the time-derivative of macroscopic polarization. The formal definition of oxidation numbers proposed in Ref. [89] for insulating crystals and in Ref. [82, 87, 90] for insulating liquids is indeed closely related to polarization theory.

Faraday laws apply to ionic liquids where the oxidation numbers—upon chemical intuition—are quite robust in the given compound and do not depend on the fluctuating environment: this happens, for instance, when the octet rule is enough to determine the conventional oxidation numbers. There are challenging cases, though, well known in the solid state, like e.g. charge-ordering phenomena in transition metal oxides. In such cases a given element may assume different oxidation numbers depending on the ligands and on coordination. As an example, Mn can be labeled with up to seven oxidation numbers, from Mn(I) to Mn(VII), distinguished from their distinct spectroscopic and magnetic signatures [85]. It is easy to guess that when the (conventionally defined) oxidation numbers do depend on the environment the liquid is not insulating and topology cannot be invoked.

4.6.2 Ionic conductivity

An outstanding 2019 paper by Grasselli and Baroni [90] has extended the scope of topology in the context of charge transport in insulating liquids. Besides the amount of transported charge, another basic observable is the value of ionic conductivity at a given temperature, which is provided within classical statistical mechanics from the fluctuation-dissipation theorem, as discussed in Sec. C.5.1. The formula (called Green-Kubo formula) is based on the autocorrelation function of the equilibrium fluctuating charge current in absence of the electric field [91, 92]: see Eq. C.32.

Owing to the adiabatic approximation, the nuclear motion is classical even within ab-initio molecular dynamics, where the quantum nature of the electrons is fully accounted for. The contribution to the current from the motion of each nucleus at a given time is quite different from the one of a scalar integer charge—the topological oxidation number times e–moving on the same trajectory. According to Ref. [90], topology warrants that, when the real current is replaced with the fictitious current carried by the said integer charges (along the ab-initio trajectories), the Green-Kubo formula yields the correct ionic conductivity.

The topological nature of ionic conduction has an interesting implication. Let us

consider for instance liquid (undissociated) water under normal conditions: it is an insulator with an high value of the static dielectric constant ($\epsilon_0 \simeq 80$). Therefore the Faraday experiment would give a negative result, in agreement with the fact that the oxidation number of a neutral molecule as a whole is zero (both topologically and according to the rules): it is then kind of obvious that ionic conductivity must vanish. This fact is apparently less obvious when analyzed from the fluctuation-dissipation-theorem viewpoint. Water is strongly infrared-active [93], the Born charge tensor of the molecule as a whole is quite sizable [94], and the fluctuating equilibrium current is therefore conspicuously nonzero. Yet, given that the topological charge of a neutral undissociated molecule is zero, topology also guarantees that the time-integrated equilibrium autocorrelation function of the current vanishes, as indeed needed on physics considerations.

Chapter 5

Macroscopic polarization

The macroscopic polarization $\mathbf{P}$ is a fundamental concept that all undergraduates learn about in elementary courses [95, 96]. In view of this, it is truly extraordinary that until the early 1990s there was no generally accepted formula for $\mathbf{P}$ in condensed matter, even as a matter of principles. $\mathbf{P}$ is an intensive vector quantities that intuitively carries the meaning of dipole per unit volume. Most textbooks [73, 97] provide a flawed definition of $\mathbf{P}$, not implementable in practical computations [98].

A genuine change of paradigm was initiated by a couple of important papers [99, 100], after which the major development was introduced by King-Smith and Vanderbilt in 1992 (paper published in 1993 [101]). Other important advances occurred during the 1990s [102, 103] and the so-called "modern theory of polarization"it is by now at a very mature stage; several reviews have appeared in the literature over the years [1, 2, 3, 5, 6, 7, 8, 20, 24].

Aiming at a computational physics readership, it is worth emphasizing that most ab-initio electronic-structure codes on the market, for dealing with either crystalline or noncrystalline materials, implement the modern theory of polarization as a standard option. A nonexaustive list includes ABINIT [104], CRYSTAL [105], QUANTUM-ESPRESSO [106], SIESTA [107], VASP [108], and CPMD [109]; the code tutorials often include an outline of polarization theory. Its implementations have been instrumental since almost three decades in the study of ferroelectric and piezoelectric materials [110, 111, 112].

The basic concepts of the modern theory of polarization also start reaching a few textbooks [113, 114], though very slowly; most of them are still plagued with erroneous concepts and statements.

5.1   Polarization and electric field

The modern theory of polarization, at least in its original form, only addresses the polarization $\mathbf{P}$ in a null macroscopic $\mathbf{E}$ field; it must be realized that, insofar as we

address an infinite system with no boundaries, the $\mathbf{E}$ field is quite arbitrary. The microscopic charge density is neutral in average and lattice periodical; the value of $\mathbf{E}$ is just an arbitrary boundary condition for the integration of Poisson's equation. The usual choice (performed within all electronic-structure codes) is to impose a lattice-periodical Coulomb potential, i.e. $\mathbf{E} = 0$. Imposing a given nonzero value of $\mathbf{E}$ is equally legitimate (in insulators), although technically more difficult [115, 116]).

When addressing a finite sample with boundaries, the $\mathbf{E}$ field is in principle measurable inside the material, without reference to what happens at the sample boundary; this is not the case of $\mathbf{D}$. In fact, $\mathbf{E}$ obtains by averaging over a macroscopic length scale the microscopic electric field $\mathbf{E}^{(\mathrm{micro})}(\mathbf{r})$, which fluctuates at the atomic scale [96]. In a macroscopically homogeneous system the macroscopic field $\mathbf{E}$ is constant, and in crystalline materials it coincides with the cell average of $\mathbf{E}^{(\mathrm{micro})}(\mathbf{r})$. A lattice-periodical potential enforces $\mathbf{E} = 0$; for a supercell calculation, this applies to the field average over the supercell, while in different regions there can be a nonzero macroscopic field.

As explained so far, there is no need of addressing finite samples and external vs. internal fields from a theoretician's viewpoint. Nonetheless a brief digression is in order, given that experiments *are* performed over finite samples, often in external fields. Suppose a finite macroscopic sample is inserted in a constant external field $\mathbf{E}^{(\mathrm{ext})}$: the microscopic field $\mathbf{E}^{(\mathrm{micro})}(\mathbf{r})$ coincides with $\mathbf{E}^{(\mathrm{ext})}$ far away from the sample, while it is different inside because of screening effects. If we choose an homogeneous sample of *ellipsoidal shape*, then the macroscopic average of $\mathbf{E}^{(\mathrm{micro})}(\mathbf{r})$, i.e. the macroscopic screened field $\mathbf{E}$, is constant in the bulk of the sample. The shape effects are embedded in the depolarization coefficients [95]: the simplest case is the extremely oblate ellipsoid, i.e. a slab of a macroscopically homogeneous dielectric; more details are given in Ref. [8]. For the slab geometry in a vanishing external field $\mathbf{E}^{(\mathrm{ext})}$ the internal field $\mathbf{E}$ vanishes when $\mathbf{P}$ is parallel to the slab (transverse polarization), while $\mathbf{E} = -4\pi\mathbf{P}$ is the depolarization field when $\mathbf{P}$ is normal to the slab (longitudinal polarization): see Fig. 5.1.

5.2 Polarization differences

Novel ideas about macroscopic polarization emerged in the early 1990s [99, 100]; these led to the modern theory, based on a Berry phase, which was founded by King-Smith and Vanderbilt soon afterwards [101]. At its foundation, the modern theory was limited to a crystalline system in an independent-electron framework (either KS or Hartree-Fock). Later, the theory was extended to correlated and/or disordered systems [102, 103].

The historical development of the theory passed through abandoning the concept of polarization "itself", addressing instead a polarization *difference*, which could be
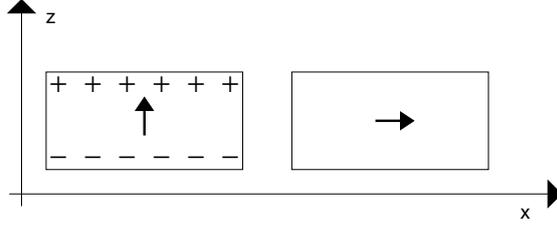
Figure 5.1: Macroscopic polarization $\mathbf{P}$ in a slab normal to z, for a vanishing external field $\mathbf{E}^{(\mathrm{ext})}$. Left: When $\mathbf{P}$ is normal to the slab, a depolarizing field $\mathbf{E} = -4\pi\mathbf{P}$ is present inside the slab, and charges at its surface, with areal density $\sigma_{\mathrm{surface}} = \mathbf{P} \cdot \mathbf{n}$ Right: When $\mathbf{P}$ is parallel to the slab, no depolarizing field and no surface charge is present.

expressed as a time-integrated adiabatic current.[100, 101] Only afterwards it was realized [117] that even polarization itself can be defined, although by means of a change of paradigm: bulk polarization is not a vector (as theretofore assumed), it is a lattice. In these Notes I will follow the historical developments, presenting polarization differences first, and polarization itself later, in Sect. 5.4.

The first calculation ever of spontaneous polarization was published in 1990 [99]. The case study was BeO: it has the simplest structure where inversion symmetry is absent (i.e. wurtzite), and furthermore its constituents are first-row atoms. The idea was to address the macroscopic polarization of a slab of finite thickness, with faces normal to the c axis, embedding it in an *ad hoc* medium which (i) has no bulk polarization for symmetry reasons, and (2) does not produce any geometrical or chemical perturbation at the interface. The optimal choice is a fictitious BeO in the zincblende structure. Because of obvious reasons, the system is periodically replicated in a supercell geometry (Fig. 5.2, top panel). The selfconsistent calculation shows well localized interface charges, of opposite sign and equal magnitudes at the two nonequivalent interfaces (Fig. 5.2, bottom panel). The interface charge is related to the difference in polarization between the two materials: $\sigma_{\mathrm{interface}} = \Delta\mathbf{P} \cdot \mathbf{n}$. The computer experiment provides the value of $\sigma_{\mathrm{interface}}$, and since $\mathbf{P}$ vanishes by symmetry in the zincblende slab, one thus obtains the bulk value of $\mathbf{P}$ in the wurtzite material. Notice that here $\mathbf{P}$ is a *longitudinal* polarization, in a depolarizing field.

It must be emphasized that the quantity really "measured" in this computer experiment is $\Delta\mathbf{P}$, *not* the polarization $\mathbf{P}$ itself. After Ref [99] was published, a study of the experimental literature showed that—contrary to an incorrect widespread belief—no experimental value of $\mathbf{P}$ in any wurtzite material exists: only estimates are available. Ref. [99] marks, as said above, a change of paradigm: polarization must be *defined* by means of *differences*, and the concept of polarization "itself"
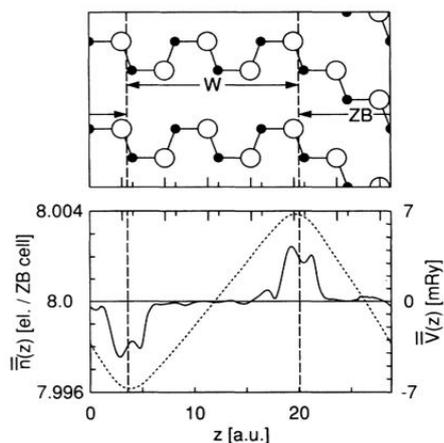
Figure 5.2: Top panel: The 14-atom BeO supercell in a vertical plane through the BeO bonds; the wurtzite (W) and zincblende (ZB) stackings are perspicuous. Bottom panel: Macroscopic averages of the valence electron density (solid) and of the electrostatic potential (dotted).
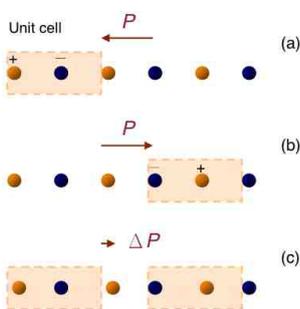
Figure 5.3: A 1d solid with infinite length. Different choices of the unit cell give different **P** values: (a), (b). On the other hand, the change of polarization Δ**P** does not depend on the choice of the unit cell (c).

must be abandoned. With hindsight, it is nowadays pretty clear that the problem, as well as its solution, exists already at the classical level: this is sketched in Fig. 5.3. Most textbooks are missing this very basic fact.

The modern theory avoids addressing the "absolute" polarization of a given equilibrium state, quite in agreement with the experiments, which invariably measure polarization *differences*. Instead, the theory addresses differences in polarization between two states of the material that can be connected by an adiabatic switching process. The time-dependent Hamiltonian is assumed to remain insulating at all times, and the polarization difference is then *defined* [100] as the time-integrated transient macroscopic current that flows through the insulating sample during the switching process:

$$\Delta \mathbf{P} = \mathbf{P}(\Delta t) - \mathbf{P}(0) = \int_0^{\Delta t} dt \, \mathbf{j}(t). \tag{5.1}$$

In the adiabatic limit $\Delta t \to \infty$ and $\mathbf{j}(t) \to 0$, while $\Delta \mathbf{P}$ stays finite. Addressing currents (instead of charges) explains the occurrence of *phases* of the wavefunctions (instead of square moduli) in the modern theory. Eventually the time integration in Eq. (5.1) will be eliminated, leading to a two-point formula involving only the initial and final states.

5.3 Independent electrons

5.3.1 The King-Smith and Vanderbilt formula

For a crystalline system of independent electrons the expression for the transient current occurring in Eq. (5.1) is precisely the same as previously derived for quantum transport, Eq. (4.77). Therefore for one band (index 0) and single occupancy in dimension d the electronic contribution to the polarization difference is

$$\Delta\mathbf{P} = -\frac{ie}{(2\pi)^d} \int_0^{\Delta t} dt \int_{\text{BZ}} d\mathbf{k} \, (\, \langle \dot{u}_{0\mathbf{k}} | \nabla_{\mathbf{k}} u_{0\mathbf{k}} \rangle - \langle \nabla_{\mathbf{k}} u_{0\mathbf{k}} | \dot{u}_{0\mathbf{k}} \rangle \,); \qquad (5.2)$$

the (classical) nuclear contribution must be added separately. We remind that crystal-cell neutrality is essential. Notice also that, given the occurrence of Bloch states, the Hamiltonian is lattice periodical at all t: this implicitly means that in Eq. (5.2) $\Delta\mathbf{P}$ is evaluated at $\mathbf{E} = 0$.

It is now expedient to introduce a dimensionless adiabatic time λ, with $|u_{j\mathbf{k}}(t)\rangle = |u_{j\mathbf{k}}(\lambda(t))\rangle$, $\lambda(0) = 0$, $\lambda(1) = \Delta t$. Eq. (5.2) becomes then, for n doubly occupied bands (index $j = 1, n$) in $3d$:

$$\begin{aligned}
\Delta\mathbf{P} &= \mathbf{P}(1) - \mathbf{P}(0) = \int_0^1 d\lambda \, \partial_\lambda \mathbf{P}(\lambda) \\
\partial_\lambda \mathbf{P} &= -\frac{2ie}{(2\pi)^3} \sum_{j=1}^n \int_{\text{BZ}} d\mathbf{k} \, (\, \langle \partial_\lambda u_{j\mathbf{k}} | \nabla_{\mathbf{k}} u_{j\mathbf{k}} \rangle - \langle \nabla_{\mathbf{k}} u_{j\mathbf{k}} | \partial_\lambda u_{j\mathbf{k}} \rangle \,).
\end{aligned} \qquad (5.3)$$

It is essential that the gap does not close, i.e. the system remain insulating, for all λ values.

The expression in Eq. (5.3) can be integrated with respect to λ to obtain

$$\mathbf{P}(\lambda) = -\frac{2ie}{(2\pi)^3} \sum_{j=1}^n \int_{\text{BZ}} d\mathbf{k} \, \langle u_{j\mathbf{k}} | \nabla_{\mathbf{k}} u_{j\mathbf{k}} \rangle : \qquad (5.4)$$

this is the (by now famous) King-Smith and Vanderbilt formula [101], yielding the polarization of the final state minus the polarization of the initial state, Eq. (5.3). To understand the meaning of the $\mathbf{k}$ integral in $3d$ we take the simple example of a simple cubic lattice of constant a, similarly to Eq. (4.78):

$$P_z(\lambda) = -\frac{2e}{(2\pi)^3} \sum_{j=1}^n \int_{-\pi/a}^{\pi/a} dk_x \int_{-\pi/a}^{\pi/a} dk_y \left[i \int_{-\pi/a}^{\pi/a} dk_z \langle u_{j\mathbf{k}} | \partial_{k_z} u_{j\mathbf{k}} \rangle \right], \qquad (5.5)$$

where the square parenthesis highlights the Berry-phase, to be compared to Eq. (3.37).

73

In first-principle implementations, the Berry phase is discretized as in Eq. (3.55), and the remaining $2d$ **k**-integral is discretized in the trivial way. The first calculation ever of the "spontaneous" polarization of a ferroelectric material ($KNbO_3$) appeared in 1993 [118], and agreed within 10% with the measured values. As said above, the Berry-phase formula is nowadays implemented in most first-principle codes.

5.3.2 The quantum of polarization

Given that every phase is defined modulo 2π, all of the two-point formulæ for $\Delta\mathbf{P}$ in terms of Berry phases are arbitrary modulo a polarization "quantum". This is the tradeoff one has to pay when switching from the curvature formula, Eq. (5.3)—where no such arbitrariness exists—to the two-point King-Smith and Vanderbilt formula, where only the connection occurs. The actual arbitrariness of $\Delta\mathbf{P}$ in $3d$ is $2e\mathbf{R}/V_{\text{cell}}$, where $\mathbf{R}$ is a lattice vector and V_{cell} is the cell volume (the 2 factor owes to double band occupancy). A similar arbitrariness of an integer times $e\mathbf{R}/V_{\text{cell}}$ occurs for the classical nuclear contribution to polarization.

The quantum arbitrariness is rarely a problem in practice. In most cases, the change in **P** that can be induced by a perturbation, such as a small sublattice displacement, is insufficient to cause **P** to change by a large fraction of the quantum. Where exceptions exist the ambiguity can be resolved by subdividing the adiabatic path into several shorter intervals, for each of which the change in **P** is unambiguous for practical purposes. Additional problems may occur in the discretized version of the Berry-phase formula; this is discussed e.g. in Ref. [8] and, in more detail, in Ref. [7].

Here we stress that the quantum ambiguity is an essential aspect of the theory. For example, for the case of a *closed cyclic* adiabatic evolution of the system, in which the parameter values $\lambda = 0$ and $\lambda = 1$ label the *same physical state* of the system, we retrieve the quantization of charge transport, discussed in Sect. 4.5, and governed by a Chern number.

5.3.3 Wannier functions

The KS (or Hartee-Fock) ground state is a Slater determinant of doubly occupied orbitals; any unitary transformation of the occupied states among themselves leaves the determinantal wavefunction invariant (apart for an irrelevant phase factor), and hence it leaves invariant any KS ground-state property.

For an insulating crystal the Bloch KS orbitals of completely occupied bands can be transformed to localized Wannier orbitals (or functions) WFs. This is known since 1937 [119], but for many years the WFs have been mostly used as a formal tool; they became a popular topic in computational electronic structure only after the 1997 seminal work of Marzari and Vanderbilt [56]. A comprehensive review

appeared as Ref. [19], and a public-domain implementation is in WANNIER90 [120]. If the crystal is metallic, the WFs can still be technically useful [121], but it must be emphasized that the ground state *cannot* be written as a Slater determinant of localized orbitals of any kind, as a matter of principle.

The transformation of the Berry phase formula in terms of WFs provides an alternative, and perhaps more intuitive, viewpoint. The formal transformation was known since the 1950s [122], although the physical meaning of the formalism was not understood until the advent of the modern theory of polarization.

The unitary transformation which defines the WF $w_{j\mathbf{R}}(\mathbf{r})$, labeled by band j and unit cell $\mathbf{R}$, within our normalization is

$$|w_{j\mathbf{R}}\rangle = \frac{V_{\text{cell}}}{(2\pi)^3} \int_{\text{BZ}} d\mathbf{k}\, e^{i\mathbf{k}\cdot\mathbf{R}} |\psi_{j\mathbf{k}}\rangle . \tag{5.6}$$

Any WF centered in $\mathbf{R}$ is just the translate of a central-cell WF, i.e.:

$$\langle\mathbf{r}|w_{j\mathbf{R}}\rangle = \langle\mathbf{r} - \mathbf{R}|w_{j\mathbf{0}}\rangle, \quad |w_{j\mathbf{0}}\rangle = \frac{V_{\text{cell}}}{(2\pi)^3} \int_{\text{BZ}} d\mathbf{k}\, |\psi_{j\mathbf{k}}\rangle. \tag{5.7}$$

The ground-state projector $\mathcal{P}$ is clearly invariant by any unitary transformation of the occupied orbitals among themselves; in terms of WFs it is written as:

$$\mathcal{P} = \frac{V_{\text{cell}}}{(2\pi)^3} \sum_{j=1}^{n} \int_{\text{BZ}} d\mathbf{k}\, |\psi_{j\mathbf{k}}\rangle\langle\psi_{j\mathbf{k}}| = \sum_{j\mathbf{R}} |w_{j\mathbf{R}}\rangle\langle w_{j\mathbf{R}}|. \tag{5.8}$$

In several circumstances, Eq. (5.8) is the most straightforward "bridge" between PBCs and OBCs; localized orbitals have a long history within quantum chemistry.

If one then defines the "Wannier centers" as $\mathbf{r}_{j\mathbf{R}} = \langle w_{j\mathbf{R}}|\mathbf{r}|w_{j\mathbf{R}}\rangle$, it is rather straightforward to prove that Eq. (5.4) is equivalent to

$$\mathbf{P}^{(\text{el})}(\lambda) = -\frac{2e}{V_{\text{cell}}} \sum_{j=1}^{n} \mathbf{r}_{j\mathbf{0}}. \tag{5.9}$$

This means that the electronic term in the macroscopic polarization $\mathbf{P}$ is (twice) the dipole of the Wannier charge distributions in the central cell, divided by the cell volume. The nuclear term is obviously similar in form to Eq. (5.9); the sum of both terms is charge neutral.

WFs are severely gauge-dependent, since the phases of the $|\psi_{j\mathbf{k}}\rangle$ appearing in Eq. (5.6) can be chosen arbitrarily. However, their centers are gauge-invariant modulo a lattice vector. Therefore $\mathbf{P}^{(\text{el})}$ in Eq. (5.9) is affected by the same "quantum" indeterminacy discussed above.

The modern theory, when formulated in terms of WFs, becomes much more intuitive, and in a sense vindicates the venerable Clausius-Mossotti viewpoint [123]:

in fact, the charge distribution is partitioned into localized contributions, each providing an electric dipole, and these dipoles yield the electronic term in $\mathbf{P}$. However, it is clear from Eq. (5.6) that the *phase* of the Bloch orbitals is essential to arrive at the right partitioning. Any decomposition based on charge only is severely nonunique and does not provide in general the right $\mathbf{P}$, with the notable exception of the extreme case of molecular crystals.

In the latter case, in fact, we may consider the set of WFs centered on a given molecule; their total charge distribution coincides—in the weakly interacting limit—with the electron density of the isolated molecule (possibly in a local field). This justifies the elementary Clausius-Mossotti viewpoint. It is worth mentioning that the dipole of a polar molecule is routinely computed in a supercell geometry via the single-point Berry phase discussed below [94]. The dipole value coincides with the one computed in the trivial way in the large supercell limit. Finite-size corrections, due to the local field (different in the two cases), can also be applied [124].

The case of alkali halides—where the model is often phenomenologically used— deserves a different comment [8]. The electron densities of isolated ions (with or without fields) are quite different from the corresponding WFs charge distributions, for instance because of orthogonality constraints: hence the Clausius-Mossotti model is *not* justified in its elementary form, despite contrary statements in the literature. For a detailed analysis, see Ref. [125].

5.4 Polarization itself

As said above, it was realized in 1993 by Vanderbilt and King-Smith [117] that even polarization itself can be defined, although by means of a change of paradigm: bulk polarization is not a vector (as theretofore assumed), it is a lattice. A counterintuitive corollary is that the polarization $\mathbf{P}$ of an inversion-symmetric crystal is not necessarily zero. Since inversion symmetry requires $\mathbf{P} = -\mathbf{P}$, the lattice must be symmetric: this may happen even if $\mathbf{P} = 0$ does not belong to the lattice. For a macroscopic bounded crystallite, the lattice ambiguity is fixed only after the sample termination is chosen [24, 126]. I am going to present a novel derivation, which only appeared in 2021 [127]; it is very general, beyond the independent-electron framework.

We start observing that the dipole of a bounded and charge-neutral sample is a very trivial quantity; the macroscopic polarization of a crystalline solid, instead, has been a challenging problem for many years. In quantum mechanics the dipole of a bounded sample is the expectation value of the position operator $\mathbf{r}$. The drawback is that solid state physics requires Born-von-Kàrmàn periodic boundary conditions (PBCs) [97], which define the Hilbert space where Schrödinger equation is solved. Unfortunately the multiplicative operator $\mathbf{r}$ is *not* a legitimate operator in the PBC

Hilbert space: it maps a state vector within the space into an entity which does not belong to the same space. Here we solve the problem by alternatively expressing the dipole of a bounded sample in an unconventional way, with no reference to the **r** operator; in this way the crystalline expression follows somewhat naturally.

5.4.1   Polarization of a bounded crystallite

We assume that N electrons are confined in a macroscopic sample of volume $\mathcal{V}$, together with a neutralizing background of point-like classical nuclei. Let $|\tilde{\Psi}_0\rangle$ be the singlet insulating ground eigenstate; the many-body wavefunction is square-integrable over $\mathbb{R}^{3N}$ and vanishes far away from the sample. If the system is macroscopically homogeneous, the electronic term in polarization has the pretty trivial expression

$$\mathbf{P}^{(\text{el})} = -\frac{e}{\mathcal{V}}\langle\tilde{\Psi}_0|\,\hat{\mathbf{r}}\,|\tilde{\Psi}_0\rangle, \qquad \hat{\mathbf{r}} = \sum_{i=1}^{N}\mathbf{r}_i; \tag{5.10}$$

the nuclear classical contribution has to be added in order to obtain a meaningful observable.

It is expedient to address the family of many-body Hamiltonians parametrized by the parameter $\boldsymbol{\kappa}$:

$$\hat{H}_{\boldsymbol{\kappa}} = \frac{1}{2m}\sum_{i=1}^{N}\left(\mathbf{p}_i + \hbar\boldsymbol{\kappa}\right)^2 + \hat{V}, \tag{5.11}$$

where $\hat{V}$ includes one-body and two-body potentials, and whose ground eigenstate is $|\tilde{\Psi}_{0\boldsymbol{\kappa}}\rangle$. In order to simplify notations we will set $\hat{H}_0 \equiv \hat{H}$ and $|\tilde{\Psi}_{n0}\rangle \equiv |\tilde{\Psi}_n\rangle$. The vector $\boldsymbol{\kappa}$, having the dimensions of an inverse length, generalizes the Hamiltonian by including a constant vector potential: it is therefore a pure gauge. The gauge-transformed eigenstates are

$$|\tilde{\Psi}_{n\boldsymbol{\kappa}}\rangle = \mathrm{e}^{-i\boldsymbol{\kappa}\cdot\hat{\mathbf{r}}}|\tilde{\Psi}_n\rangle. \tag{5.12}$$

We pause at this point to stress an important semantical issue. The choice of the (arbitrary) $\boldsymbol{\kappa}$ value in Eq. (5.11) fixes the gauge in the Hamiltonian. Once this fixed, there is an additional freedom in choosing the arbitrary phase factor in front of each eigenstate: even this second choice goes under the name of gauge choice. The expression of any physical observable must be gauge-invariant in both senses. I therefore alert the reader that, in the following, it is essential to realize in which context the term "gauge" is used.

By taking the $\boldsymbol{\kappa}$ derivative of Eq. (5.12) one transforms Eq. (5.10) into

$$\mathbf{P}^{(\text{el})} = -\frac{ie}{\mathcal{V}}\langle\tilde{\Psi}_{0\boldsymbol{\kappa}}|\partial_{\boldsymbol{\kappa}}\tilde{\Psi}_{0\boldsymbol{\kappa}}\rangle, \tag{5.13}$$

at any $\boldsymbol{\kappa}$-value; in view of the subsequent developments, we set $\boldsymbol{\kappa} = 0$ in the following:

$$\mathbf{P}^{(\mathrm{el})} = -\frac{ie}{\mathcal{V}} \langle \tilde{\Psi}_0 | \partial_{\boldsymbol{\kappa}} \tilde{\Psi}_0 \rangle. \tag{5.14}$$

From the arguments in Ch. 3 it should be clear that $i \langle \tilde{\Psi}_{0\boldsymbol{\kappa}} | \partial_{\boldsymbol{\kappa}} \tilde{\Psi}_{0\boldsymbol{\kappa}} \rangle$ is a Berry connection evaluated at $\boldsymbol{\kappa} = 0$. Since Eq. (5.11) at two different $\boldsymbol{\kappa}$'s yields two different Hamiltonians, an equally acceptable gauge-transformed eigenstate would be

$$|\tilde{\Psi}_{n\boldsymbol{\kappa}}\rangle = \mathrm{e}^{i\phi(\boldsymbol{\kappa})} \mathrm{e}^{-i\boldsymbol{\kappa}\cdot\hat{\mathbf{r}}} |\tilde{\Psi}_n\rangle, \tag{5.15}$$

with an arbitrary $\phi(\boldsymbol{\kappa})$. The physical observable obtains from Eq. (5.14) when the gauge of Eq. (5.12) is enforced; it is not allowed to adopt therein the most general gauge of Eq. (5.15). In fact, a Berry connection by itself *cannot* define a physical observable. In the present case Eq. (5.14) acquires its physical meaning only after the above specific gauge fixing. I stress that here is the conceptual novelty of the present work: a definition of the dipole of a bounded sample where no use is made of the position operator $\mathbf{r}$. The same definition and the same gauge fixing—Eq. (5.18) below—can be exported to the PBC crystalline case.

5.4.2 Unbounded crystal

We adopt the same Hamiltonian as in Eq. (5.11), but now within the PBC Hilbert space: the many-body wavefunction is periodic in the cubic "supercell" of side L in each electronic variable independently, and normalized to one therein. Each Cartesian coordinate is then equivalent to the angle $\varphi_i = 2\pi x_i/L$, and analogously for y_i and z_i. The potential $\hat{V}$ enjoys the same periodicity: this means that the macroscopic field $\boldsymbol{\mathcal{E}}$ inside the sample vanishes. We will indicate the eigenstates as $|\Psi_{n\boldsymbol{\kappa}}\rangle$ without a tilde, in order to distinguish them from those of the bounded crystallite; as stressed above, the multiplicative $\hat{\mathbf{r}}$ operator is "forbidden" in the PBC Hilbert space [103].

In order to address polarization, we need to ensure beforehand that the ground state is insulating. The many-body velocity operator is

$$\hat{\mathbf{v}}_{\boldsymbol{\kappa}} = \frac{1}{m} \sum_{i=1}^{N} (\mathbf{p}_i + \hbar\boldsymbol{\kappa}) = \frac{1}{\hbar} \partial_{\boldsymbol{\kappa}} \hat{H}_{\boldsymbol{\kappa}}, \tag{5.16}$$

hence by Hellmann-Feynman theorem the macroscopic current density is

$$\mathbf{j}_{\boldsymbol{\kappa}} = -\frac{e}{\hbar L^3} \langle \Psi_{0\boldsymbol{\kappa}} | \partial_{\boldsymbol{\kappa}} \hat{H}_{\boldsymbol{\kappa}} | \Psi_{0\boldsymbol{\kappa}} \rangle = -\frac{e}{\hbar L^3} \partial_{\boldsymbol{\kappa}} E_{0\boldsymbol{\kappa}} , \tag{5.17}$$

where $E_{0\boldsymbol{\kappa}}$ is the ground-state energy. Given that an insulator does not sustain a dc current, the ground-state energy is $\boldsymbol{\kappa}$-independent (the opposite is true in metals).

The Hamiltonian of Eq. (5.11) was first introduced in 1964 in a milestone paper by W. Kohn, who noticed that PBCs violate gauge-invariance in the conventional sense [128]. If we try the same transformation as in Eq. (5.12), the quantity $e^{-i\boldsymbol{\kappa}\cdot\hat{\mathbf{r}}}|\Psi_0\rangle$ is a solution of Schrödinger equation with energy $\mathcal{E}_0$, but it does not obey PBCs and therefore does not belong to the Hilbert space. At an arbitrary $\boldsymbol{\kappa}$, the genuine PBC eigenstates $|\Psi_{n\boldsymbol{\kappa}}\rangle$ have a nontrivial $\boldsymbol{\kappa}$-dependence. There is, however, a discrete set of special $\boldsymbol{\kappa}$ vectors for which

$$|\Psi_{0\boldsymbol{\kappa}}\rangle = e^{-i\boldsymbol{\kappa}\cdot\hat{\mathbf{r}}}|\Psi_0\rangle \tag{5.18}$$

obeys PBCs and yields therefore the ground eigenstate of $\hat{H}_{\boldsymbol{\kappa}}$: $\boldsymbol{\kappa} = \frac{2\pi}{L}(\ell, m, n)$, with integer (ℓ, m, n).

In order to define polarization, we proceed by adopting the analogue of Eq. (5.14), and in the analogous gauge. We start from the identity

$$\partial_{\boldsymbol{\kappa}}\ln\langle\Psi_0|\Psi_{0\boldsymbol{\kappa}}\rangle = \frac{\langle\Psi_0|\partial_{\boldsymbol{\kappa}}\Psi_{0\boldsymbol{\kappa}}\rangle}{\langle\Psi_0|\Psi_{0\boldsymbol{\kappa}}\rangle}, \tag{5.19}$$

where $|\Psi_{0\boldsymbol{\kappa}}\rangle$ is the PBC ground eigenstate of $\hat{H}_{\boldsymbol{\kappa}}$; since $\langle\Psi_0|\partial_{\boldsymbol{\kappa}}\Psi_0\rangle$ is purely imaginary, a leading-order expansion in $\boldsymbol{\kappa}$ yields

$$i\langle\Psi_0|\partial_{\boldsymbol{\kappa}}\Psi_0\rangle\cdot\boldsymbol{\kappa} \simeq -\mathrm{Im}\,\ln\langle\Psi_0|\Psi_{0\boldsymbol{\kappa}}\rangle. \tag{5.20}$$

We pause to observe that multivaluedness debuts here. In fact Eq. (5.20) relates two phase angles: a differential angle on the left, and a finite angle difference on the right. While a differential angle is single valued, a finite angle is defined modulo 2π; upon replacing the former with the latter we are going to define a multivalued observable. We stress once more that multivaluedness is not a mathematical artifact; it is a necessary feature of polarization within PBCs [24].

Next we pick a vector $\boldsymbol{\kappa}_1$ in the special set: $\boldsymbol{\kappa}_1 = \frac{2\pi}{L}(1, 0, 0)$, and we replace the derivative in Eq. (5.14) with a finite difference, in the large-sample limit:

$$P_x^{(\mathrm{el})} = \frac{e}{2\pi L^2}\mathrm{Im}\,\ln\langle\Psi_0|\Psi_{0\boldsymbol{\kappa}_1}\rangle. \tag{5.21}$$

As it stands, Eq. (9.5) is gauge-dependent and cannot express an observable: it is in fact a discretized Berry connection. Eq. (9.5) only acquires physical meaning when we fix the gauge by adopting the one of Eqs. (5.12) and (5.18) (with no extra phase factor):

$$
\begin{aligned}
P_x^{(\mathrm{el})} &= \frac{e}{2\pi L^2}\mathrm{Im}\,\ln\langle\Psi_0|\,e^{-i\boldsymbol{\kappa}_1\cdot\hat{\mathbf{r}}}\,|\Psi_0\rangle \\
&= \frac{e}{2\pi L^2}\mathrm{Im}\,\ln\langle\Psi_0|\,e^{-i\frac{2\pi}{L}\sum x_i}\,|\Psi_0\rangle.
\end{aligned} \tag{5.22}
$$

We have thus arrived at the main message of the present work: the bounded-crystallite formula, Eq. (5.14), and the crystalline formula, Eq. (5.22), are essentially the same formula, within the same gauge, in two different frameworks.

The replacement of $|\Psi_{0\boldsymbol{\kappa}_1}\rangle$ in Eq. (9.5) with $\mathrm{e}^{-i\boldsymbol{\kappa}_1\cdot\hat{\mathbf{r}}}|\Psi_0\rangle$ in Eq. (5.22) is allowed in insulators only. We remind that $|\Psi_{0\boldsymbol{\kappa}}\rangle$ obtains by following the ground state $|\Psi_0\rangle$ when the $\boldsymbol{\kappa}$ vector in $\hat{H}_{\boldsymbol{\kappa}}$ is adiabatically turned on; in the metallic case—as shown by Kohn [128]—the energy $E_{0\boldsymbol{\kappa}}$ of such state *does* depend on $\boldsymbol{\kappa}$, and therefore $|\Psi_{0\boldsymbol{\kappa}_1}\rangle$ is orthogonal to $\mathrm{e}^{-i\boldsymbol{\kappa}_1\cdot\hat{\mathbf{r}}}|\Psi_0\rangle$. We have shown above that in the insulating case the state $|\Psi_{0\boldsymbol{\kappa}_1}\rangle$ has instead the same energy as $\mathrm{e}^{-i\boldsymbol{\kappa}_1\cdot\hat{\mathbf{r}}}|\Psi_0\rangle$, and therefore the two states may be identified.

The well known Eq. (5.22), sometimes dubbed "single-point Berry phase", was originally obtained in Ref [103] by considering a many-body Hamiltonian which is adiabatically varied in time, and showing that the time derivative of Eq. (5.22) coincides with the macroscopic current density $j_x^{(\mathrm{el})}(t)$ which flows through the insulating sample. Here I have derived the same result via a different logic: polarization itself obtains without addressing currents at all, starting instead from an unconventional definition of the dipole of a bounded sample.

Finally, the nuclear term in polarization can be added to Eq. (5.22) in a very compact form. If the nuclei of charge Z_ℓ sit at sites $\mathbf{R}_\ell$ in the supercell, the expression is

$$P_x = \frac{e}{2\pi L^2}\mathrm{Im}\,\ln\,\langle\Psi_0|\,\mathrm{e}^{i\frac{2\pi}{L}(\sum_\ell Z_\ell X_\ell - \sum_i x_i)}\,|\Psi_0\rangle, \tag{5.23}$$

where $X_\ell = R_{\ell,x}$. Owing to charge neutrality, polarization is invariant by translation of the coordinate origin (as it must be). It is argued that Eq. (5.23) also holds when the quantum nature of the nuclei is considered.

5.5 Polarization as a multivalued observable

Bulk polarization is a lattice, not a vector, and in fact the main entry of Eqs. (5.22) and (5.23) is the multivalued function "Im ln". But it is also clear that for a three-dimensional system these equations cannot be accepted as they stand in the large-sample limit: the prefactor goes in fact to zero. We have not exploited crystalline symmetry yet.

By definition, whenever a material is crystalline, a uniquely defined lattice can be associated with the real sample. The lattice is a "mathematical construction" [97], uniquely defined—by means of an appropriate average—even in cases with correlation, finite temperature, quantum nuclei, chemical disorder (i.e. crystalline alloys, a.k.a. solid solutions), where the actual wavefunction may require a supercell (multiple of the primitive lattice cell).

We consider—without loss of generality—a simple cubic lattice of constant a,

where the supercell side L is an integer multiple of a: $L = Ma$. Suppose the potential $\hat{V}$ in the Hamiltonian is adiabatically varied in time; we define the phase angle

$$\gamma_x(t) = \text{Im ln}\, \langle \Psi_0(t) |\, e^{i\frac{2\pi}{L}(\sum_\ell Z_\ell X_\ell - \sum_i x_i)}\, |\Psi_0(t)\rangle, \qquad (5.24)$$

where $|\Psi_0(t)\rangle$ is the adiabatic ground eigenstate. The current flowing across a section of area L^2 normal to x is

$$I_x(t) = L^2 \dot{P}_x(t) = \frac{e}{2\pi}\dot{\gamma}_x(t). \qquad (5.25)$$

Owing to cristalline periodicity, The current $I_x(t)$ is the sum of M^2 identical currents, each flowing through a microscopic section of area a^2; one can therefore define a reduced crystalline phase angle $\gamma_x^{(\text{crystal})}$ such that $\dot{\gamma}_x(t) = M^2\dot{\gamma}_x^{(\text{crystal})}(t)$. The crystalline polarization is thus expressed in terms of $\gamma_x^{(\text{crystal})}$ as

$$P_x = \frac{e}{2\pi a^2}\gamma_x^{(\text{crystal})}; \qquad (5.26)$$

the case of independent electrons is presented in detail in the next Section.

A generic lattice is dealt with by means of a coordinate transformation [1]; the bulk value of $\mathbf{P}$ is then ambiguous modulo $e\mathbf{R}/\mathcal{V}_{\text{cell}}$, where $\mathbf{R}$ is a lattice vector and $\mathcal{V}_{\text{cell}}$ is the volume of a primitive cell. The quantity $e\mathbf{R}/\mathcal{V}_{\text{cell}}$ goes under the name of polarization "quantum". By definition a primitive cell is a minimum-volume one [97]: this choice is mandatory in order to make $\mathbf{P}$ an unambiguously defined multivalued observable. Finally we observe that the modulo ambiguity is only removed when the termination of the bounded sample is specified; it is also required that even the surfaces, as well as the bulk, are insulating [24]. Insofar as the crystalline system is unbounded the modulo ambiguity cannot be removed.

5.6 Polarization of a band insulator revisited

Within mean field (either Hartree-Fock or Kohn-Sham) the ground eigenstate $|\Psi_0\rangle$ in the Schrödinger representation is a Slater determinant of $N/2$ doubly occupied orbitals; in the crystalline case translational symmetry allows choosing the orbitals in the Bloch form. For the sake of simplicity we get rid of trivial factors of two, by considering a Slater determinant of singly occupied orbitals (so-called "spinless electrons"); furthermore we consider the contribution to $P_x^{(\text{el})}$ of a single occupied band.

In the simple cubic case, as dealt with above, the Bloch vectors are:

$$\mathbf{k}_m = \frac{2\pi}{Ma}(m_1, m_2, m_3), \quad m_s = 0, 1, \ldots, M - 1, \qquad (5.27)$$

81

where $m \equiv (m_1, m_2, m_3)$. The Bloch orbitals $|\psi_{\mathbf{k}_m}\rangle = \mathrm{e}^{i\mathbf{k}_m \cdot \mathbf{r}}|u_{\mathbf{k}_m}\rangle$ are normalized over the crystal cell of volume a^3. It is expedient to define the auxiliary Bloch orbitals $|\phi_{\mathbf{k}_m}\rangle = \mathrm{e}^{i\frac{2\pi}{L}x}|\psi_{\mathbf{k}_m}\rangle$, and $|\Phi_0\rangle$ as their Slater determinant; we also define $\mathbf{q} = (\frac{2\pi}{Ma}, 0, 0)$. Then

$$\langle \Psi_0 | \mathrm{e}^{i\sum_i \mathbf{q}\cdot\mathbf{r}_i} |\Psi_0\rangle = \langle \Psi_0 | \Phi_0\rangle = \frac{1}{M^{3N}}\det \mathcal{S}, \tag{5.28}$$

where $\mathcal{S}$ is the $N \times N$ overlap matrix of the orbitals, in a different normalization:

$$\begin{aligned}
\mathcal{S}_{mm'} &= M^3 \langle \psi_{\mathbf{k}_m} | \phi_{\mathbf{k}_{m'}}\rangle = M^3 \langle u_{\mathbf{k}_m} | \mathrm{e}^{i(\mathbf{q}+\mathbf{k}_{m'}-\mathbf{k}_m)\cdot\mathbf{r}} | u_{\mathbf{k}_{m'}}\rangle \\
&= M^3 \langle u_{\mathbf{k}_m} | u_{\mathbf{k}_{m'}}\rangle \delta_{\mathbf{q}+\mathbf{k}_{m'}-\mathbf{k}_m} = M^3 \langle u_{\mathbf{k}_m} | u_{\mathbf{k}_m - \mathbf{q}}\rangle \delta_{mm'}.
\end{aligned} \tag{5.29}$$

The normalization factors cancel: we have in fact

$$\langle \Psi_0 | \mathrm{e}^{i\frac{2\pi}{L}\sum_i x_i} |\Psi_0\rangle = \frac{1}{M^{3N}}\det \mathcal{S} = \prod_{m_1,m_2,m_3=0}^{M-1} \langle u_{\mathbf{k}_m} | u_{\mathbf{k}_m - \mathbf{q}}\rangle, \tag{5.30}$$

$$\begin{aligned}
\gamma_x^{(\text{crystal})} &= \frac{1}{M^2}\mathrm{Im} \ln \langle \Psi_0 | \mathrm{e}^{-i\frac{2\pi}{L}\sum_i x_i} |\Psi_0\rangle \\
&= -\frac{1}{M^2}\sum_{m_2,m_3=0}^{M-1} \mathrm{Im} \ln \prod_{m_1=0}^{M-1} \langle u_{\mathbf{k}_m} | u_{\mathbf{k}_m - \mathbf{q}}\rangle.
\end{aligned} \tag{5.31}$$

This is indeed the single-band version of the discretized Berry-phase formula routinely implemented in ab-initio electronic-stucture codes for computing macroscopic polarization [24]; the classical nuclear term has to be added. Indeed, the same equation has been originally obtained by King-Smith and Vanderbilt [101] upon discretization of Eq. (5.5).

5.6.1 The surface charge theorem

The early occurrences of the theorem of quantization of the surface charge [38, 39, 40, 41] are discussed above, Sect. 2.3. The topological nature of this theorem was first realized by Niu [129] in 1986; here we follow the treatment of Vanderbilt and King-Smith [117] (see also Ref. [5]).

According to elementary electrostatics the macroscopic bound surface charge density σ_{surface} residing on the surface of a sample is related to the polarization in the interior by $\sigma_{\text{surface}} = \hat{\mathbf{n}} \cdot \mathbf{P}$, where $\hat{\mathbf{n}}$ is the surface normal. One defines the bound charge σ_{surface} by saying that no free charge is present, but what, precisely, does this mean? The surface must be insulating, with the electron chemical potential lying in a gap that is common to both bulk and surface. But this is not a unique prescription,

since there can be a surface band which is entirely occupied or entirely empty. The two cases differ by a polarization quantum in the corresponding $\mathbf{P}$ value. In fact, given that the bulk polarization $\mathbf{P}$ is arbitrary modulo $e\mathbf{R}/V_{\text{cell}}$, it follows that the charge per surface area is defined modulo a quantum

$$\sigma_{\text{surface}} = \hat{\mathbf{n}} \cdot \mathbf{P} \quad \text{modulo} \ \frac{e}{A_{\text{surface}}} \ . \tag{5.32}$$

An equivalent formulation of the surface charge theorem can be arrived at by means of WFs. The WF approach is most perspicuous for quasi-$1d$ systems (e.g. insulating polymers); a pedagogical presentation is in Ref. [126]. Notice that all terminations are "insulating" in a polymer, there are no "surface" conducting states to speak of.

The bulk-surface correspondence encoded in Eq. (5.32) is an outstanding manifestation of topology in condensed matter physics: the surface charge of an insulating surface is "topologically protected". The actual value of σ_{surface} among the discrete allowed values is then determined by energy considerations.

For a centrosymmetric crystal it is tempting to guess that $\mathbf{P}(\lambda)$ in the Berry-phase formula, Eq. (5.4), vanishes for any λ. Instead, this is not the case: centrosymmetry only dictates that $\mathbf{P} = -\mathbf{P}$ modulo a quantum, and therefore Eq. (5.4) yields $\mathbf{P}$ equal to an integer or half integer multiple of the quantum $e\mathbf{R}/V_{\text{cell}}$ (for single band occupancy). Then from Eq. (5.32) it follows that the charge per surface cell may only be an integer or half integer, as first discovered many years ago, and previously discussed in Sect. 2.3. Therein, it was observed that this important theorem is often ignored even by specialists in surface physics. A thorough analysis of polar surfaces, in the light of the present theorem, has appeared in a couple of 2011 papers: by Stengel [130] and by Bristowe, Littlewood and Artacho [131].

Among the values dictated by topology, Nature chooses the minimum-energy one. If the electric field vanishes outside the solid, a charged surface implies a nonvanishing field inside. This has an extensive energy cost (proportional to the square of the field times the volume). If the bulk of the solid is centrosymmetric, the surface charge is zero; or otherwise the surface is metallic.

Quasi-$1d$ "crystalline" systems (i.e. stereoregular polymers) are more interesting. Therein: (1) the energy cost of the field is nonextensive and (2) the "surface" (i.e. the termination) is zero-dimensional and cannot be "metallic". Therefore different values of the surface quantum can be actually realized for the same bulk: we illustrate this with Figs. 5.4 and 5.5, taken from Ref. [126].

Alternant trans-polyacetylene is a centrosymmetric quasi-$1d$ insulating crystal: its end charges are quantized in units of half an election charge. We consider two different terminations, shown in Fig. 5.4: notice that in both cases the molecule as a whole is *not* centrosymmetric, although the bulk is. The end charges are trivially related to the dipole per unit length (or per monomer). Fig. 5.5 shows that, in the long system limit, the end charges are either zero or one, depending on the
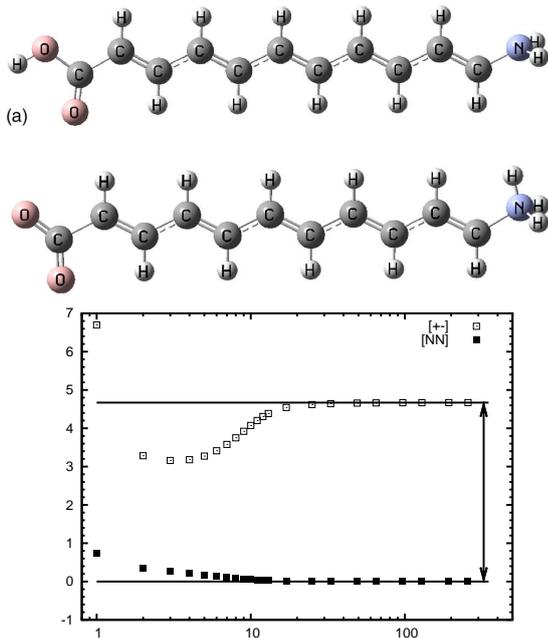
Figure 5.4: A centrosymmetric insulating quasi-$1d$ "crystal" with two different terminations: alternant trans-polyacetylene. Here the "bulk" is five-monomer long.

Figure 5.5: Quantization of the end charges in polyacetylene, after Ref. [126]. Dipole per monomer as a function of the number of monomers in the chain, for the two different terminations.

termination. The structure is either "neutral" or "charge-transfer"; the end groups are "donor" (NH_2) and "acceptor" (COOH). The figure also shows that at finite sizes both molecules are polar; the quantization is exact in principle in the infinite system limit; in the present case it is attained at about 10-20 monomers. This length is essentially the exponential decay length of the one-body density matrix of bulk polyacetylene.

In the polyacetylene case considered so far topology mandates the end charges to be zero modulo 1 (in units of e). Other cases where instead the end charge is 1/2 modulo 1 are also considered in Ref. [126]. The actual value of the end charges depends on the relative ionicity of the end groups vs. the bulk. Counterintuitively, while the ionicity varies on a continuous scale, the end charges may only vary by an integer. More about this is said below, Sect. 5.7.

5.6.2 The single-point Berry phase in the noncrystalline case

The key ingredient for computing the infrared spectrum of amorphous or liquid systems is the power spectrum of the autocorrelation function of the macroscopic polarization $\langle \mathbf{P}(t) \cdot \mathbf{P}(0) \rangle$. Since Car-Parrinello simulations are customarily performed using only $\mathbf{k} = 0$ in the (supercell) BZ, we need to analyze the single-point version of the Berry-phase formula for polarization, Eq. (5.5).

We consider a simple cubic supercell of side L; in the large L limit the BZ integral of any function $f(\mathbf{k})$ is approximated as

$$\int_{\text{BZ}} d\mathbf{k}\, f(\mathbf{k}) \;\rightarrow\; \frac{(2\pi)^3}{L^3} f(0). \tag{5.33}$$

We start with the Berry phase, i.e. with the $1d$ integral in square parenthesis in Eq. (5.5):

$$\gamma_z = i \int_{-\frac{\pi}{L}}^{\frac{\pi}{L}} dk_z \langle u_{j\mathbf{k}} | \partial_{k_z} u_{j\mathbf{k}} \rangle = i \int_0^{\frac{2\pi}{L}} dk_z \langle u_{j\mathbf{k}} | \partial_{k_z} u_{j\mathbf{k}} \rangle \rightarrow -\text{Im} \ln \det S(\mathbf{k}_1, \mathbf{k}_2), \tag{5.34}$$

where $\mathbf{k}_1 = (0,0,0)$ and $\mathbf{k}_2 = (0,0,2\pi/L)$. In Eq. (5.34) we have used the discretized Berry phase, Eq. (3.55), with only one factor in the matrix product. Then, as in Sect. 3.9.1, we notice that $|u_{j\mathbf{k}_2}\rangle = \mathrm{e}^{-i\frac{2\pi z}{L}} |u_{j\mathbf{k}_1}\rangle$: therefore the overlap matrix in Eq. (5.34) becomes

$$S_{jj'}(\mathbf{k}_1, \mathbf{k}_2) = \langle u_j | \mathrm{e}^{-i\frac{2\pi z}{L}} | u_{j'} \rangle, \tag{5.35}$$

where all the orbitals $|u_j\rangle = |\psi_j\rangle$ are evaluated at $\mathbf{k} = 0$. We then approximate even the remaining integrals in Eq. (5.5) with a single point. At any time during the simulation the electronic term in the polarization is thus

$$P_z^{(\text{el})}(t) = -\frac{e}{\pi L^2} \gamma_z = \frac{e}{\pi L^2} \text{Im} \ln \det S. \tag{5.36}$$

The nuclear (or core) contribution has a very simple form. If z_m is the instantaneous z coordinate of the m-th nucleus, and eZ_m the corresponding charge, the total polarization is

$$P_z(t) = \frac{e}{\pi L^2} \text{Im} \ln \det S + \frac{e}{L^3} \sum_m Z_m z_m. \tag{5.37}$$

This the expression currently used in computing power spectra and infrared spectra [93], and, more generally, whenever a single $\mathbf{k}$ point is used in the first-principle simulations [132, 94]. As already observed, even the nuclear term is affected by the quantum indeterminacy: we can therefore express it equivalently into an "Im log" form. It is easy to show that Eq. (5.37) can be equivalently written as

$$P_z(t) = \frac{e}{2\pi L^2} \text{Im} \ln \left[(\det S)^2 \mathrm{e}^{i\frac{2\pi}{L} \sum_m Z_m z_m} \right], \tag{5.38}$$

where the square of the determinant owes to the double spin occupancy.

Eq. (5.38) is nothing else than the independent-electron formulation of Eq. (5.22). We have already observed that the polarization quantum in Eq. (5.37) is e/L^2, which vanishes in the $L \rightarrow \infty$ limit: therefore polarization itself is ill-defined in the noncrystalline case. Nonetheless thre is no problem in the present context, and in

85

fact Eq. (5.36) is routinely used for evaluating infrared spectra of amorphous and liquid materials at finite temperature; some more details are given in Sect. D.11.

The key point is that the $L \to \infty$ limit is not actually needed; for an accurate description of a given material, it is enough to assume a *finite L*, actually larger than the relevant correlation lengths and diffusion lengths in the material. For any given L, the quantum e/L^2 sets an upper limit to the magnitude of a polarization difference accessible via the Berry phase. The larger are the correlation lengths, the smaller is the accessible $\Delta \mathbf{P}$. This is no problem at all in practice, either when evaluating static derivatives by numerical differentiation, such as e.g. in Ref. [132, 94], or when performing Car-Parrinello simulations [93]. In the latter case Δt is a Car-Parrinello time step (a few a.u.), during which the polarization varies by a tiny amount, much smaller than the quantum (the typical size of a large simulation cell nowadays is $L \simeq 50$ bohr). Whenever needed, the drawback may be overcome by splitting Δt into several smaller intervals. More about computing infrared spectra via molecular dynamics simulations will be said in Sect. D.11.

5.6.3 Kohn-Sham polarization vs. real polarization

All of the independent-electron formulæ discussed in Sect. 5.3 are exact for noninteracting electrons, but the obvious aim is to implement them with KS orbitals, in a given density-functional theory (DFT) framework. Since macroscopic polarization applies to insulators only, we stress that we mean "KS insulator" throughout: that is, we assume that the KS spectrum is gapped. In the class of "simple" (i.e. computationally friendly) materials a genuine insulator is also a KS insulator, although pathological cases (computationally unfriendly) do exist.

Having specified this, the key issue is then: Does the KS polarization coincide with the physical many-body one? The answer is subtle, and is different whether one chooses either "open" boundary conditions, as appropriate for molecules and clusters, or periodic boundary conditions (Born-von Kármán), as invariably done in the present Notes.

Within open boundary conditions the KS orbitals vanish at infinity, as well as the charge density of the sample. $\mathbf{P}$ is then the first moment of the charge density, divided by the sample volume. The basic tenet of DFT is that the microscopic density of the fictitious noninteracting KS system coincides with the density of the interacting system. Therefore the exact $\mathbf{P}$ coincides by definition with the one obtained from the KS orbitals.

Matters are quite different within periodic boundary conditions: we have seen above that $\mathbf{P}$ is *not* a function of the charge density, hence the value of $\mathbf{P}$ obtained from the KS orbitals, in general, is not the correct many-body $\mathbf{P}$. This was first shown in 1995 by Gonze, Ghosez, and Godby [133], and later discussed by several authors. A complete account of the issue can be found in Refs. [6, 64]. Here we

just mention that the exact **P** is provided by Eq. (5.22), while the KS **P** is provided by Eq. (5.36), where the KS orbitals enter Eq. (5.35); both expressions are to be evaluated in the large-L limit. The two expressions are clearly different whenever the ground wave function is not a Slater determinant.

Therefore **P** cannot be exactly expressed within DFT, but the exact functional is obviously inaccessible, and even sometimes pathological. The practical issue is whether the current popular functionals provide an accurate approximation to the experimental values of **P** in a large class of materials.

A vast first-principle literature has accumulated over the years by either linear-response theory [134]—not reviewed here—or by the modern theory. The errors are typically of the order of 10-20% on permittivity, and much less on most other properties (infrared spectra, piezoelectricity, ferroelectricity) for many different materials. It is unclear which part of the error is to be attributed to DFT per se, and which part is to be attributed to the *approximations* to DFT. For a very recent (2018) pedagogical presentation of the issue see Ref. [64].

5.7 Polarization as a $\mathbb{Z}_2$ topological invariant

We consider in this Section only crystalline systems in $1d$ or quasi-$1d$ (i.e. stereoregular polymers): polarization P has the dimensions of a charge, and therefore P/e is dimensionless. We further limit ourselves to the *centrosymmetric* case.

We start rewriting the most compact single-point formula, in the $1d$ case, Eq. (5.23):

$$P/e = \frac{1}{2\pi}\text{Im ln } \langle\Psi_0|e^{i\frac{2\pi}{L}(\sum_\ell Z_\ell X_\ell - \sum_i x_i)}|\Psi_0\rangle; \qquad (5.39)$$

we remind that P is well defined only for a charge-neutral system, and indeed Eq. (5.39) includes both electronic and nuclear contributions to P.

For a centrosymmetric system the matrix element in Eq. (5.39) is real, hence its phase is either 0 or π (mod 2π). We also remind that, in general, Eq. (5.39) is meaningful in the large-system limit: N electrons in a periodic box of length L, with N/L constant. In a centrosymmetric system the matrix element is real at any N: the limit is no longer necessary.

It is expedient to consider $2P/e$, which therefore assumes the values of either 0 mod 2, or 1 mod 2. This establishes a $\mathbb{Z}_2$ classification of $1d$ centrosymmetric materials: they are either $\mathbb{Z}_2$-even ($P = 0$ mod e) or $\mathbb{Z}_2$-odd ($P = e/2$ mod e).

We have not yet proved that such classification is indeed topological. Let us consider the two simplest nonprimitive (binary) centrosymmetric lattices in $1d$. The top sketch refers to an idealized molecular crystal, or to an alternant polymer, like polyacetilene (compare to Fig. 5.4); the bottom sketch refers to an idealized
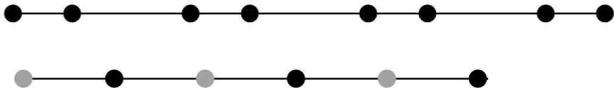
Figure 5.6: Two paradigmatic centrosymmetric lattices in $1d$

ionic crystal. The former is $\mathbb{Z}_2$-even (see Fig. 5.5), while the latter is $\mathbb{Z}_2$-odd. Both system can be described at the simplest level by an extreme tight-binding model Hamiltonian (Hückel-like), whose only parameters are the first-neighbor hoppings t and the onsite energies ϵ. The $\mathbb{Z}_2$-even model has alternating t's and constant ϵ; the $\mathbb{Z}_2$-odd has constant t and alternating ϵ's. It is a straightforward exercise to show that one cannot "continuously deformate" the Hamiltonian (and its ground state) from one case into the other without closing the gap, while conserving centrosymmetry. In modern jargon, the $\mathbb{Z}_2$ invariant is "protected" by centrosymmetry. Provided we don't close the gap, we may also deformate the tight-binding ground state into a first-principle one: the classification is independent of the theory level.

Since PBCs are at the root of Eq. (5.39), we have addressed so far an unbounded sample: the $\mathbb{Z}_2$ invariant is a bulk property, but is determined only modulo a "quantum". Real systems are instead bounded: polarization remains ambiguous until the termination of the sample has been specified. In that case Nature chooses the minimum-energy eigenstate, while respecting the topology constraint (see Sect. 5.6.1). We also observe that for a bounded sample topology only appears in the large-N limit (Fig. 5.5), while the PBC expression of Eq. (5.39) is quantized even at finite N.

One further comment about the ionic case: bottom sketch in Fig. 5.6. If we identify the structure with a lattice of classical point charges $\pm e$, it is obvious that the dipole per unit length of a bounded sample is $\pm e/2$, depending on the termination. When replacing the classical point charges with real anions and cations, one expects the dipole per unit length to be *about* $\pm e/2$, but only in the limit of a strongly ionic system. Conterintuitively, topology mandates that—in the large system limit—the dipole per unit length is $\pm e/2$ even in the case of very weak ionicity, i.e. a binary chain made of Ga and As atoms.

The proof that the end charge in centrosymmetric linear polymers is topological can be equivalently arrived at by Wannier-function counting: see Ref. [126]. Similar arguments also prove the topological nature of the soliton charge in polyacetilene, first discovered by Su, Schrieffer, and Heeger in 1979 [135]. If we insist that we want a singlet wavefunction—as everywhere in this Section—then the soliton charge can only be $\pm e$. But we may relax this condition, adopting single occupancy (instead of double) for one of the Wannier functions in the soliton region: the soliton is then neutral, but carries spin $\pm 1/2$.

88

The $\mathbb{Z}_2$ is addressed again below, Sect. 7.5.5, for a model interacting many-electron $1d$ system. It will be shown that by increasing the strength of the interaction the system undergoes a topological quantum transition, from $\mathbb{Z}_2$-odd to $\mathbb{Z}_2$-even, crossing a metallic state.

Chapter 6

Chern-Simons geometric phase

6.1 Axion term in magnetoelectric response

We have stressed in Sect. 3.10.4 that differential geometry and algebraic topology share several common features in $2n + 1$ dimensions. Macroscopic polarization is essentially a $1d$ phenomenon whose electronic contribution is a geometrical property of the electronic ground state: the BZ integral of the trace of a Chern-Simons 1-form, Eq. (3.57). A question then naturally arises: does it exist (in insulators) a $3d$ phenomenon whose geometrical expression is the BZ integral of the trace of a Chern-Simons 3-form, Eq. (3.72)? The positive answer was provided by Qi, Hughes, and Zhang in 2008 [62]; see also Refs. [63, 136] for a more understandable presentation. The geometrical observable enters the theory of the magnetoelectric response, outlined in in Appendix A, and particularly in A.6.

The magnetoelectric response tensor is—as any response—expressed via linear-response theory and Kubo formulæ, but it includes a term which is a *pure ground-state* term, and does not require perturbation theory: the so-called "axion" term, a pseudoscalar. Its expression is

$$\alpha^{\mathrm{CS}} = \frac{1}{4\pi^2} \frac{e^2}{\hbar c} \theta = \frac{1}{2\pi} \frac{e^2}{hc} \theta,$$
(6.1)

where θ is an angle [137, 136, 60], and "CS" stays for Chern-Simons. The literature often adopts SI units, where the magnetoelectric susceptibility is no longer dimensionless; in this case

$$\alpha^{\mathrm{CS}} = \frac{e^2}{2\pi h} \theta \qquad \text{(SI units)}.$$
(6.2)

The θ angle is a bulk property defined modulo 2π; therefore α^{CS} is a bulk property with a "quantum" of arbitrariness, which (in Gaussian units, adopted here) is proportional to the fine-structure constant. When we address a bounded crystallite,

the value of α^{CS} is no longer ambiguous once the surface termination of the $3d$ sample has been specified. The analogy of α^{CS} with macroscopic polarization—and of θ with the corresponding Berry phase—is evident.

The main ingredient is the nonAbelian Berry connection, Eq. (3.56), which is a 1-form with matrix coefficients

$$\mathscr{A}_{\alpha,jj'}(\mathbf{k}) = i\langle u_{jk}|\partial_\alpha u_{j'k}\rangle, \tag{6.3}$$

which for any Cartesian coordinate α are Hermitian matrices in the band indices jj'. Since spin-orbit coupling is essential in magnetoelectric materials, the Bloch states $|u_{jk}\rangle$ are actually two-component spinors (a.k.a. spinorbitals), and the matrix dimensions are even. In the following we will omit the band-and-spin indices: the symbol $\mathscr{A}_\alpha(\mathbf{k})$ will indicate the Hermitian matrix. The symbol "tr" indicates the matrix trace over the band indices, while instead the symbol "Tr" indicates the trace over the Hilbert space. The nonAbelian connection is therefore the differential form

$$A = i\langle u_{j\mathbf{k}}|\partial_{\mathbf{k}} u_{j'\mathbf{k}}\rangle \cdot d\mathbf{k} = \mathscr{A}_\alpha(\mathbf{k})dk^\alpha. \tag{6.4}$$

The $\mathbf{k}$ space expression for θ in terms of the nonAbelian Berry connection looks formidable, but is in fact a known formula in differential geometry; the compact expressions used by mathematicians have been presented in Sect. 3.10.4. As said above, θ is the BZ integral of the trace of the nonAbelian Chern-Simons 3-form [137, 136]:

$$\theta = -\frac{1}{4\pi}\epsilon_{\alpha\gamma\beta}\int_{\text{BZ}} d\mathbf{k}\,\text{tr}\,\left[\mathscr{A}_\alpha(\mathbf{k})\partial_\beta\mathscr{A}_\gamma(\mathbf{k}) - \frac{2i}{3}\mathscr{A}_\alpha(\mathbf{k})\mathscr{A}_\beta(\mathbf{k})\mathscr{A}_\gamma(\mathbf{k})\right].. \tag{6.5}$$

In the Abelian case the connection matrices are diagonal and real: hence the second term in the trace vanishes. In the simplest Abelian case (i.e. two-band) the Chern-Simons geometric phase, Eq. (6.5), has a very simple expression in terms of the connection $\boldsymbol{\mathcal{A}}(\mathbf{k})$ (having two spin components) written in vector form:

$$\theta = -\frac{1}{4\pi}\int_{\text{BZ}} d\mathbf{k}\,\boldsymbol{\mathcal{A}}(\mathbf{k})\cdot\nabla_\mathbf{k}\times\boldsymbol{\mathcal{A}}(\mathbf{k}) = -\frac{1}{4\pi}\int_{\text{BZ}} d\mathbf{k}\,\boldsymbol{\mathcal{A}}(\mathbf{k})\cdot\boldsymbol{\Omega}(\mathbf{k}), \tag{6.6}$$

where $\boldsymbol{\Omega}(\mathbf{k})$ is the $3d$ Berry curvature written in vector form as well.

6.1.1 $\mathbb{Z}_2$ topological insulators in $3d$

The Chern-Simons term in the free energy is proportional to $\mathbf{E}\cdot\mathbf{B}$, ergo odd under time-reversal and under inversion. This implies $\theta = -\theta$ if the crystal has any of the above symmetries. Notice also that in order to have a genuine magnetoelectric response both symmetries must be broken (Appendix A).

Since θ in an unbounded insulating sample is only defined modulo 2π, this means that both $\theta = 0$ and $\theta = \pi$ are possible in T-invariant (TI) crystals (regardless of inversion symmetry) and in centrosymmetric crystals (regardless of T-symmetry). Nontrivial ($\theta = \pi$) insulators of the former class are called "strong topological insulators"; those of the latter class are called "axion insulators" [138].

In both cases all linear-response (Kubo) terms accounting for magnetoelectric coupling vanish, and the angle θ becomes topological: a $\mathbb{Z}_2$ invariant. This is in analogy with γ in $1d$ centrosymmetric insulators, mentioned above and thoroughly discussed in Sec. 5.7.

The invariant θ equals π for crystals in the class of strong $\mathbb{Z}_2$ topological insulators, and therefore Eq. (6.5) provides an alternative way of detecting topological order. Despite the difficulties in discretizing Eq. (6.5)—illustrated in Sec. 6.1.2 below—the value of $\theta = \pi$ has been numerically verified for Bi_2Se_3 [60], which is a paradigmatic material: its $\mathbb{Z}_2$-odd character has been assessed via other theoretical tools, and experimentally as well.

Notice that $\theta = \pi$, according to Eq. (6.1), would yield a magnetoelectric susceptibility equal (in Gaussian units) to $1/4\pi$ times the fine structure constant, i.e. $\simeq 6 \times 10^{-4}$, a rather large value compared to common magnetoelectrics.

The θ value addressed so far is evaluated as a BZ integral, Eq. (6.5), for an unbounded sample. Next we ideally address a bounded crystallite cut from a $\mathbb{Z}_2$-odd crystal, where the bulk T symmetry is preserved at the surface as well. If the whole surface were insulating, the arguments given above would guarantee a nonvanishing (and large) magnetoelectric susceptibility. This is clearly absurd: any T-invariant finite system has a null magnetoelectric susceptibility (all terms of it, not only the Chern-Simons term). The solution is obvious: the surface must be metallic. It is known by independent arguments, in fact, that $\mathbb{Z}_2$-odd T-invariant insulators have topologically protected metallic states at their surfaces.

The case of axion insulators, where the topological invariant is protected by inversion symmetry (and not by time reversal) deserves a separate discussion, in what concerns their surface properties. In fact it is impossible to preserve inversion symmetry at a surface; the nontrivial consequences of this fact are discussed in Ref. [138]. Last but not least, axion insulators remain hypothetical at the time of writing (2019): no material of this class has been synthesized so far.

6.1.2 Numerical considerations

The analytical BZ integrals for the geometrical properties all require twice differentiable Bloch orbitals $|u_{j\mathbf{k}}\rangle$ throughout the Brillouin zone. In numerical work the BZ integrals are discretized over a $\mathbf{k}$-point mesh; the orbitals $|u_{j\mathbf{k}}\rangle$ are then *not* a differentiable function of $\mathbf{k}$. The phase factors are erratic, and further unitary mixing of the occupied states at a given $\mathbf{k}$ is usual. One therefore needs replacing

the analytical formulas with their discretized counterpart, which must converge to the respective analytical expression in the dense-$\mathbf{k}$ limit.

In the present Notes we have described in detail the discretization of the Berry phase γ in Sect. 3.9.4 and of the Chern number C_1 in 4.1.4; more will be said below in Sec. 8.3.4. In all these cases the expressions are *numerically* gauge invariant. By this we mean that the expressions can be directly implemented with the raw Bloch states (as provided by Hamiltonian diagonalizations) where the gauge is erratic. A naive discretization—based on replacing the $\mathbf{k}$-derivatives in the integrands in Eqs. (6.8) and (6.10) with finite differences—fails because the raw $|u_{j\mathbf{k}}\rangle$ are not differentiable. The discretizations discussed in Secs. 3.9.1 and 8.3.4 follow a different path and are unaffected by the erratic gauge.

In the case of the Chern-Simons phase θ a numerically gauge-invariant discretization has not be found so far (except in the Abelian case). Therefore the gauge must be regularized before the BZ integral, Eq. (6.5), is discretized [60, 61].

6.2 Polarization and Chern number revisited

In order to emphasize analogies and differences with some of the previous results, here we recast them in terms of the nonAbelian Berry connection.

We have shown in Ch. 5 that the *electronic* contribution to the polarization of a $1d$ system can be written in the form

$$P_x = -e\frac{\gamma}{2\pi}, \tag{6.7}$$

where γ is the Berry phase. Notice that here we define γ for double occupancy, i.e. inluding the factor of two. Polarization is a well defined observable only for charge-neutral systems. The Berry phase can be generalized to include the nuclear contribution as well: Eq. (6.7) defines the total polarization if we set

$$\gamma = 2\int_{BZ} dk \, \mathrm{tr}\,[\mathscr{A}_x(k)] + \mathrm{Im}\,\ln\,\mathrm{e}^{-i\frac{2\pi}{a}\sum_m Z_m x_m^{(n)}}. \tag{6.8}$$

In Eq. (6.8) Z_m are the dimensionless nuclear charges, and $x_m^{(n)}$ the nuclear coordinates in the $1d$ unit cell of size a. Thanks to charge neutrality, Eq. (6.8) in invariant by translation of the origin (each of the two terms is not such). As we have already stressed, the electronic term in Eq. (6.8) he BZ integral of the trace of a Chern-Simons 1-form.

Next we switch to the Berry curvature, which is a 2-form. We have previously outlined the fact that features of differential geometry and algebraic topology are quite different in even vs. odd dimensions. The Berry curvature is called a Chern

1-form (in general Chern n-forms are indeed $2n$-forms). Chern-forms are gauge-covariant, and their trace is gauge-invariant, differently from Chern-Simons forms. The BZ-integral of the trace of a Chern form is quantized, i.e. (when properly normalized) is a $\mathbb{Z}$ invariant.

The Chern invariant in $2d$ has been defined in Eq. (3.49); we recast it here in terms of the nonAbelian Berry connection as

$$2\pi C_1 = i \int_{\text{BZ}} d\mathbf{k} \text{ tr } [\partial_x \mathscr{A}_y(k) - \partial_y \mathscr{A}_x(k)] = i\varepsilon_{\alpha\beta} \int_{\text{BZ}} d\mathbf{k} \text{ tr } [\partial_\alpha \mathscr{A}_\beta(k)]. \qquad (6.9)$$

This is written for spin channel, or for "spinless electrons", as everywhere in these Notes. In terms of the nonAbelian Berry curvature matrix $\mathbf{\Omega}_{\alpha\beta}(\mathbf{k})$ the expression is

$$2\pi C_1 = i \int_{\text{BZ}} d\mathbf{k} \text{ tr } [\mathbf{\Omega}_{xy}(\mathbf{k})]. \qquad (6.10)$$

Notice that in the nonAbelian case

$$\mathbf{\Omega}_{\alpha\beta}(\mathbf{k}) = \partial_\alpha \mathscr{A}_\beta(\mathbf{k}) - \partial_\beta \mathscr{A}_\alpha(\mathbf{k}) - i[\mathscr{A}_\alpha(\mathbf{k}), \mathscr{A}_\beta(\mathbf{k})], \qquad (6.11)$$

but the commutator does not contribute to the trace in Eq. (6.9), which therefore looks the same in either the Abelian or nonAbelian case.

The right-hand member integral in Eq. (6.9) can be interpreted as an angle, equal to an integer number of 2π. The form of Eq. (6.9) emphasizes the analogies, but also the differences, with both the polarization phase angle γ, Eq. (6.8) and the Chern-Simons phase, Eq. (6.5). The invariant has no "modulo" arbitrariness, hence it provides a $\mathbb{Z}$ (not $\mathbb{Z}_2$) topological classification; it can be nonzero only in crystals which lack T-symmetry

6.2.1 Open boundary conditions

Insofar as we adopt PBCs the system is unbounded and P_x is well defined only modulo e. If we consider instead a finite realization of the same $1d$ Hamiltonian with length L, its OBCs dipole d_x divided by L assumes—in the large-L limit—one of the quantized values dictated by the bulk: see e.g. Fig. 5.4. If we adopt Eq. (6.7) even in the OBC case, the dimensionless polarization can be written as

$$\tilde{\gamma} = -\frac{2\pi d_x}{eL} = 2\pi \left[-\frac{2}{L}\text{Tr }\{xP\} + \frac{1}{L}\sum_m Z_m x_m^{(n)} \right], \qquad (6.12)$$

where P is the ground-state projector and "Tr" is the trace over the Hilbert space.

In the special case where the bulk of the $1d$ system is centrosymmetric, the phase γ becomes topological, as thoroughly discussed in Sec. 5.7; the same happens to

$\tilde{\gamma}$ in the large-L limit. Centrosymmetric systems in $1d$ admit a $\mathbb{Z}_2$ classification: $\mathbb{Z}_2$-even means $\gamma = 0 \bmod 2\pi$, $\mathbb{Z}_2$-odd means $\gamma = \pi \bmod 2\pi$,

Next we switch to the axion term, for spinful electrons. The full density matrix is therefore a projector in the space of two-component spinors, at variance with the one entering Eq. (6.8), which projects on a single spin channel (doubly occupied). The Chern-Simons phase has a rather simple expression in $\mathbf{r}$ space for a bounded $3d$ crystallite within OBCs. According to the literature [136] its expression in terms of the ground-state projector is

$$\tilde{\theta} = -\frac{4\pi^2}{3V}\varepsilon_{\alpha\gamma\beta}\text{Im Tr } \{r_\alpha Pr_\beta Pr_\gamma P\}, \qquad (6.13)$$

where V is the crystallite volume, in the large-V limit; a crucial requirement is that the crystallite has an *insulating boundary*. As for the case of polarization, the modulo 2π ambiguity in Eq. (6.13) is fixed only after the actual sample termination is specified.

We notice that the ground state projector P enters the expression for $\tilde{\gamma}$ once, and it appears three times in the expression for $\tilde{\theta}$: this is reminiscent of their $\mathbf{k}$-space counterparts γ and θ, which were obtained from a Chern-Simons 1-form and 3-form, respectively. The Chern number can be written in a form which superficially looks similar to Eqs. (6.12) and (6.13). In fact it will be shown below in these Notes that

$$2\pi C_1 = 4\pi^2 \varepsilon_{\alpha\beta} \text{ Im Tr}_\text{V} \{r_\alpha Pr_\beta P\}, \qquad (6.14)$$

where Tr_V is the trace per unit volume, taken in the inner region of the system only, and *not* on the whole bounded sample (the reasons for this caveat will be discussed below); instead, the traces in Eqs. (6.12) and (6.13) are over the whole sample. A basic feature of Eq. (6.14) is that it has no modulo indetermination and is therefore a $\mathbb{Z}$ invariant, independent of the actual sample termination.

6.2.2 Why even and odd dimensions are different

At variance with Eqs. (6.12) and (6.13), the ground state projector P enters Eq. (6.14) twice, i.e. an *even* number of times: this makes a huge difference, which is reminiscent of the huge difference between odd-n-forms and even-n-forms in reciprocal space. All of the OBCs expressions in Sect. 6.2.1 are for a bounded crystallite, with square-integrable orbitals (or spinorbitals); they cannot be adopted as such within PBCs, because the position operator $\mathbf{r}$ becomes ill-defined therein [103]. It turns out that this is not a serious drawback when the number of projectors is even, i.e. in the case of Eq. (6.14).

We define the complementary projector $Q = I - P$; then using $P^2 = P$ we recast Eq. (6.14) as

$$2\pi C_1 = 4\pi^2 \epsilon_{\alpha\beta} \text{ Im Tr}_\text{V} \{Pr_\alpha Pr_\beta P\} = -4\pi^2 \epsilon_{\alpha\beta} \text{ Im Tr}_\text{V} \{Pr_\alpha Qr_\beta P\}; \qquad (6.15)$$

then we notice that the operator $Pr_\alpha Qr_\beta P$ is bounded and lattice periodical. This is the major virtue making C_1 boundary insensitive, with no modulo arbitrariness. In fact $Pr_\alpha Qr_\beta P$ is well defined even within PBCs: when evaluated over Bloch states, it only requires off-diagonal matrix elements of $\mathbf{r}$, which are well defined. This is a staple of linear-response theory [134], which exploits the commutator $[H, \mathbf{r}] = -i\hbar\mathbf{v}$:

$$\langle\psi_{j\mathbf{k}}|\,\mathbf{r}\,|\psi_{j'\mathbf{k}}\rangle = i\hbar\frac{\langle\psi_{j\mathbf{k}}|\,\mathbf{v}\,|\psi_{j'\mathbf{k}}\rangle}{\epsilon_{j'\mathbf{k}} - \epsilon_{j'\mathbf{k}}}, \qquad j \neq j'. \tag{6.16}$$

A similar manipulation does not help when the number of P operators is odd: in Eq. (6.12) the unbounded operator xP cannot be made bounded anyhow. In Eq. (6.13), replacement of only one of the P operators gives some clue about the $1d$-$3d$ analogy. In fact

$$\theta = \frac{4\pi^2}{3V}\epsilon_{\alpha\gamma\beta}\text{Im Tr}\,\{r_\alpha(Pr_\beta Qr_\gamma P)\,\}, \tag{6.17}$$

where we have exploited antisymmetry. The inserted parenthesis isolate–by the associative property—the operator $Pr_\beta Qr_\gamma P$, which is bounded and lattice periodical, as said above. But this is multiplied by the unbounded operator r_α before the trace is taken: the drawback is therefore not much different from the case of $1d$ polarization.

Chapter 7

Theory of the insulating state

The electron was discovered by J. J. Thomson in 1896. Soon afterwards it became clear that the electron is responsible for electrical conduction in metals, and conversely for the lack of conduction in insulators. The two opposite properties were explained in the early 1900s by two somewhat opposite models for the electron behavior: Drude's (1900) and Lorentz's (1906), both obviously based on classical mechanics (Fig. 7.1). According to the Drude model the electrons in a metal roam over macroscopic distances, hindered by some dissipative mechanism (in order to guarantee Ohm's law). According to the Lorentz model instead the electrons in an insulator are modelled as (charged) harmonic oscillators: they cannot carry a steady-state current, and they polarize instead.

The advent of quantum mechanics solved many open problems, including this one. All electrons are of one kind; Bloch theorem, together with the Pauli principle, provides a simple discrimination between metals and insulators. Bloch theorem appeared in 1928 [139], and the main explanation of the insulating/metallic behavior
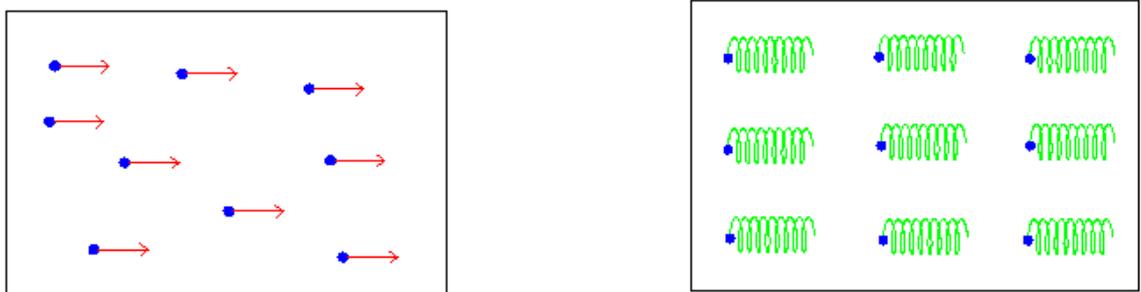

Figure 7.1: Schematic view of metals vs. insulators in classical physics. Left sketch: Drude model for metals, where electrons roam freely over macroscopic distances, hindered only by some dissipative mechanism. Right sketch: Lorentz model for insulators, where each electron is tied (by an harmonic force) to a particular center.

was apparently first proposed by Wilson in a couple of 1931 papers [140] (Peierls did also claim priority). The single-particle spectrum of a lattice-periodical Hamiltonian is in general gapped, and the electron count determines where the Fermi level lies. If it crosses a band one has a conductor: an applied electric field induces free acceleration of the electrons (at $T = 0$ in absence of dissipation). If the Fermi level lies instead in a gap, one has an insulator: in presence of a field the electronic system polarizes, but no steady-state current flows for $T \to 0$. This is what all undergraduates learn nowadays: band structure explains the insulating/conducting behavior of most common crystalline materials across the periodic table.

At the root of band theory are two basic assumptions: the electrons are noninteracting (in a mean-field sense), and the solid is crystalline. By the late 1950s, however, it became clear that there are solids to which this description does not apply: their insulating behavior is due to completely different mechanisms. The works of Mott in 1949 [141] and of Anderson in 1958 [142] opened new avenues in condensed matter physics. In the materials which we now call Mott insulators the insulating behavior is due to electron correlation [143], while in those called Anderson insulators it is due to lattice disorder [144]. P. W. Anderson and N. Mott were among the recipients of the 1977 Nobel prize for their previous work on the insulating state.

Therefore in the early 1960s it became desirable to understand the insulating state in a somewhat more general way than the textbook approach, based on band theory. In 1964 Walter Kohn published a milestone paper, bearing the same title as the present Chapter [128]. Kohn defined the insulating state making neither reference to electronic excitations nor to Fermi-level properties: the qualitative difference between insulators and metals manifests itself also in a different organization of the electrons in their many-body *ground state*. Even before the system is excited by any probe, a different organization of the electrons is present in the ground state and this is the key feature discriminating between insulators and metals [128, 145]. According to Kohn, the electrons in the insulating state satisfy a many-electron localization condition; this kind of localization must be defined in a subtle way given that, for instance, the Hamiltonian eigenstates in a band insulator are obviously not localized. According to the original Kohn's formulation, the insulating behaviour arises whenever the ground-state wavefunction of an extended system breaks up into a sum of contributions which are localized in essentially disconnected regions of the many-electron configuration space.

A series of more recent papers [9, 21, 146, 147, 148, 149, 150, 151] has established Kohn's pioneering viewpoint on a sound formal and computational basis, rooted in geometrical concepts. In the modern reformulation of the theory, Kohn's localization is measured by an appropriate quantity—having the dimensions of a squared length—which converges to a finite value in all kinds of insulators, and diverges

in all kinds of metals. These developments followed (and were inspired by) the modern theory of polarization.

The ultimate polarization formula, capable of addressing any insulator (including cases with disorder and electron-electron interaction), was published by the present author in 1998 [103]. As shown above, the electronic term in polarization can be cast in terms of a "single-point Berry phase" γ, Eq. (5.22), which is the phase of a complex number $\mathfrak{z}_N$, expressed as a ground-state expectation value, and whose modulus is no larger than one. Whenever the modulus of $\mathfrak{z}_N$ goes to zero in the large-system limit, then its phase—hence bulk polarization—is ill defined: Resta and Sorella therefore postulated in 1999 [146] that the novanishing/vanishing of $|\mathfrak{z}_N|$ characterizes the insulating/metallic nature of a condensed many-electron system. Such bold postulate marks the starting point of the modern theory of the insulating state; the postulate is based on the key observation that the static polarization of a metallic sample is *qualitatively* different from the polarization of an insulating one. In the metallic case the macroscopic polarization is trivial, material-independent: it completely screens any static field (Faraday-cage effect). In the insulating case, instead, the macroscopic polarization is nontrivial and material-dependent, both with and without an applied field.

Before Ref. [146] the focus of the theory of the insulating state was invariably on dc conductivity only, disregarding the alternative polarization characterization. The most recent papers—and the present Notes as well—also address conductivity issues in their relationship with polarization and with the behavior of $|\mathfrak{z}_N|$.

7.1 Quadratic spread of the Wannier functions

As shown in Sect. 5.3.3 the electronic term in macroscopic polarization is proportional to the first moment of the Wannier distribution, Eq. (5.9); the theory of the insulating state has instead some important relationship with the *second cumulant moment* of this distribution, also called "quadratic spread" by Marzari and Vanderbilt [56, 19] (hereafter quoted as MV). It is defined as

$$\Omega = \sum_{j=1}^{n_{\mathrm{b}}} [\, \langle w_{j\mathbf{0}}|\, r^2\, |w_{j\mathbf{0}}\rangle - |\langle w_{j\mathbf{0}}|\, \mathbf{r}\, |w_{j\mathbf{0}}\rangle|^2\,] = \sum_{m=1}^{n_{\mathrm{b}}} (\, \langle r^2\rangle_{j\mathbf{0}} - |\mathbf{r}_{j\mathbf{0}}|^2). \qquad (7.1)$$

This is *not* gauge invariant, but has a gauge-invariant lower bound (a minimum only in 1d), invariably called Ω_{I} in the modern literature, defined as the Cartesian trace of the BZ-integrated metric tensor:

$$\Omega_{\mathrm{I}} = V_{\mathrm{cell}} \sum_{\alpha=1}^{d} \int_{\mathrm{BZ}} \frac{d\mathbf{k}}{(2\pi)^d}\ g_{\alpha\alpha}(\mathbf{k}); \qquad (7.2)$$

the metric tensor $g_{\alpha\beta}(\mathbf{k})$ has been defined above, Eqs. (3.43) and (3.47). Notice that Ω_{I}—as well as Ω, Eq. (7.1)—is an *extensive* quantity (when we consider e.g. supercells of increasing size). Notice also that it refers to single band occupation: it is therefore a "spinless" quantity.

The gauge-invariant quadratic spread Ω_{I} enters in a fundamental way the modern theory of the insulating state, when it addresses the special case of a band insulator (i.e. a crystalline system of independent electrons). In Sect. 7.2.6 we will define the intensive quantity λ^2 as the quadratic spread per electron and per Cartesian coordinate: the squared Resta-Sorella localization length, Eq. (7.43) below. We anticipate that, in the special case of an isotropic band insulator, λ^2 is related to the MV gauge-invariant quadratic spread as $\lambda^2 = \Omega_{\mathrm{I}}/(n_{\mathrm{b}}d)$, where n_{b} is the number of occupied bands and d is the dimension.

The MV gauge-invariant quadratic spread Ω_{I} entered the literature as a formal auxiliary quantity. But—after the groundbreaking paper by Michael Berry [33, 25]— a basic tenet of quantum mechanics is that any gauge-invariant quantity is, at least in principle, an observable. The establishment of Ω_{I} as a physical observable is due to Souza, Wilkens, and Martin in 2000 [147]: by means of a kind of fluctuation-dissipation theorem, they related Ω_{I} to the ω-dependent conductivity in insulators (see Sect. 7.3.3).

7.1.1 Metals

The ground-state projector, Eq. (3.45), can be generalized to deal with the metallic case as well:

$$\mathcal{P}_{\mathbf{k}} = \sum_j \theta(\mu - \epsilon_{j\mathbf{k}})|u_{j\mathbf{k}}\rangle\langle u_{j\mathbf{k}}|, , \tag{7.3}$$

where μ is the Fermi level and $\epsilon_{j\mathbf{k}}$ are band energies. Like for insulators, $\mathcal{P}_{\mathbf{k}}$ is invariant for unitary transformations of the $|u_{j\mathbf{k}}\rangle$ at a given $\mathbf{k}$, although in the metallic case the number of occupied bands is $\mathbf{k}$-dependent.

If we try to adopt the expression of Eq. (3.47) for the metric-curvature tensor, we need to evaluate the $\mathbf{k}$-derivative of $\mathcal{P}_{\mathbf{k}}$:

$$\partial_\alpha \mathcal{P}_{\mathbf{k}} = - \sum_j \delta(\mu - \epsilon_{j\mathbf{k}})\partial_\alpha \epsilon_{j\mathbf{k}} \, |u_{j\mathbf{k}}\rangle\langle u_{j\mathbf{k}}| \tag{7.4}$$

$$+ \sum_j \theta(\mu - \epsilon_{j\mathbf{k}})(|u_{j\mathbf{k}}\rangle\langle \partial_\alpha u_{j\mathbf{k}}| + |\partial_\alpha u_{j\mathbf{k}}\rangle\langle u_{j\mathbf{k}}|).$$

The δ-like singularity at the Fermi surface vanishes in insulators; as for the remaining contributions, they are smooth in insulators and piecewise continuous in metals. The diagonal elements in Eq. (3.47) are thus highly singular, proportional to the *square* of a Dirac δ at the $\mathbf{k}$ vectors where a band crosses the Fermi level.

We may thus say that $\Omega_{\mathrm{I}} = \infty$ in the metallic case, while it is always finite for an insulator. This is just an anticipation of what will be the leitmotiv in the following of the present Chapter: a geometrical quantity, having the dimensions of a squared length, discriminates between insulators and metals.

7.2 Conductivity and Drude weight

7.2.1 Generalities

Misleading and/or incorrect beliefs about what the Drude weight is sometimes appear in the literature. The presentation given here is inspired by Refs. [152, 153, 154, 151].

The phenomenological definition of the insulating state is based on conductivity: a macroscopically homogeneous material is insulating whenever its dc longitudinal conductivity vanishes, i.e. when the real symmetric part of the conductivity tensor $\sigma_{\alpha\beta}^{(+)}(\omega)$ goes to zero for $\omega \to 0$.

Longitudinal conductivity is an intensive material property whose most general form can be written as

$$\sigma_{\alpha\beta}^{(+)}(\omega) = D_{\alpha\beta} \left[\delta(\omega) + \frac{i}{\pi\omega} \right] + \sigma_{\alpha\beta}^{(\mathrm{regular})}(\omega) = \sigma_{\alpha\beta}^{(\mathrm{Drude})}(\omega) + \sigma_{\alpha\beta}^{(\mathrm{regular})}(\omega), \qquad (7.5)$$

where the constant $D_{\alpha\beta}$ goes under the name of Drude weight. The Drude weight can be defined as [128]:

$$D_{\alpha\beta} = \pi \lim_{\omega \to 0} \omega \, \mathrm{Im} \, \sigma_{\alpha\beta}^{(+)}(\omega). \qquad (7.6)$$

The insulating behavior of a material implies that $D_{\alpha\beta} = 0$ and that the real symmetric part of $\sigma_{\alpha\beta}^{(\mathrm{regular})}(\omega)$ goes to zero for $\omega \to 0$ at zero temperature. A compact and meaningful way for discriminating between metals and insulators is therefore the Souza-Wilkens-Martin integral [147]

$$I_{\mathrm{SWM}} = \int_0^\infty \frac{d\omega}{\omega} \sum_{\alpha=1}^d \mathrm{Re} \, \sigma_{\alpha\alpha}(\omega), \qquad (7.7)$$

which diverges for all metals and converges for all insulators; more about I_{SWM} will be said in Sect. 7.3.3.

The conductivity obeys the the f-sum rule

$$\int_0^\infty d\omega \, \mathrm{Re} \, \sigma_{\alpha\alpha}(\omega) = \frac{D_{\alpha\alpha}}{2} + \int_0^\infty d\omega \, \mathrm{Re} \, \sigma_{\alpha\alpha}^{(\mathrm{regular})}(\omega) = \frac{\omega_{\mathrm{p}}^2}{8} = \frac{\pi e^2 n}{2m}, \qquad (7.8)$$

where n is the electron density and ω_{p} is the plasma frequency. For free electrons (a gas of noninteracting electrons in a flat potential) $\sigma_{\alpha\beta}^{(\mathrm{regular})}(\omega)$ vanishes, while $D_{\alpha\beta}$ assumes the same value as in classical physics [155], i.e $D_{\alpha\beta} = \pi e^2 (n/m)\delta_{\alpha\beta}$: this explains the extraordinary longevity of Drude theory, developed in the year 1900. Given Eq. (7.8), switching on the potential (one-body and two-body) has the effect of transferring some spectral weight from the Drude peak into the regular term; in the case of insulators, the Drude peak vanishes.

Dissipation can be included fenomenologically in the Drude term by adopting a single-relaxation-time approximation, exactly as in the classical textbook case [155], i.e.

$$\sigma_{\alpha\beta}^{(\mathrm{Drude})}(\omega) = \frac{i}{\pi}\frac{D_{\alpha\beta}}{\omega + i/\tau}, \tag{7.9}$$

whose $\tau \to \infty$ limit coincides with first term in the expression in Eq. (7.5).

In the special case of a band metal (i.e. a crystalline system of non interacting electrons) $\sigma_{\alpha\beta}^{(\mathrm{regular})}(\omega)$ is a linear-response property which accounts for interband transitions, and is nonvanishing only at frequencies higher than a finite threshold. In disordered and/or correlated systems instead the selection rule breaks down: the Drude weight may vanish, and $\sigma_{\alpha\beta}^{(\mathrm{regular})}(0)$ may be nonzero.

The Drude weight $D_{\alpha\beta}$ (also called charge stiffness) is *not* a linear-response property. It must be regarded as a ground-state property which accounts for the (inverse) inertia of the many-electron system in the adiabatic limit, and provides an effective value of n/m, where the free-electron value is modified by the one- and two-body potentials: this is clearly shown in Eqs. (7.30) and (7.72) below. As said above, the free-electron Drude weight is an upper limit for the actual value of $D_{\alpha\alpha}$. In the case of a band metal $D_{\alpha\beta}$ can be equivalently expressed as a Fermi-surface integral, by means of an integration by parts: it acquires then the meaning of an "intraband" term: see Eq. (7.31) below.

7.2.2 Kohn's expression for the Drude weight

In his milestone 1964 paper, Kohn proposes a very general expression for the Drude weight, which applies to disordered and/or correlated systems as well; it reduces to the standard expressions given e.g. in Ref. [153] in the special case of a band metal: see Eq. (7.27) below.

We address an interacting (and possibly disordered) N-electron system. We consider—following the milestone Kohn's paper [128]—the family of many-body Hamiltonians parametrized by $\boldsymbol{\kappa}$ as

$$\hat{H}_{\boldsymbol{\kappa}} = \frac{1}{2m}\sum_{i=1}^{N}|\mathbf{p}_i + \hbar\boldsymbol{\kappa}|^2 + \hat{V}; \tag{7.10}$$

102

the potential $\hat{V}$ includes one-body (possibly disordered) and two-body (electron-electron) contributions. Equation (7.10) is exact in the nonrelativistic, infinite-nuclear-mass limit. The vector $\boldsymbol{\kappa}$, having the dimensions of an inverse length, is called "flux" or "twist". Setting $\boldsymbol{\kappa} \neq 0$ amounts to a gauge transformation. The electrons are confined in a cubic box of volume L^d and the eigenstates $|\Psi_{n\boldsymbol{\kappa}}\rangle$ are normalized to one in the hypercube of volume L^{Nd}; we will adopt the simplifying notation $|\Psi_{n0}\rangle = |\Psi_n\rangle$. The same Hamiltonian has been addressed before in Sect. 5.4.2.

Bulk properties of condensed matter obtain from the thermodynamic limit: $N \to \infty$, $L \to \infty$, N/L^d constant. Since the following formulæ will comprise $\boldsymbol{\kappa}$-derivatives evaluated at $\boldsymbol{\kappa} = 0$, it is important to stress that the differentiation is performed first, and the thermodynamic limit afterwards.

Two kinds of boundary conditions can be adopted for the given Hamiltonian: either periodic (PBCs) or "open" (OBCs). The latter case is dealt with below in Sect. 7.4; here we adopt Born-von-Kàrmàn PBCs over each electron coordinate $\mathbf{r}_i$ independently, whose Cartesian components $r_{i,\alpha}$ are then equivalent to the angles $2\pi r_{i,\alpha}/L$. The potential $\hat{V}$ enjoys the same periodicity, which implies that the electric field averages to zero over the sample. As noticed by W. Kohn in 1964 [128], PBCs violate gauge invariance in the conventional sense: for instance, the ground state energy $E_{0\boldsymbol{\kappa}}$ actually *depends* on $\boldsymbol{\kappa}$ in metals, and in metals only. It has been show before, in Sect. 5.4.2, that $E_{0\boldsymbol{\kappa}} = E_0$ *does not* depend on $\boldsymbol{\kappa}$ in insulators.

Kohn's expression for the Drude weight is:

$$D_{\alpha\beta} = \frac{\pi e^2}{\hbar^2 L^d} \frac{\partial^2 E_{0\boldsymbol{\kappa}}}{\partial \kappa_\alpha \partial \kappa_\beta}\bigg|_{\boldsymbol{\kappa}=0}. \tag{7.11}$$

We remind that it is crucial to set $\boldsymbol{\kappa} = 0$ in the derivative *before* the thermodynamic limit is taken: this ensures that we are following the ground state adiabatically [152].

7.2.3 Kubo formulæ for conductivity

We remind the fundamentals of linear-response theory, presented in more detail in Appendix C; the general response function $\chi(\omega)$ is defined as

$$f_{\text{output}}(\omega) = \chi(\omega)\, f_{\text{input}}(\omega). \tag{7.12}$$

Within quantum mechanics at zero temperature, we define $\chi(t)$ by means of a perturbation in the Hamiltonian $\Delta\hat{H} = -\delta(t)\hat{A}$ (the "kick"), acting on the system in its ground state. The response is measured as the expectation value of another operator $\hat{B}$. Without loss of generality we simplify notations by assuming that

$$\langle\Psi_0|\,\hat{A}\,|\Psi_0\rangle = 0, \quad \langle\Psi_0|\,\hat{B}\,|\Psi_0\rangle = 0. \tag{7.13}$$

Time-dependent perturbation theory leads to the Kubo formula for the generalized susceptibility, which we write in the ω domain by adopting the compact notations due to Zubarev [67, 156, 157]:

$$\chi(\omega) = -\langle\langle \hat{B}|\hat{A}\rangle\rangle_\omega; \tag{7.14}$$

$$\langle\langle \hat{B}|\hat{A}\rangle\rangle_\omega = \frac{1}{\hbar}\lim_{\eta\to 0^+}\sideset{}{'}\sum_{n\neq 0}\left(\frac{\langle\Psi_0|\hat{B}|\Psi_n\rangle\langle\Psi_n|\hat{A}|\Psi_0\rangle}{\omega - \omega_{0n} + i\eta} - \frac{\langle\Psi_0|\hat{A}|\Psi_n\rangle\langle\Psi_n|\hat{B}|\Psi_0\rangle}{\omega + \omega_{0n} + i\eta}\right). \tag{7.15}$$

The positive infinitesimal η ensures causality, and we remind that

$$\lim_{\eta\to 0^+}\frac{1}{\omega \pm i\eta} = \mathscr{P}\frac{1}{\omega} \mp i\pi\delta(\omega), \tag{7.16}$$

where $\mathscr{P}$ indicates the principal part. We draw attention to the fact that the sign conventions adopted in this work agree with Zubarev [67, 156] and Chandler [158], but are opposite the the ones of McWeeny [157] and other textbooks.

In order to address conductivity it is essential to adopt PBCs: there cannot be any steady state current in a bounded sample within OBCs. Furthermore, since the multiplicative position $\hat{\mathbf{r}}$ is not a legitimate operator within PBCs [103], it is mandatory to adopt the vector-potential gauge for the macroscopic electric field $\boldsymbol{\mathcal{E}}$: the perturbation in the Hamiltonian is therefore an ω-dependent vector potential $\delta\mathbf{A}$, constant in space.

The current carried by a generic state $|\Psi\rangle$ after the perturbation is switched on is therefore

$$\mathbf{j} = -\frac{e}{L^d}\langle\Psi|\,\hat{\mathbf{v}}\,|\Psi\rangle - \frac{e^2 N}{mcL^d}\delta\mathbf{A}, \tag{7.17}$$

where $\hat{\mathbf{v}}$ is the many-body velocity, Eq. (1.18), at $\boldsymbol{\kappa} = 0$; the two terms in Eq. (7.17) are generally called "paramagnetic" and "diamagnetic", respectively. Expansion of the Hamiltonian to first order in the perturbing vector potential $\delta\mathbf{A}$ yields

$$\Delta\hat{H} = \frac{e}{c}\delta\mathbf{A}\cdot\hat{\mathbf{v}}. \tag{7.18}$$

If we set $\boldsymbol{\mathcal{E}}$ and $\delta\mathbf{A}$ along the β direction, the linearly induced current in the α direction is

$$\begin{aligned} j_\alpha &= -\frac{e^2 N}{mcL^d}\delta A\,\delta_{\alpha\beta} - \frac{e}{L^d}\langle\langle\,\hat{v}_\alpha\,|\,\frac{e}{c}\delta A\,\hat{v}_\beta\rangle\rangle_\omega\\ &= -\frac{e^2}{cL^d}\left(\frac{N}{m}\delta_{\alpha\beta} + \langle\langle\hat{v}_\alpha|\hat{v}_\beta\rangle\rangle_\omega\right)\delta A(\omega), \end{aligned} \tag{7.19}$$

where we are restoring the ω dependence. The term in δA^2, being constant in space, has zero matrix elements; it is also second order in $\boldsymbol{\mathcal{E}}$.

In order to arrive at conductivity we need to express $\delta A(\omega)$ in Eq. (7.19) in terms of $\mathcal{E}(\omega)$. In the time domain their relationship is $\mathcal{E} = -\frac{1}{c}\partial \delta A/\partial t$; a naive integration would yield $\delta A(\omega) = -ic\mathcal{E}(\omega)/\omega$, but this violates causality. The correct integration yields:

$$\delta A(\omega) = -\frac{ic\mathcal{E}(\omega)}{\omega + i\eta} = c\mathcal{E}(\omega)\left[\frac{1}{i\omega} - \pi\delta(\omega)\right]. \tag{7.20}$$

Therefore the current, as expressed directly in terms of the field intensity, is

$$j_\alpha(\omega) = \sigma_{\alpha\beta}(\omega)\,\mathcal{E}_\beta(\omega) = -\frac{e^2}{L^d}\left(\frac{N}{m}\delta_{\alpha\beta} + \langle\langle\hat{v}_\alpha|\hat{v}_\beta\rangle\rangle_\omega\right)\left[\frac{1}{i\omega} - \pi\delta(\omega)\right]\mathcal{E}_\beta(\omega). \tag{7.21}$$

We then write the Kubo formula as

$$\langle\langle v_\alpha|v_\beta\rangle\rangle_\omega = \frac{1}{\hbar}\lim_{\eta\to 0^+}{\sum_{n\neq 0}}'\left(\frac{\mathcal{R}_{n,\alpha\beta} + i\,\mathcal{I}_{n,\alpha\beta}}{\omega - \omega_{0n} + i\eta} - \frac{\mathcal{R}_{n,\alpha\beta} - i\,\mathcal{I}_{n,\alpha\beta}}{\omega + \omega_{0n} + i\eta}\right), \tag{7.22}$$

$$\mathcal{R}_{n,\alpha\beta} = \mathrm{Re}\,\langle\Psi_0|\hat{v}_\alpha|\Psi_n\rangle\langle\Psi_n|\hat{v}_\beta|\Psi_0\rangle, \quad \mathcal{I}_{n,\alpha\beta} = \mathrm{Im}\,\langle\Psi_0|\hat{v}_\alpha|\Psi_n\rangle\langle\Psi_n|\hat{v}_\beta|\Psi_0\rangle, \tag{7.23}$$

where $\mathcal{R}_{n,\alpha\beta}$ is symmetric and $\mathcal{I}_{n,\alpha\beta}$ antisymmetric. The longitudinal conductivity is the symmetric part $\sigma_{\alpha\beta}^{(+)}(\omega)$ of the tensor. Upon exploiting Eq. (7.16) we eventually get

$$D_{\alpha\beta} = \frac{\pi e^2}{L^d}\left(\frac{N}{m}\delta_{\alpha\beta} - \frac{2}{\hbar}{\sum_{n\neq 0}}'\frac{\mathcal{R}_{n,\alpha\beta}}{\omega_{0n}}\right), \tag{7.24}$$

$$\mathrm{Re}\,\sigma_{\alpha\beta}^{(\mathrm{regular})}(\omega) = \frac{\pi e^2}{\hbar L^d}{\sum_{n\neq 0}}'\frac{\mathcal{R}_{n,\alpha\beta}}{\omega_{0n}}[\,\delta(\omega - \omega_{0n}) + \delta(\omega + \omega_{0n})\,], \tag{7.25}$$

$$\mathrm{Im}\,\sigma_{\alpha\beta}^{(\mathrm{regular})}(\omega) = \frac{2e^2}{\hbar L^d}{\sum_{n\neq 0}}'\frac{\mathcal{R}_{n,\alpha\beta}}{\omega_{0n}}\frac{\omega}{\omega_{0n}^2 - \omega^2}. \tag{7.26}$$

A straightforward calculations proves that Eq. (7.24) is indeed equivalent to Kohn's expression for the Drude weight, Eq. (7.11). The second term in parenthesis is the correction to the free-electron Drude weight due to the one- and two-body potential $\hat{V}$; in the case of a band metal Eq. (7.24) reduces to Eq. (7.30).

It is trivial to verify that Eqs. (7.24) and (7.25) obey the f-sum rule, Eq. (7.8); the two terms $\sigma_{\alpha\beta}^{(\mathrm{Drude})}(\omega)$ and $\sigma_{\alpha\beta}^{(\mathrm{regular})}(\omega)$ obey the Kramers-Kronig relationships *separately*; we also remind that only longitudinal conductivity $\sigma_{\alpha\beta}^{(+)}(\omega)$ is addressed for the time being; transverse conductivity $\sigma_{\alpha\beta}^{(-)}$ will be addressed in Ch. 8.

At any finite size L the spectrum is discrete and the system is gapped, while in a metal the gap closes in the large-L limit. It is therefore necessary to regularize

the singular sums in Eqs. (7.25) and (7.26); this can be done in the following way [159]. One starts assuming a finite value of η in the Kubo formula, Eq. (7.15), with η much larger than the level spacing; then one takes the $L \to \infty$ limit first, and the $\eta \to 0^+$ limit afterwards. There is no need of regularizing in order to adddress sum rules, like the f-sum rule and the SWM sum rule (discussed below), because they are integrated properties.

The basic expressions of of Eqs. (7.24) and (7.25) apply to both metals and insulators. In the latter case the Drude weight vanishes, and $\sigma_{\alpha\beta}^{(+)}(\omega) = \sigma_{\alpha\beta}^{(\text{regular})}(\omega)$; for a gapped insulator $\sigma_{\alpha\beta}^{(+)}(\omega)$ vanishes for $\omega < \epsilon_{\text{gap}}/\hbar$.

7.2.4 Semiclassical theory of electron transport

In the simple case of a band metal (with double band occupancy) Eq. (7.11) becomes the Brillouin-zone (BZ) integral [153]:

$$D_{\alpha\beta} = 2\pi e^2 \sum_j \int_{\text{BZ}} \frac{d\mathbf{k}}{(2\pi)^d} \, \theta(\mu - \epsilon_{j\mathbf{k}}) \, m_{j,\alpha\beta}^{-1}(\mathbf{k}), \tag{7.27}$$

where μ is the Fermi level, $\epsilon_{j\mathbf{k}}$ are band energies, and the effective inverse mass tensor of band j is

$$m_{j,\alpha\beta}^{-1}(\mathbf{k}) = \frac{1}{\hbar^2} \frac{\partial^2 \epsilon_{j\mathbf{k}}}{\partial k_\alpha \partial k_\beta}. \tag{7.28}$$

There is an alternative and very meaningful expression equivalent to Eq. (7.28), where the deviation of $m_{j,\alpha\beta}^{-1}(\mathbf{k})$ from its free-electron value appears as a geometrical term. To the best of the author's knowledge, this expression is due to Gao, Yang, and Niu in a 2015 paper [160]. We remind that the orbitals $|u_{j\mathbf{k}}\rangle$ are eigenstates of $\mathcal{H}_\mathbf{k} = e^{-i\mathbf{k}\cdot\mathbf{r}} \mathcal{H} e^{i\mathbf{k}\cdot\mathbf{r}}$, hence the identity $\langle u_{j\mathbf{k}} | (\mathcal{H}_\mathbf{k} - \epsilon_{j\mathbf{k}}) | u_{j\mathbf{k}} \rangle \equiv 0$ holds. Taking two derivatives, one arrives at

$$m_{j,\alpha\beta}^{-1}(\mathbf{k}) = \frac{1}{m}\delta_{\alpha\beta} - \frac{2}{\hbar^2}\text{Re} \, \langle \partial_\alpha u_{j\mathbf{k}} | (\mathcal{H}_\mathbf{k} - \epsilon_{j\mathbf{k}}) | \partial_\beta u_{j\mathbf{k}} \rangle, \tag{7.29}$$

$$D_{\alpha\beta} = \pi e^2 \frac{n}{m}\delta_{\alpha\beta} - \frac{4\pi e^2}{\hbar^2} \sum_j \int_{\text{BZ}} \frac{d\mathbf{k}}{(2\pi)^d} \, \theta(\mu - \epsilon_{j\mathbf{k}}) \, \text{Re} \, \langle \partial_\alpha u_{j\mathbf{k}} | (\mathcal{H}_\mathbf{k} - \epsilon_{j\mathbf{k}}) | \partial_\beta u_{j\mathbf{k}} \rangle :$$

$$\tag{7.30}$$

the first term on the r.h.s. is the free-electron Drude weight, while the second one is the geometrical correction due to the crystalline potential. For free electrons the $|u_{j\mathbf{k}}\rangle$ are $\mathbf{k}$-independent, ergo the correction vanishes.

For insulators, the BZ integral in Eq. (7.27) trivially vanishes. As anticipated above, the Drude weight of a band metal can be equivalently expressed as a Fermi-surface integral, by means of an integration by parts: it acquires then the meaning

of an "intraband" term [153]:

$$D_{\alpha\beta} = -2\pi e^2 \sum_j \int_{\mathrm{BZ}} \frac{d\mathbf{k}}{(2\pi)^d} \, f'(\epsilon_{j\mathbf{k}}) \, v_{j\alpha}(\mathbf{k}) v_{j\beta}(\mathbf{k}), \qquad v_{j\alpha}(\mathbf{k}) = \frac{1}{\hbar} \frac{\partial \epsilon_{j\mathbf{k}}}{\partial k_\alpha} \qquad (7.31)$$

where at zero temperature the Fermi occupation function is $f(\epsilon) = \theta(\mu - \epsilon)$. The formula as given is for a simple Fermi surface; in general there may be multiple bands that cross the Fermi level and Fermi surfaces having complex topology, for example including disconnected sections. The integration by parts requires then some care (see e.g. Ref. [161]).

The Fermi-surface expression shows explicit agreement with the spirit of Landau's Fermi-liquid theory, which holds that charge transport in metals involves only quasiparticles with energies within $k_{\mathrm{B}}T$ of the Fermi level. Indeed, Eq. (7.31) can be endowed with a relaxation time $\tau(\epsilon)$, yielding the same formula for dc conductivity as given by the popular Ashcroft-Mermin textbook [155]:

$$\mathrm{Re}\ \sigma_{\alpha\beta}^{(+)}(0) = -2e^2 \sum_j \int_{\mathrm{BZ}} \frac{d\mathbf{k}}{(2\pi)^d} \, f'(\epsilon_{j\mathbf{k}}) \, \tau(\epsilon_{j\mathbf{k}}) \, v_{j\alpha}(\mathbf{k}) v_{j\beta}(\mathbf{k}). \qquad (7.32)$$

Notice that the textbook formula is derived in a purely semiclassical way; here we have proved that it coincides with the exact one (obviously, for band metals only).

7.2.5 Adiabatic vs. nonadiabatic inertia of the many-electron system

It is immediate to retrieve from Eq. (7.28) the inverse of the electron mass m in the special case of free electrons. In the general case Eqs. (7.11) and (7.27) still measure the inertia of the many-electron system, albeit in the adiabatic limit only ($D_{\alpha\beta}$ is an $\omega \to 0$ property). It is interesting to compare the adiabatic inertia of the many-electron system to the nonadiabatic one in response to an instantaneous "kick": the two quantities only coincide in the free-electron case.

It is expedient to introduce an "effective" value of n/m, by means of the relationship

$$D_{\alpha\beta} = \pi e^2 (n/m)_{\alpha\beta}^*, \qquad (7.33)$$

where we remind that for free electrons $D_{\alpha\beta} = \pi e^2 (n/m) \delta_{\alpha\beta}$, hence $(n/m)_{\alpha\alpha}^* \leq n/m$.

Conductivity is the current response to a macroscopic field:

$$j_\alpha(\omega) = \sigma_{\alpha\beta}(\omega) \mathcal{E}_\beta(\omega), \qquad (7.34)$$

where the sum on the repeated indices is understood. Suppose we apply an instantaneous pulse $\boldsymbol{\mathcal{E}}(t) = \tilde{\boldsymbol{\mathcal{E}}} \, \delta(t)$, where $\tilde{\boldsymbol{\mathcal{E}}}$ has the dimensions of a field times a time.

We start replacing the full $\sigma_{\alpha\beta}(\omega)$ in Eq. (7.34) with $\sigma_{\alpha\beta}^{\text{(Drude)}}(\omega)$; in the simple case of a band metal this amounts to neglecting the interband transitions. The response is therefore identical to the one of a free-electron gas of density $n_{\alpha\beta}^*$, ergo the current at $t = 0^+$ is

$$j_\alpha^{\text{(Drude)}}(0^+) = e^2 (n/m)_{\alpha\beta}^* \, \tilde{\mathcal{E}}_\beta, \qquad (7.35)$$

and in the single-relaxation-time approximation, the longitudinal steady-state current is

$$j_\alpha = \sigma_{\alpha\beta}^{\text{(Drude)}}(0) \, \mathcal{E}_\beta = \tau e^2 (n/m)_{\alpha\beta}^* \, \mathcal{E}_\beta. \qquad (7.36)$$

The analogy with the textbook classical case [155] is perspicuous.

Next we are going to study the full nonadiabatic response to the instantaneous pulse, using the full $\sigma_{\alpha\beta}(\omega)$. The Fourier transform of the pulse $\tilde{\mathcal{E}} \, \delta(t)$ is $\tilde{\mathcal{E}}$ (ω-independent), hence the linearly induced current is

$$\begin{aligned} j_\alpha(\omega) &= \sigma_{\alpha\beta}(\omega) \, \tilde{\mathcal{E}}_\beta, \\ j_\alpha(t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega \, \sigma_{\alpha\beta}(\omega) \, \tilde{\mathcal{E}}_\beta \, \mathrm{e}^{-i\omega t}. \end{aligned} \qquad (7.37)$$

The evaluation of the integral is a bit tricky. Since $\sigma(\omega)$ is causal, $j_\alpha(t) = 0$ for $t < 0$, hence we may replace $j_\alpha(t) = [j_\alpha(t) + j_\alpha(-t)] \, \theta(t)$. It then follows

$$j_\alpha(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega \, \sigma_{\alpha\beta}(\omega)(\mathrm{e}^{i\omega t} + \mathrm{e}^{-i\omega t}) \, \tilde{\mathcal{E}}_\beta, \qquad t > 0. \qquad (7.38)$$

Since the imaginary part of $\sigma(\omega)$ is odd, we have

$$j_\alpha(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} d\omega \, \mathrm{Re} \, \sigma_{\alpha\beta}(\omega) \mathrm{e}^{i\omega t} \, \tilde{\mathcal{E}}_\beta, \qquad t > 0. \qquad (7.39)$$

Next we exploit the f-sum rule, Eq. (7.8):

$$\int_{-\infty}^{\infty} d\omega \, \mathrm{Re} \, \sigma_{\alpha\beta}(\omega) = \delta_{\alpha\beta} \, \omega_{\mathrm{p}}/4 = \pi e^2 (n/m) \delta_{\alpha\beta} \qquad (7.40)$$

$$j_\alpha(0^+) = e^2 (n/m) \, \mathcal{E}_\alpha. \qquad (7.41)$$

This is similar in form to Eq. (7.35), but here $n = N/V$ includes all the electrons, even the core ones (provided the integral includes ultraviolet and x-ray regions of the spectrum). Notice that, instead, $(n/m)_{\alpha\beta}^*$ in Eq. (7.35) only includes valence electrons: in fact completely filled bands do not contribute to Eq. (7.27).

The full response in Eq. (7.41) is identical to the classical response of a system of free electrons. And in fact at $t = 0^+$ the electrons have not yet responded to anything other than the instantaneous pulse [153], which probes the bare inertia of the many-electron system. As said above, only the ground-state adiabatic contribution is included in Eq. (7.35), while instead Eq. (7.41) includes coupling to the excited states, as required by time-dependent perturbation theory.

7.2.6 The insulating state according to Resta and Sorella

As discussed above, the main observable that discriminates an insulator from a metal is the vanishing vs. nonvanishing of dc conductivity. Soon after the development of the modern theory of polarization in the 1990s, Resta and Sorella [146], hereafter quoted as RS, realized that besides conductivity, another observable sharply characterizes the insulating state: macroscopic polarization **P**. In fact in a metal **P** is trivial and material-independent: it completely screens any static field (Faraday-cage effect). In the insulating case, instead, **P** is nontrivial and material-dependent, both with and without an applied field. By the late 1990s all the mysteries and misconceptions about polarization were unveiled and a change of paradigm occurred; the same change of paradigm could also provide a clue on the genuine nature of the insulating state.

The ultimate formulation of the theory of polarization appeared in 1998 [103], where the electronic term in **P** was cast in term of a single-point Berry phase; the expression for a quasi-1d system of N electrons in a periodic box (along x) of length L was given above. We reproduce it here for the sake of clarity:

$$\gamma^{(\text{el})} = \text{Im} \ln \mathfrak{z}_N, \quad \mathfrak{z}_N = \langle \Psi_0 | \, e^{-i\frac{2\pi}{L}\sum_j x_j} \, | \Psi_0 \rangle, . \tag{7.42}$$

The ground state $|\Psi_0\rangle$ is very general: it may include disorder and correlation. The electronic Berry phase, Eq. (7.42), is well defined insofar as the complex number $\mathfrak{z}_N$ does not vanish in the large-N limit (at constant N/L). In the original paper RS postulated that the vanishing/nonvanishing of $\mathfrak{z}_N$ characterizes the metallic/insulating state, on the ground—as said above—that bulk macroscopic polarization is ill defined in metals and well defined in insulators.

In his 1964 paper W. Kohn stated verbatim that "insulating characteristics are a strict consequence of electron localization (in an appropriate sense) and do not require an energy gap"; he did not, however, provide a practical tool to actually measure such localization in a given many-body wavefunction. Kohn's paper remained somewhat neglected and little cited for many years. In 1999 RS proposed to exploit the complex number $\mathfrak{z}_N$ in Eq. (7.42), whose modulus is no greater than one, to define a "localization length" λ, in the spirit of Kohn's vision, via the relationship

$$\lambda^2 = -\frac{1}{4\pi^2} \lim_{N \to \infty} \frac{L^2}{N} \ln |\mathfrak{z}_N|^2, \tag{7.43}$$

which was postulated to diverge in any metal and to converge in any insulator. In agreement with Kohn's pioneering viewpoint, localization is at the root of the insulating behavior. Several arguments were provided in support of our original postulate (some of them detailed below), although none of these arguments was actually related to conductivity. The original paper also conjectured that the approach can be extended to 2d and 3d systems, where λ^2 becomes a Cartesian

tensor. Later, Souza, Wilkens, and Martin [147] have shown that the RS localization tensor has the meaning of second cumulant moment of the electron distribution (see Sect. 7.3). Because of this, the notation $\langle r_\alpha r_\beta \rangle_c$ is adopted for the localization tensor in most of the subsequent literature.

In the isotropic case Eq. (7.43) holds as it stands, where L is the linear dimension of the system. As anticipated above in the special case of an isotropic band insulator in dimension d one has $\lambda^2 = \Omega_I/(n_b d)$, where n_b is the number of occupied bands and Ω_I is the MV gauge-invariant quadratic spread, Eq. (7.2).

We are going to provide next the link between the RS formula and conductivity. It has been proved in Sect. 5.4.2 that in a crystalline system (possibly correlated) the electronic Berry phase yielding the electronic polarization $\mathbf{P}^{(el)}$ has the same expression as in Eq. (7.42) even for $2d$ and $3d$ systems, provided the $|\Psi_0\rangle$ therein is the ground eigenstate of Eq. (7.10) at $\boldsymbol{\kappa} = 0$, periodic over a cubic box of side L. The d-dependent prefactor has been discussed above [102, 127].

We consider again the many body Hamiltonian of Eq. (7.10) within PBCs over a cubic box of volume L^d. As said above, the ground eigenstate $|\Psi_{\boldsymbol{\kappa}}\rangle$ has a nontrivial $\boldsymbol{\kappa}$-dependence: we consider a special value $\boldsymbol{\kappa}_1 = \frac{2\pi}{L}\mathbf{e}_1$, where $\mathbf{e}_1$ is the unit vector in the x-direction. For this special $\boldsymbol{\kappa}_1$ the effect of the gauge can be gauged away; in fact the state vector

$$|\Phi_{0\boldsymbol{\kappa}_1}\rangle = e^{-i\boldsymbol{\kappa}_1 \cdot \hat{\mathbf{r}}}|\Psi_0\rangle \tag{7.44}$$

obeys PBCs, and is and eigenstate of $\hat{H}_{\boldsymbol{\kappa}_1}$ with eigenvalue E_0, similarly to the OBCs case. We remind that $\hat{\mathbf{r}} = \sum_i \mathbf{r}_i$. Now the issue is whether $|\Phi_{0\boldsymbol{\kappa}_1}\rangle$ coincides or not with the genuine $|\Psi_{0\boldsymbol{\kappa}_1}\rangle$, obtained by following the ground state adiabatically while $\boldsymbol{\kappa}$ is switched on continuously. The issue has been discussed already en passant, in Sect. 5.4.2.

We assume an isotropic system, where $D_{\alpha\beta} = D\,\delta_{\alpha\beta}$. Eq. (7.11) shows that whenever the Drude weight D is nonzero the state $|\Psi_{0\boldsymbol{\kappa}_1}\rangle$ has an energy higher than E_0: it is therefore an excited eigenstate of $\hat{H}_{\boldsymbol{\kappa}_1}$, orthogonal to $|\Phi_{0\boldsymbol{\kappa}_1}\rangle$. If instead $D = 0$, then the state $|\Psi_{0\boldsymbol{\kappa}_1}\rangle$ coincides—apart for a phase factor—with $|\Phi_{0\boldsymbol{\kappa}_1}\rangle$ (we are assuming a nondegenerate ground state):

$$\langle \Phi_{0\boldsymbol{\kappa}_1}|\Psi_{0\boldsymbol{\kappa}_1}\rangle = \langle\Psi_0|\,e^{i\boldsymbol{\kappa}_1\cdot\hat{\mathbf{r}}}\,|\Psi_{0\boldsymbol{\kappa}_1}\rangle = 0, \quad D \neq 0, \tag{7.45}$$

$$\langle \Phi_{0\boldsymbol{\kappa}_1}|\Psi_{0\boldsymbol{\kappa}_1}\rangle = \langle\Psi_0|\,e^{i\boldsymbol{\kappa}_1\cdot\hat{\mathbf{r}}}\,|\Psi_{0\boldsymbol{\kappa}_1}\rangle = e^{-i\gamma}, \quad D = 0. \tag{7.46}$$

We only address the modulus of the matrix elements in the following. To leading order in $1/L$, we may replace $|\Psi_{0\boldsymbol{\kappa}_1}\rangle$ with $|\Psi_0\rangle$, hence:

$$|\mathfrak{z}_N| = |\langle\Psi_0|\,e^{i\boldsymbol{\kappa}_1\cdot\hat{\mathbf{r}}}\,|\Psi_0\rangle| \to 0, \quad D \neq 0, \tag{7.47}$$

$$|\mathfrak{z}_N| = |\langle\Psi_0|\,e^{i\boldsymbol{\kappa}_1\cdot\hat{\mathbf{r}}}\,|\Psi_0\rangle| \to 1, \quad D = 0. \tag{7.48}$$

Therefore in the large-L limit $|\mathfrak{z}_N|$ vanishes whenever the Drude weight is nonzero, while it converges to 1 from below in insulators. The RS expression of Eq. (7.43),

having the dimensions of a squared length, is positive and finite in insulators in the large-system limit. According to RS the dimensionless quantity $N^{2/d-1} \ln |\mathfrak{z}_N|^2$ converges (in insulators) to a finite limit, ergo λ^2, Eq. (7.43), is intensive in any dimension d. In the present review we prove the intensiveness of λ^2 (in any d) in the special case of a band insulator, Eq. (7.55) below. In the general case, the intensiveness (again in any dimension d) follows from the equivalence to the many-body metric, Eq. (7.65) below. By similar considerations, it is easy to see that λ^2 instead diverges whenever $D > 0$.

7.2.7 Independent electrons

A crystalline system of independent electrons is either a band metal or a band insulator: in both cases $|\Psi_0\rangle$ is a Slater determinant of Bloch orbitals. The RS square length λ^2 discriminates very sharply between the two cases [21, 146]. We proceed in analogy to our previous treatment of quasi-$1d$ insulators, for the a simple cubic lattice (with no loss of generality). Let a be the lattice constant, $L = Ma$ the Born-von-Kàrmàn periodicity, and μ the Fermi level; $|\psi_{m\mathbf{k}}\rangle$ are the Bloch orbitals with energy $\epsilon_{m\mathbf{k}}$. Our main ingredient will be is the overlap matrix of Eq. (3.40), also called "connection matrix", which we rewrite here for the sake of convenience

$$S_{mm'}(\mathbf{k}, \mathbf{k}') = \langle u_{m\mathbf{k}} | u_{m'\mathbf{k}'} \rangle. \tag{7.49}$$

The d-dimensional analogue of Eq. (3.55) reads

$$|\Psi_0\rangle = \frac{1}{\sqrt{M^{Nd}}} \, \mathsf{A} \prod_{\epsilon_{m\mathbf{k}_s} < \mu} \prod_{s=1}^{M^d} \psi_{m\mathbf{k}_s}^\uparrow \psi_{m\mathbf{k}_s}^\downarrow, \tag{7.50}$$

where now s is a vector index running on a regular grid of integers; in $3d$ the grid is:

$$s \equiv (s_1, s_2, s_3), \quad s_1, s_2, s_3 = 1, 2 \ldots M, \quad \mathbf{k}_s = \frac{2\pi}{L}(s_1 \mathsf{e}_1 + s_2 \mathsf{e}_2 + s_3 \mathsf{e}_3). \tag{7.51}$$

The $\sqrt{1/M^{Nd}}$ factor in Eq. (7.50) owes to the different normalizations: $|\Psi_0\rangle$ is normalized over the Born-von-Kàrmàn cell of volume L^d, while the Bloch orbitals $|\psi_{m\mathbf{k}}\rangle$ are normalized over the crystal cell of volume a^3. Using then spin factorization and the sparseness of the overlap matrix we arrive, *in the insulating case*, at

$$\mathfrak{z}_N^* = \langle \Psi_0 | \, e^{i\boldsymbol{\kappa}_1 \cdot \hat{\mathbf{r}}} \, | \Psi_0 \rangle = \prod_{s=1}^{M^d} [\det S(\mathbf{k}_s, \mathbf{k}_s + \boldsymbol{\kappa}_1)]^2, \tag{7.52}$$

where the $1/M^{Nd}$ factor has disappeared.

Instead, it is rather simple to prove that $\mathfrak{z}_N$ vanishes in the metallic case, and therefore $\lambda^2 = \infty$ in a band metal, even at *finite* N [21, 146]. Instead in presence of interaction and/or disorder the selection rule breaks down. The RS localization length λ^2 diverges only asymptotically. This is shown very perspicuously in the paradigmatic case case of the Mott metal-insulator transition discussed below.

Coming back to the insulating case, Eq. (7.52) implies

$$\ln |\mathfrak{z}_N|^2 = 2 \ln \prod_{s=1}^{M^d} \det S(\mathbf{k}_s, \mathbf{k}_s - \boldsymbol{\kappa}_1)\, S(\mathbf{k}_s, \mathbf{k}_s + \boldsymbol{\kappa}_1), \qquad (7.53)$$

and the RS localization length becomes

$$\lambda^2 = -\frac{L^2}{4\pi^2 N} \ln |\mathfrak{z}_N|^2 = -\frac{2L^2}{4\pi^2 N} \ln \prod_{s=1}^{M^d} \det S(\mathbf{k}_s, \mathbf{k}_s - \boldsymbol{\kappa}_1)\, S(\mathbf{k}_s, \mathbf{k}_s + \boldsymbol{\kappa}_1), \quad (7.54)$$

where the large-M limit is understood.

At this point we make contact with the quantum metric and the gauge-invariant quadratic spread. Starting from Eq. (7.54), one arrives at:

$$\lambda^2 = \frac{a^d}{n_{\mathrm{b}}} \int_{\mathrm{BZ}} \frac{d\mathbf{k}}{(2\pi)^d}\, g_{xx}(\mathbf{k}) = \frac{\Omega_{\mathrm{I}}}{n_{\mathrm{b}} d}; \qquad (7.55)$$

the proof is given in Ref. [21]

7.3 Geometry of the many-body ground state

We generalize a little bit Kohn's Hamiltonian as

$$\hat{H}_{\boldsymbol{\kappa}} = \frac{1}{2m} \sum_{i-1}^{N} \left[\mathbf{p}_i + \frac{e}{c} \mathbf{A}(\mathbf{r}_i) + \hbar \boldsymbol{\kappa} \right]^2 + \hat{V}, \qquad (7.56)$$

where the vector potential $\mathbf{A}(\mathbf{r})$ summarizes all T-breaking terms, as e.g. those due to spin-orbit coupling to a background of local moments; furthermore in order to simplify notations we will set in the following $\hat{H}_0 \equiv \hat{H}$, $|\Psi_{n0}\rangle \equiv |\Psi_n\rangle$, $E_{n0} \equiv E_n$.

If the state vector is a differentiable function of the twist $\boldsymbol{\kappa}$, then the differential phase and the differential distance define the Berry connection and the quantum metric, respectively:

$$\varphi_{\boldsymbol{\kappa},\boldsymbol{\kappa}+d\boldsymbol{\kappa}} = \mathcal{A}_\alpha(\boldsymbol{\kappa}) d\kappa_\alpha, \qquad\qquad D^2_{\boldsymbol{\kappa},\boldsymbol{\kappa}+d\boldsymbol{\kappa}} = g_{\alpha\beta}(\boldsymbol{\kappa}) d\kappa_\alpha d\kappa_\beta, \qquad (7.57)$$

$$\mathcal{A}_\alpha(\boldsymbol{\kappa}) = i\langle \Psi_{\boldsymbol{\kappa}} | \partial_{\kappa_\alpha} \Psi_{\boldsymbol{\kappa}} \rangle, \quad g_{\alpha\beta}(\boldsymbol{\kappa}) = \mathrm{Re}\, \langle \partial_{\kappa_\alpha} \Psi_{\boldsymbol{\kappa}} | \partial_{\kappa_\beta} \Psi_{\boldsymbol{\kappa}} \rangle - \langle \partial_{\kappa_\alpha} \Psi_{\boldsymbol{\kappa}} | \Psi_{\boldsymbol{\kappa}} \rangle \langle \Psi_{\boldsymbol{\kappa}} | \partial_{\kappa_\beta} \Psi_{\boldsymbol{\kappa}} \rangle;$$

summation over repeated Cartesian indices is understood (here and throughout). The Berry curvature is defined as the curl of the connection:

$$\tilde{\mathbf{\Omega}}_{\alpha\beta}(\boldsymbol{\kappa})d\kappa_\alpha d\kappa_\beta = [\partial_{\kappa_\alpha}\mathcal{A}_\beta(\boldsymbol{\kappa}) - \partial_{\kappa_\beta}\mathcal{A}_\alpha(\boldsymbol{\kappa})]d\kappa_\alpha d\kappa_\beta = -2\,\mathrm{Im}\langle\partial_\alpha\Psi_{\boldsymbol{\kappa}}|\partial_\beta\Psi_{\boldsymbol{\kappa}}\rangle d\kappa_\alpha d\kappa_\beta. \tag{7.58}$$

All of the above forms are extensive.

The connection is a 1-form and is gauge-dependent; the metric and the curvature are 2-forms and are gauge invariant. The above fundamental quantities are defined in terms of the state vectors solely; we will also address a 2-form which involves the Hamiltonian as well. Suppose that H is the Hamiltonian and E_0 its ground eigenvalue: we will consider

$$\mathcal{G} = \langle\Psi|\,(H - E_0)\,|\Psi\rangle, \tag{7.59}$$

which vanishes for $|\Psi\rangle = |\Psi_0\rangle$; an essential feature of $\mathcal{G}$ is that it is invariant by translation of the energy zero. The geometrical quantity of interest is the gauge-invariant 2-form which obtains by varying $|\Psi\rangle$ in the neighborhood of $|\Psi_0\rangle$.

7.3.1 Metric and the Resta-Sorella theory

We start defining the metric per electron as

$$\mathfrak{g}_{\alpha\beta}(N) = \frac{1}{N}(\ \mathrm{Re}\ \langle\partial_{\kappa_\alpha}\Psi_0|\partial_{\kappa_\beta}\Psi_0\rangle - \langle\partial_{\kappa_\alpha}\Psi_0|\Psi_0\rangle\langle\Psi_0|\partial_{\kappa_\beta}\Psi_0\rangle\ );. \tag{7.60}$$

For the sake of clarity, we make contact with the notations adopted in most of the previous literature. The RS localization tensor, a.k.a. second cumulant moment of the electron distribution is

$$\langle r_\alpha r_\beta\rangle_\mathrm{c} = \lim_{N\to\infty}\mathfrak{g}_{\alpha\beta}(N). \tag{7.61}$$

We are addressing this limit next: Eq. (7.65).

Notice that in Eq. (7.60)—as recommended by Kohn [128]—the number of electrons N in $|\Psi_0\rangle$ is kept constant during the differentiation, and the $N \to \infty$ limit is taken afterwards. Instead in the RS limiting process N varies while $|\boldsymbol{\kappa}_1|$ tends to zero. Nonetheless we are going to prove that the RS localization length λ^2 coincides with $\mathfrak{g}_{\alpha\alpha}$ in the large-N limit. The metric is by definition the infinitesimal distance:

$$\mathcal{D}^2_{0,\boldsymbol{\kappa}} = -\ln\ |\langle\Psi_0|\Psi_{\boldsymbol{\kappa}}\rangle|^2 \simeq N\mathfrak{g}_{\alpha\beta}(N)\kappa_\alpha\kappa_\beta; \tag{7.62}$$

by setting $\boldsymbol{\kappa} = \boldsymbol{\kappa}_1 = \frac{2\pi}{L}\mathrm{e}_1$ and $|\Psi_{0\boldsymbol{\kappa}_1}\rangle = \mathrm{e}^{-i\boldsymbol{\kappa}_1\cdot\hat{\mathbf{r}}}|\Psi_0\rangle$ we get

$$\mathcal{D}^2_{0\boldsymbol{\kappa}_1} \simeq \frac{4\pi^2 N}{L^2}\mathfrak{g}_{\alpha\alpha}(N) \simeq -\ln|\mathfrak{z}_N|^2, \tag{7.63}$$

113

$$\mathfrak{g}_{\alpha\alpha}(N) \simeq -\frac{L^2}{4\pi^2 N} \ln |\mathfrak{z}_N|^2 \tag{7.64}$$

where, at a given N, the difference between the two expressions is of order $1/L$ or higher; we finally get:

$$\mathfrak{g}_{\alpha\alpha} = \lim_{N\to\infty} \mathfrak{g}_{\alpha\alpha}(N) = -\frac{1}{4\pi^2} \lim_{N\to\infty} \frac{L^2}{N} \ln |\mathfrak{z}_N|^2 = \lambda^2. \tag{7.65}$$

We have already proved that the RS localization length diverges whenever the Drude weight is nonzero. After Eq. (7.65), an equivalent statement is that the convergence/divergence of the many-body metric, Eq. (7.60), discriminates between the insulating/metallic behavior. The relationship between λ^2 and the metric per electron also proves that λ^2 is an intensive quantity.

For the special case of a band insulator (with n_{b} doubly occupied bands) the many-body metric $\mathfrak{g}_{\alpha\beta}$ is related to the **k**-space metric as

$$\mathfrak{g}_{\alpha\beta} = \frac{V_{\mathrm{cell}}}{n_{\mathrm{b}}} \int_{\mathrm{BZ}} \frac{d\mathbf{k}}{(2\pi)^d} \, g_{\alpha\beta}(\mathbf{k}); \tag{7.66}$$

the proof is in Eqs. (7.55) and (7.65), albeit for the diagonal elements only.

It is also expedient to cast the metric in the form of a sum over states:

$$\mathfrak{g}_{\alpha\beta}(N) = \frac{1}{N} \sum_{n\neq 0} \frac{\langle \Psi_0 | \hat{v}_\alpha | \Psi_n \rangle \langle \Psi_n | \hat{v}_\beta | \Psi_0 \rangle}{\omega_{0n}^2}. \tag{7.67}$$

This is proved by inserting a complete set of states into $\langle \partial_{\kappa_\alpha} \Psi_0 | \partial_{\kappa_\beta} \Psi_0 \rangle$, to get

$$\langle \partial_{\kappa_\alpha} \Psi_0 | \partial_{\kappa_\beta} \Psi_0 \rangle = \sum_{n\neq 0} (\langle \partial_{\kappa_\alpha} \Psi_0 | \Psi_n \rangle \langle \Psi_n | \partial_{\kappa_\beta} \Psi_0 \rangle + \langle \partial_{\kappa_\alpha} \Psi_0 | \Psi_0 \rangle \langle \Psi_0 | \partial_{\kappa_\beta} \Psi_0 \rangle), \tag{7.68}$$

and then exploiting the $\boldsymbol{\kappa} \cdot \hat{\mathbf{p}}$ expansion [162, 163]

$$|\Psi_0(\boldsymbol{\kappa})\rangle \simeq |\Psi_0\rangle - \boldsymbol{\kappa} \cdot \sum_{n\neq 0} |\Psi_n\rangle \frac{\langle \Psi_n | \, \hat{\mathbf{v}} \, | \Psi_0 \rangle}{\omega_{0n}}, \tag{7.69}$$

$$|\partial_{\boldsymbol{\kappa}} \Psi_0(\boldsymbol{\kappa})\rangle = - \sum_{n\neq 0} |\Psi_n\rangle \frac{\langle \Psi_n | \, \hat{\mathbf{v}} \, | \Psi_0 \rangle}{\omega_{0n}}. \tag{7.70}$$

7.3.2 Drude weight revisited

Kohn's Drude weight can be recast in a geometrical form, having a rather perspicuous meaning; we arrive at an equivalent geometrical form starting from the

identity $\langle \Psi_{0\boldsymbol{\kappa}} | (\hat{H}_{\boldsymbol{\kappa}} - E_{0\boldsymbol{\kappa}}) | \Psi_{0\boldsymbol{\kappa}} \rangle \equiv 0$, taking two derivatives, and setting $\boldsymbol{\kappa} = 0$:

$$\frac{\partial^2 E_{0\boldsymbol{\kappa}}}{\partial \kappa_\alpha \partial \kappa_\beta} = \frac{N\hbar^2}{m} \delta_{\alpha\beta} - 2 \operatorname{Re} \langle \partial_{\kappa_\alpha} \Psi_{0\boldsymbol{\kappa}} | (\hat{H}_{\boldsymbol{\kappa}} - E_{0\boldsymbol{\kappa}}) | \partial_{\kappa_\beta} \Psi_{0\boldsymbol{\kappa}} \rangle \qquad (7.71)$$

$$D_{\alpha\beta} = \frac{\pi e^2 N}{m L^d} \delta_{\alpha\beta} - \frac{2\pi e^2}{\hbar^2 L^d} \operatorname{Re} \langle \partial_{\kappa_\alpha} \Psi_0 | (\hat{H} - E_0) | \partial_{\kappa_\beta} \Psi_0 \rangle, \qquad (7.72)$$

Therein, the first term measures the free-electron acceleration; the second (geometrical) one measures how much such acceleration is hindered by the one-body and two-body potentials. The geometrical term is zero even for the *interacting* electron gas; whenever instead the one-body potential is not flat, then both one-body and two-body terms in $\hat{V}$ concur in hindering the free acceleration. Eq. (7.72) is clearly the many-body analogue of Eq. (7.30).

7.3.3   The sum rule of Souza, Wilkens, and Martin

As anticipated above, SWM have proposed the integral

$$I_{\text{SWM}} = \int_0^\infty \frac{d\omega}{\omega} \sum_{\alpha=1}^d \operatorname{Re} \sigma_{\alpha\alpha}(\omega) \qquad (7.73)$$

as a discriminant between metals and insulators: in fact it diverges for all metals and converges for all insulators.

Let us discuss the insulating case first: there is no Drude peak, and $\sigma_{\alpha\beta}^{(+)}(\omega) = \sigma_{\alpha\beta}^{(\text{regular})}(\omega)$. Replacement of Eq. (7.25) into the integral yields

$$I_{\text{SWM}} = \frac{\pi e^2}{\hbar L^d} \sum_{n \neq 0} \sum_{\alpha=1}^d \frac{\langle \Psi_0 | \hat{v}_\alpha | \Psi_n \rangle \langle \Psi_n | \hat{v}_\alpha | \Psi_0 \rangle}{\omega_{0n}^2}; \qquad (7.74)$$

comparison to Eq. (7.67) yields the outstanding expression for the many-body metric—and equivalently for the RS localization tensor—in terms of conductivity, found by SWM in 2000:

$$\mathfrak{g}_{\alpha\beta} = \frac{\hbar L^d}{\pi e^2 N} \int_0^\infty \frac{d\omega}{\omega} \operatorname{Re} \sigma_{\alpha\beta}^{(+)}(\omega), \qquad (7.75)$$

where the large-N limit is understood. This relationship proves that the metric tensor is indeed an observable. At the independent-particle level, the SWM sum rule yields the gauge-invariant quadratic spread Ω_{I}, which is—as anticipated above—an observable as well:

$$\Omega_{\text{I}} = \frac{\hbar V_{\text{cell}}}{\pi e^2} \int_0^\infty \frac{d\omega}{\omega} \sum_{\alpha=1}^d \operatorname{Re} \sigma_{\alpha\alpha}(\omega) = \frac{\hbar V_{\text{cell}}}{\pi e^2} I_{\text{SWM}}. \qquad (7.76)$$

115

For an insulator with a spectral gap the f-sum rule, Eq. (7.8), provides an upper bound:

$$\mathfrak{g}_{\alpha\alpha} = \frac{\hbar L^d}{\pi e^2 N} \int_{\epsilon_{\text{gap}}/\hbar}^{\infty} \frac{d\omega}{\omega} \text{Re } \sigma_{\alpha\alpha}(\omega) < \frac{\hbar^2 L^d}{\pi e^2 N \epsilon_{\text{gap}}} \int_{\epsilon_{\text{gap}}/\hbar}^{\infty} d\omega \text{ Re } \sigma_{\alpha\alpha}(\omega) = \frac{\hbar^2}{m\epsilon_{\text{gap}}}.$$
(7.77)

In the metallic case I_{SWM} obviously diverges, owing to the Drude $\delta(\omega)$ term. In the special case of a band metal a SWM-like sum rule holds, where only the regular part of the conductivity is used:

$$I_{\alpha\beta}^{(\text{regular})} = \int_0^{\infty} \frac{d\omega}{\omega} \text{Re } \sigma_{\alpha\beta}^{(\text{regular})}(\omega) = \frac{\pi e^2}{\hbar} \int_{\text{BZ}} \frac{d\mathbf{k}}{(2\pi)^d} g_{\alpha\beta}(\mathbf{k}),$$
(7.78)

where in the metallic case we define the Bloch metric tensor of the occupied manifold as

$$
\begin{aligned}
g_{\alpha\beta}(\mathbf{k}) &= \text{Re } \sum_j \theta(\mu - \epsilon_{j\mathbf{k}}) \langle \partial_\alpha u_{j\mathbf{k}} | \partial_\beta u_{j\mathbf{k}} \rangle \\
&- \sum_{jj'} \theta(\mu - \epsilon_{j\mathbf{k}}) \theta(\mu - \epsilon_{j'\mathbf{k}}) \langle \partial_\alpha u_{j'\mathbf{k}} | u_{j\mathbf{k}} \rangle \langle u_{j\mathbf{k}} | \partial_\beta u_{j'\mathbf{k}} \rangle.
\end{aligned}
$$
(7.79)

In conclusion within PBCs the SWM sum rule is not an useful independent criterion to discriminate whether the ground state is insulating or metallic: in fact it is completely equivalent to assess whether the Drude weight is vanishing or not vanishing. Instead, when PBCs are adopted, the RS value of λ^2 is the most useful discriminant for the insulating state: see Sects. 7.5.5 and 7.5.6.

Matters are different within OBCs: as shown in Sect. 7.4.5 the large-system limit of the ground state metric—owing to the OBC SWM sum rule—discriminates very effectively between insulators and metals.

7.4 Bounded samples within open boundary conditions

We adopt here the same Kohn's Hamiltonian as above: Eqs. (5.11) and (7.10) in the time-reversal symmetric case, and Eq. (7.56) to deal even with case where the time-reversal symmetry is spontaneously broken. In this Section we switch from the PBCs adopted so far to OBCs: the cubic box confines the electrons in an infinite potential well; we indicate as $|\tilde{\Psi}_n(\boldsymbol{\kappa})\rangle$ the eigenstates. They are square-integrable over $\mathbb{R}^{Nd}$, and the position operator $\hat{\mathbf{r}} = \sum_i \mathbf{r}_i$ is the ordinary multiplicative operator. Within OBCs the effect of the gauge is easily "gauged away": the ground-state energy is $\boldsymbol{\kappa}$-independent, while the ground state is $|\tilde{\Psi}_{0\boldsymbol{\kappa}}\rangle = e^{-i\boldsymbol{\kappa}\cdot\hat{\mathbf{r}}}|\tilde{\Psi}_0\rangle$.

7.4.1 Many-body geometry within OBCs

The OBCs Berry connection has been already discussed above in Sect. 5.4.1; the OBCs curvature trivially vanishes. Only the metric deserves a thorough discussion.

The $\boldsymbol{\kappa}$-derivative of $|\tilde{\Psi}_0(\boldsymbol{\kappa})\rangle$ is

$$|\partial_{\boldsymbol{\kappa}}\tilde{\Psi}_0\rangle = -i\,\hat{\mathbf{r}}|\tilde{\Psi}_0\rangle; \tag{7.80}$$

hence the many-body metric, defined as in Eq. (7.60), is

$$\mathfrak{g}_{\alpha\beta}(N) = \frac{1}{N}(\,\langle\tilde{\Psi}_0|\hat{r}_\alpha\hat{r}_\beta|\tilde{\Psi}_0\rangle - \langle\tilde{\Psi}_0|\hat{r}_\alpha|\tilde{\Psi}_0\rangle\langle\tilde{\Psi}_0|\hat{r}_\beta|\tilde{\Psi}_0\rangle\,). \tag{7.81}$$

This is clearly a real symmetric tensor, whereas its PBC counterpart, Eq. (7.60), may be endowed with an imaginary antisymmetric part if time-reversal symmetry is absent; this is the Berry curvature, discussed in Ch. 8. It is not difficult to prove that the OBC metric can be cast as a sum over states, formally identical to its PBCs counterpart, Eq. (7.67):

$$\mathfrak{g}_{\alpha\beta}(N) = \frac{1}{N}\sum_{n\neq 0}\frac{\langle\tilde{\Psi}_0|\hat{v}_\alpha|\tilde{\Psi}_n\rangle\langle\tilde{\Psi}_n|\hat{v}_\beta|\tilde{\Psi}_0\rangle}{\omega_{0n}^2}. \tag{7.82}$$

The OBCs many-body metric per electron, Eq. (7.81), is clearly a second cumulant moment of the dipole; alternatively, it measures the quadratic quantum fluctuations of the polarization in the system ground state. An equivalent expression for $\mathfrak{g}_{\alpha\beta}(N)$ is in terms of the one-body density $n(\mathbf{r})$ and the two-body density $n^{(2)}(\mathbf{r},\mathbf{r}')$ is [149]:

$$\begin{aligned}
\mathfrak{g}_{\alpha\beta}(N) &= \frac{1}{2N}\int d\mathbf{r}\,d\mathbf{r}'\,(\mathbf{r}-\mathbf{r}')_\alpha(\mathbf{r}-\mathbf{r}')_\beta[\,n(\mathbf{r})n(\mathbf{r}') - n^{(2)}(\mathbf{r},\mathbf{r}')\,]. \\
&= -\frac{1}{2N}\int d\mathbf{r}\,d\mathbf{r}'\,(\mathbf{r}-\mathbf{r}')_\alpha(\mathbf{r}-\mathbf{r}')_\beta\,n(\mathbf{r})\,n_{\mathrm{xc}}(\mathbf{r},\mathbf{r}'), \tag{7.83}
\end{aligned}$$

where $n_{\mathrm{xc}}(\mathbf{r},\mathbf{r}')$ is by definition the exchange-correlation hole density. Therefore $\mathfrak{g}_{\alpha\beta}(N)$ is a second moment of the exchange-correlation hole, averaged over the sample. The fact that $\mathfrak{g}_{\alpha\beta}(N)$ includes only relative coordinates (and not absolute ones) is a desirable feature when the limit of a large system is addressed.

For any finite N the integration in Eq. (7.83) obviously converges, owing to the boundedness of the ground eigenstate $|\tilde{\Psi}_0\rangle$. The large-N limit of $\mathfrak{g}_{\alpha\beta}(N)$ discriminates, as in the PBCs case, between insulators and metals; there is, however, an important subtlety. Adoption of PBCs implies that the electric field $\boldsymbol{\mathcal{E}}$ averages to zero over the sample. Within OBCs, instead, depolarization fields affect the ground-state fluctuations, Eqs. (7.81) and (7.83). Such fields depend on the shape

117

of the sample, and the $N \to \infty$ limit requires extra care: the issue is discussed in Refs. [149, 164]. Therein it is shown that the problem is relevant only for genuinely interacting electrons, and does not exist at the independent-electron level (either Hartree-Fock or Kohn-Sham), where the large-N limit of the OBCs metric is shape-independent.

For independent electrons the ground state is a Slater determinant of square-integrable spinorbitals. All ground-state observables (and the wavefunction itself) are a function of the one-body density matrix; this in turn—for a singlet ground state—is twice the ground-state projector

$$\mathcal{P} = \sum_{j=1}^{N/2} |\varphi_j\rangle\langle\varphi_j| = \sum_{\epsilon_j < \mu} |\varphi_j\rangle\langle\varphi_j|, \tag{7.84}$$

where $|\varphi_j\rangle$ are the spinless orbitals, and ϵ_j their energies.

There are several equivalent expressions for $\mathfrak{g}_{\alpha\beta}(N)$ in terms of $\mathcal{P}$. One of them obtains straightforwardly from Eq. (7.81):

$$\mathfrak{g}_{\alpha\beta}(N) = \frac{2}{N}(\,\langle r_\alpha r_\beta\rangle - \langle r_\alpha\rangle\langle r_\beta\rangle\,) = \frac{2}{N}(\,\text{Tr}\,\{\mathcal{P}r_\alpha r_\beta\} - \text{Tr}\,\{\mathcal{P}r_\alpha \mathcal{P}r_\beta\}\,); \tag{7.85}$$

since it is a second cumulant moment, in the following we are using the same symbol λ^2 of its PBC counterpart, i.e. the RS squared localization length.

A useful equivalent form is

$$\mathfrak{g}_{\alpha\beta}(N) = -\frac{2}{N}\text{Tr}\,\{\mathcal{P}\,[\mathcal{P}, r_\alpha]\,[\mathcal{P}, r_\beta]\} = -\frac{2}{N}\int d\mathbf{r}\,\langle\mathbf{r}|\,\mathcal{P}\,[\mathcal{P}, r_\alpha]\,[\mathcal{P}, r_\beta]\,|\mathbf{r}\rangle. \tag{7.86}$$

Starting instead from Eq. (7.83) we get the alternative expression

$$\lambda^2 = \mathfrak{g}_{\alpha\beta}(N) = \frac{1}{N}\int d\mathbf{r}\,d\mathbf{r}'\,(\mathbf{r} - \mathbf{r}')_\alpha(\mathbf{r} - \mathbf{r}')_\beta\,|\langle\mathbf{r}|\mathcal{P}|\mathbf{r}'\rangle|^2 \quad \text{(double occup.)}; \tag{7.87}$$

for future use, we write the analogous expression for single occupancy

$$\lambda^2 = \mathfrak{g}_{\alpha\beta}(N) = \frac{1}{2N}\int d\mathbf{r}\,d\mathbf{r}'\,(\mathbf{r} - \mathbf{r}')_\alpha(\mathbf{r} - \mathbf{r}')_\beta\,|\langle\mathbf{r}|\mathcal{P}|\mathbf{r}'\rangle|^2 \quad \text{(single occup.)}. \tag{7.88}$$

At finite size both spectra (OBCs and PBCs) are discrete, and different between them. In the large-system limit both become continuous, and they coincide; the density of states also coincides. We argue that the two metrics coincide as well; we are going to show that they provide indeed an identical value in the special case of a crystal of independent electrons, either band metal or band insulator.

7.4.2 OBC vs. PBC metrics (independent electrons)

Let us start with the OBCs metric first. In the large-crystallite limit Eq. (7.87) becomes

$$\lim_{N \to \infty} \mathfrak{g}_{\alpha\beta}(N) = \frac{1}{n_{\mathrm{b}}} \int_{\mathrm{cell}} d\mathbf{r} \int_{\mathrm{all\,space}} d\mathbf{r}' \, (\mathbf{r} - \mathbf{r}')_\alpha (\mathbf{r} - \mathbf{r}')_\beta \, |\langle \mathbf{r} | \mathcal{P} | \mathbf{r}' \rangle|^2, \qquad (7.89)$$

where $2n_{\mathrm{b}}$ is the number of electrons per crystallite cell. It is clear that the convergence/divergence of $\mathfrak{g}_{\alpha\beta}(N)$ depends on the convergence/divergence of the inner integral, which in turn depends on the asymptotic decay of $|\langle \mathbf{r} | \mathcal{P} | \mathbf{r}' \rangle|$ for $|\mathbf{r} - \mathbf{r}'| \to \infty$. It is well known that such decay is quasi-exponential (i.e. exponential times a power) for all band insulators [165]: it is therefore obvious that Eq. (7.89) converges to a finite value. In the metallic case, instead, the density-matrix decay is power-law; the divergence of Eq. (7.89) can be verified in $d = 1, 2, 3$ for the simplest metal of all, i.e. the free-electron gas, for which the density matrix is analytically known [166]. As for Eq. (7.87)—where the integration is performed over a bounded sample—simulations and heuristic arguments altogether [167, 168, 150] suggest that the metallic divergence of $\mathfrak{g}_{\alpha\beta}(N)$ is of order of the linear dimension L of the system in $d = 1, 2$ or 3.

Let us switch next to an unbounded sample within PBCs. In the metallic case, as expected, $\mathfrak{g}_{\alpha\beta} = \infty$: see Eq. (7.4) and the following text. For insulators instead we have the finite value

$$\mathfrak{g}_{\alpha\beta} = \frac{V_{\mathrm{cell}}}{n_{\mathrm{b}}} \int_{\mathrm{BZ}} \frac{d\mathbf{k}}{(2\pi)^d} \, g_{\alpha\beta}(\mathbf{k}) = \frac{V_{\mathrm{cell}}}{n_{\mathrm{b}}} \int_{\mathrm{BZ}} \frac{d\mathbf{k}}{(2\pi)^d} \, \mathrm{Re}\,\mathrm{Tr}\,\{\mathcal{P}_{\mathbf{k}}(\partial_\alpha \mathcal{P}_{\mathbf{k}})(\partial_\beta \mathcal{P}_{\mathbf{k}})\}, \qquad (7.90)$$

where Eqs. (7.66) and (3.47) have been used. We are going to prove that the value of $\mathfrak{g}_{\alpha\beta}$ as provided by Eq. (7.90) coincides with the large-N limit of Eq. (7.86).

The single-particle density matrix (a.k.a. ground-state projector) for the unbounded crystalline insulator is, in the Schrödinger representation,

$$\langle \mathbf{r} | \mathcal{P} | \mathbf{r}' \rangle = V_{\mathrm{cell}} \int_{\mathrm{BZ}} \frac{d\mathbf{k}}{(2\pi)^d} \, \mathrm{e}^{i\mathbf{k} \cdot (\mathbf{r} - \mathbf{r}')} \langle \mathbf{r} | \mathcal{P}_{\mathbf{k}} | \mathbf{r}' \rangle; \qquad (7.91)$$

we start noticing that the integrand is periodical over the reciprocal lattice, and therefore the BZ integral of its $\mathbf{k}$-gradient vanishes:

$$i(\mathbf{r} - \mathbf{r}')\langle \mathbf{r} | \mathcal{P} | \mathbf{r}' \rangle + V_{\mathrm{cell}} \int_{\mathrm{BZ}} \frac{d\mathbf{k}}{(2\pi)^d} \, \mathrm{e}^{i\mathbf{k} \cdot (\mathbf{r} - \mathbf{r}')} \langle \mathbf{r} | \, \partial_{\mathbf{k}} \mathcal{P}_{\mathbf{k}} \, | \mathbf{r}' \rangle = 0, \qquad (7.92)$$

we also remind that Eq. (7.92) is a well behaved expression only in insulators. The first term therein is i times $[\mathbf{r}, \mathcal{P}]$: a lattice periodical operator (unlike $\mathbf{r}$ itself):

$$[\mathbf{r}, \mathcal{P}] = \frac{iV_{\mathrm{cell}}}{(2\pi)^3} \int_{\mathrm{BZ}} \frac{d\mathbf{k}}{(2\pi)^d} \, \mathrm{e}^{i\mathbf{k} \cdot (\mathbf{r} - \mathbf{r}')} \partial_{\mathbf{k}} \mathcal{P}_{\mathbf{k}}. \qquad (7.93)$$

The trace of Eq. (7.92) can therefore be cast as

$$\int_{\text{BZ}} \frac{d\mathbf{k}}{(2\pi)^d} \operatorname{Tr} \{\partial_\alpha \mathcal{P}_\mathbf{k}\} = -\frac{i}{V_{\text{cell}}} \int_{\text{cell}} d\mathbf{r} \langle \mathbf{r}| [r_\alpha, \mathcal{P}] |\mathbf{r}\rangle; \tag{7.94}$$

using similar arguments it is not difficult to prove that, for an unbounded sample within PBCs,

$$\int_{\text{BZ}} \frac{d\mathbf{k}}{(2\pi)^d} \operatorname{Tr} \{\mathcal{P}_\mathbf{k}(\partial_\alpha \mathcal{P}_\mathbf{k})(\partial_\beta \mathcal{P}_\mathbf{k})\} = \frac{1}{V_{\text{cell}}} \int_{\text{cell}} d\mathbf{r} \langle \mathbf{r}| \mathcal{P} [r_\alpha, \mathcal{P}] [r_\beta, \mathcal{P}] |\mathbf{r}\rangle. \tag{7.95}$$

Notice that the three BZ integrals entering the product on the l.h.s.—from Eqs. (7.91) and (7.93)—eventually contract to a single BZ integral: this owes to the fact that both $\langle \mathbf{r}| \mathcal{P}_\mathbf{k} |\mathbf{r}'\rangle$ and $\langle \mathbf{r}| \partial_\mathbf{k} \mathcal{P}_\mathbf{k} |\mathbf{r}'\rangle$ are lattice-periodical in $\mathbf{r}$ and $\mathbf{r}'$ *separately*. Details about such transformation can be found in the Appendix of Ref. [169].

Replacing Eq. (7.95) into Eq. (7.90) we get

$$\mathfrak{g}_{\alpha\beta} = \frac{1}{n_{\text{b}}} \int_{\text{cell}} d\mathbf{r} \operatorname{Re} \langle \mathbf{r}| \mathcal{P} [r_\alpha, \mathcal{P}] [r_\beta, \mathcal{P}] |\mathbf{r}\rangle. \tag{7.96}$$

We are now ready to compare this with the analogous expression of Eq. (7.86), where the integrand is the same but the integral is over the whole crystallite. Neglecting boundary terms in the large-crystallite limit, one clearly has

$$\frac{1}{n_{\text{b}}} \int_{\text{cell}} d\mathbf{r} \operatorname{Re} \langle \mathbf{r}| \mathcal{P} [r_\alpha, \mathcal{P}] [r_\beta, \mathcal{P}] |\mathbf{r}\rangle = \frac{2}{N} \int_{\text{crystallite}} d\mathbf{r} \langle \mathbf{r}| \mathcal{P} [r_\alpha, \mathcal{P}] [r_\beta, \mathcal{P}] |\mathbf{r}\rangle, \tag{7.97}$$

if the cell is chosen in the bulk region of the crystallite. Notice that we have omitted to indicate the real part in the r.h.s. of Eq. (7.97): the OBCs expression is always real symmetric. Instead the PBCs metric-curvature tensor may be endowed with an imaginary antisymmetric part; this is discussed in Ch. 8. Eq. (7.97) concludes our proof: the PBCs metric of an insulator, expressed as a BZ integral in Eq. (7.90), coincides in the large-N limit with the OBCs metric of a bounded crystallite, expressed as a double integral over the electron coordinates, as in Eq. (7.86).

7.4.3 Conductivity of a bounded sample within OBCs

A bounded sample cannot sustain a steady-state current, and therefore the concept of Drude weight is *apparently* meaningless (see below, Sect. 7.4.4). An oscillating field $\mathcal{E}$ induces charge sloshing and an oscillating macroscopic polarization $\mathbf{P}$, which has a finite dc limit in insulators, and diverges in metals: in fact $\partial \mathbf{P}/\partial \mathcal{E}$ is the static dielectric susceptibility, formally infinite in metals.

The OBC electronic term in polarization (Sect. 5.4.1) is

$$\mathbf{P}^{(\mathrm{el})} = -\frac{e}{\mathcal{V}} \langle \tilde{\Psi}_0 | \, \hat{\mathbf{r}} \, | \tilde{\Psi}_0 \rangle; \tag{7.98}$$

since $\hat{\mathbf{r}}$ is a legitimate operator within OBCs, one could adopt the scalar potential gauge; we prefer here instead to continue with the vector potential gauge, thus exploiting some of the previous formulæ.

The Kubo formula for the induced polarization, analogue of Eq. (7.19), is

$$P_\alpha(\omega) = -\frac{e^2}{cL^d} \langle\langle \hat{r}_\alpha | \hat{v}_\beta \rangle\rangle_\omega \, \delta A(\omega), \tag{7.99}$$

and proceeding in the same way as above we get the analogue of Eq. (7.21):

$$P_\alpha(\omega) = -\frac{e^2}{L^d} \langle\langle \hat{r}_\alpha | \hat{v}_\beta \rangle\rangle_\omega \left[\frac{1}{i\omega} - \pi\delta(\omega) \right] \mathcal{E}_\beta(\omega). \tag{7.100}$$

At this point we define the OBC conductivity $\tilde{\sigma}_{\alpha\beta}(\omega)$ by means of the relationship $\mathbf{j}(t) = \partial \mathbf{P}(t)/\partial t$, i.e.

$$\tilde{\sigma}_{\alpha\beta}(\omega)\mathcal{E}_\beta(\omega) == -i\omega P_\alpha(\omega) = \frac{e^2}{L^d} \langle\langle \hat{r}_\alpha | \hat{v}_\beta \rangle\rangle_\omega \mathcal{E}_\beta(\omega), \tag{7.101}$$

where the multiplication by $-i\omega$ cancels the $\delta(\omega)$ term in Eq. (7.100): as expected, there is no Drude peak (at finite size). The spectral decomposition of the Kubo formula, Eq. (7.15), yields

$$\langle\langle \hat{r}_\alpha | \hat{v}_\beta \rangle\rangle_\omega = \frac{1}{\hbar} \lim_{\eta \to 0^+} \sum_{n \neq 0}{}' \left(\frac{\langle \tilde{\Psi}_0 | \hat{r}_\alpha | \tilde{\Psi}_n \rangle \langle \tilde{\Psi}_n | \hat{v}_\beta | \tilde{\Psi}_0 \rangle}{\omega - \omega_{0n} + i\eta} - \frac{\langle \tilde{\Psi}_0 | \hat{v}_\beta | \tilde{\Psi}_n \rangle \langle \tilde{\Psi}_n | \hat{r}_\alpha | \tilde{\Psi}_0 \rangle}{\omega + \omega_{0n} + i\eta} \right). \tag{7.102}$$

We then exploit the usual commutator,

$$\langle \Psi_0 | \hat{\mathbf{r}} | \Psi_n \rangle = i \frac{\langle \Psi_0 | \hat{\mathbf{v}} | \Psi_n \rangle}{\omega_{0n}}, \tag{7.103}$$

to obtain

$$
\begin{aligned}
\langle\langle \hat{r}_\alpha | \hat{v}_\beta \rangle\rangle_\omega &= \frac{i}{\hbar} \lim_{\eta \to 0^+} \sum_{n \neq 0}{}' \left(\frac{\langle \tilde{\Psi}_0 | \hat{v}_\alpha | \tilde{\Psi}_n \rangle \langle \tilde{\Psi}_n | \hat{v}_\beta | \tilde{\Psi}_0 \rangle}{\omega_{0n}(\omega - \omega_{0n} + i\eta)} + \frac{\langle \tilde{\Psi}_0 | \hat{v}_\beta | \tilde{\Psi}_n \rangle \langle \tilde{\Psi}_n | \hat{v}_\alpha | \tilde{\Psi}_0 \rangle}{\omega_{0n}(\omega + \omega_{0n} + i\eta)} \right) \\
&= \frac{i}{\hbar} \lim_{\eta \to 0^+} \sum_{n \neq 0}{}' \left(\frac{\mathcal{R}_{n,\alpha\beta} + i\mathcal{I}_{n,\alpha\beta}}{\omega_{0n}(\omega - \omega_{0n} + i\eta)} + \frac{\mathcal{R}_{n,\alpha\beta} - i\mathcal{I}_{n,\alpha\beta}}{\omega_{0n}(\omega + \omega_{0n} + i\eta)} \right), \quad (7.104)
\end{aligned}
$$

where we are using the same notations as in Eq. (7.23). Notice, however, that the OBCs matrix elements—and their relative selection rules—are *very different* from the PBCs ones.

We finally get

$$\operatorname{Re} \tilde{\sigma}_{\alpha\beta}(\omega) = \frac{\pi e^2}{\hbar L^d} {\sum_{n\neq 0}}' \frac{\mathcal{R}_{n,\alpha\beta}}{\omega_{0n}} [\, \delta(\omega - \omega_{0n}) + \delta(\omega + \omega_{0n})\,], \qquad (7.105)$$

$$\operatorname{Im} \tilde{\sigma}_{\alpha\beta}(\omega) = \frac{2e^2}{\hbar L^d} {\sum_{n\neq 0}}' \frac{\mathcal{R}_{n,\alpha\beta}}{\omega_{0n}} \frac{\omega}{\omega_{0n}^2 - \omega^2}.$$

Despite having a similar form, Eqs. (7.25) and (7.105) do not coincide.: eigenvalues, eigenvectors, and selection rules are different. Most notably in the metallic case $\tilde{\sigma}_{\alpha\beta}(\omega)$ saturates the f-sum rule, Eq. (7.8), while $\sigma_{\alpha\beta}^{(\text{regular})}(\omega)$ does not: a fraction of the spectral weight goes into the Drude peak.

The major difference between PBCs and OBCs conductivity in a metal can be understood addressing the simplest metal of all: the free electron gas. Therein, as said above, $\sigma_{\alpha\beta}^{(\text{regular})}$ vanishes because of the selection rules: the offdiagonal matrix elements of the velocity between plane waves are zero. Therefore in PBCs all of the spectral weight goes into the Drude peak. These selection rules break down in OBCs, and there is no genuine Drude peak.

7.4.4 Drude weight in bounded samples within OBCs

Within OBCs D apparently does not exist, given that a bounded sample does not support dc currents; the apparent paradox was addressed in Ref. [170], and more recently in Ref. [151]. As said above, D measures the free acceleration induced by a constant field, i.e. the adiabatic inverse inertia of the many-electron system. In the case of a bounded crystallite a constant field induces no current: it simply polarizes the sample (Faraday cage effect). But an oscillating low-frequency field induces forced oscillations, which are dominated by the many-electron inertia. The response carries therefore the same essential information as the response to a constant field within PBCs. Technically, D within OBCs originates from the low-frequency sector of the nonzero-frequency Kubo formula. The root of the PBC vs. OBC difference is in the different selection rules for the intraband transitions (in the simple case of noninteracting electrons).

For free electrons (noninteracting electrons in a flat potential) within PBCs the interband contributions vanish, and the whole response is intraband: there are no transitions at $\omega \neq 0$. When the same system is confined in an infinite potential well, all of the poles of Eq. (7.105) are at finite frequency, and $\operatorname{Re} \tilde{\sigma}_{\alpha\beta}(\omega)$ saturates the f-sum rule. Simulations performed in Ref. [151] show that the spectral weight

concentrates in a region whose width goes to zero with $1/L$, and that the sum of the residui yield a very accurate D value: ergo for this paradigmatic system all the poles of Eq. (7.105) coalesce into a single pole at $\omega = 0$. It is thus reassuring that in the $L \to \infty$ limit OBCs yield the same result as PBCs; in fact irrelevance of the boundary conditions in the thermodynamic limit is a basic tenet of statistical mechanics and condensed matter physics.

As discussed in detail above, the PBC longitudinal conductivity of a general metallic system, Eq. (7.5), is the sum of two terms: the Drude term $\sigma_{\alpha\beta}^{(\text{Drude})}(\omega)$ and the regular term $\sigma_{\alpha\beta}^{(\text{regular})}(\omega)$. When the same system is addressed within OBCs, at finite size all poles are at finite frequency; nonetheless these poles can be partitioned in two classes, which can be separated by means of their $L \to \infty$ behavior. Some of the poles coalesce into a zero-frequency pole, thus accounting for the D value; the other poles provide the same $\sigma_{\alpha\beta}^{(\text{regular})}(\omega)$ as for the PBC case. Actual simulations for a the test case of a band metal in $1d$ perspicuously confirm this view [151].

7.4.5 Souza-Wilkens-Martin within OBCs

We evaluate the SWM integral in terms of the OBC conductivity, Eq. (7.73), for a bounded sample:

$$I_{\text{SWM}} = \int_0^\infty \frac{d\omega}{\omega} \sum_{\alpha=1}^d \text{Re } \tilde{\sigma}_{\alpha\alpha}(\omega) = \frac{\pi e^2}{\hbar L^d} \sum_{n \neq 0} \sum_{\alpha=1}^d \frac{\langle \tilde{\Psi}_0 | \hat{v}_\alpha | \tilde{\Psi}_n \rangle \langle \tilde{\Psi}_n | \hat{v}_\alpha | \tilde{\Psi}_0 \rangle}{\omega_{0n}^2}. \quad (7.106)$$

Using then Eq. (7.82), I_{SWM} assumes the form of a gauge invariant intensive ground-state property:

$$I_{\text{SWM}} = \frac{\pi e^2 N}{\hbar L^d} \sum_{\alpha=1}^d \mathfrak{g}_{\alpha\alpha}(N). \quad (7.107)$$

At any finite sample size I_{SWM} is finite; the convergence/divergence of the OBC metric in the large-sample limit discriminates in a very effective way between insulators and metals, as shown below in Sects. 7.5.1, 7.5.3, and 7.5.4. The ultimate reason while this happens has been understood only recently [151].

The sum over the excitations in Eq. (7.106) is dominated by the low-frequency contributions. As explained above, it was indeed found that some of the OBC poles in the metallic case and in the large-system limit accumulate at the $\omega = 0$ frequency, thus accounting for the Drude weight (or equivalently for the inverse inertia of the many-electron metallic system), yielding a divergent I_{SWM}. There is not such an accumulation in the insulating case: I_{SWM} remains finite, proportional to the PBC value of the RS λ^2.

7.5 Band insulators, Anderson insulators, Mott insulators, and more

7.5.1 Band insulator: model ionic crystal in $1d$

Let us start with a simple tight-binding (a.k.a. Hückel) Hamiltonian in $1d$:

$$H = \sum_{j\sigma} [\, (-1)^j \Delta\, c^\dagger_{j\sigma} c_{j\sigma} - t(c^\dagger_{j\sigma} c_{j+1\sigma} + \text{H.c.}) \,] \qquad (7.108)$$

where $t > 0$ is the first neighbor hopping ($\beta = -t$ in most chemistry literature) and H.c. stays for Hermitian conjugate. This is the "parent" Hamiltonian for those used below to describe model $1d$ disordered and correlated systems, Eqs. (7.113) and (7.115), respectively.

This toy model schematizes a binary ionic crystal; the band structure is

$$\epsilon(k) = \pm\sqrt{\Delta^2 + 4t^2 \cos^2 ka/2}, \qquad (7.109)$$

where a is the lattice constant and k is the Bloch vector; the gap is equal to 2Δ. The corresponding density of states is shown in Fig. 7.3 for $t = 1$ and $\Delta = 0.25$; at the band edges it shows van Hove singularities, which in $1d$ have the character of $1/\sqrt{\epsilon}$ divergences. The system is insulating at half filling except for $\Delta = 0$; it is metallic at any filling different from $1/2$.

The OBC metric is the tight-binding version of Eq. (7.88), i.e.

$$\lambda^2 = \frac{a^2}{2N} \sum_{j,j'=1}^{N} P^2_{jj'}(j - j')^2. \qquad (7.110)$$

This is a monothonical function of t/Δ; it is easily verified that it vanishes in the extreme ionic case ($t = 0$). λ^2 has been evaluated in Ref. [167] from Eq. (7.110) as a function of N: at filling $1/4$ (metallic) it diverges linearly with N, i.e. like the linear size L of the system, in qualitative agreement with the case of the $1d$ electron gas [166]; instead at filling $1/2$ (insulating) it converges like $1/N$ to 0.99 (for the given t/Δ value and $a = 1$).

The system is also metallic at half filling for $\Delta = 0$. In this case the ground-state projector has a simple analytical form:

$$P_{jj} = \frac{1}{2}; \quad P_{jj'} = 0 \text{ for even } |j' - j| = 2s, \qquad (7.111)$$

$$P_{jj'} = \frac{(-1)^s}{\pi(2s + 1)} \text{ for odd } |j' - j| = 2s + 1, \qquad (7.112)$$

which clearly implies divergence of Eq. (7.110) linear with N, again similarly to λ^2 for the $1d$ electron gas. Notice that Eq. (7.111) is a manifestation of the Coulson-Rushbrooke theorem [171].
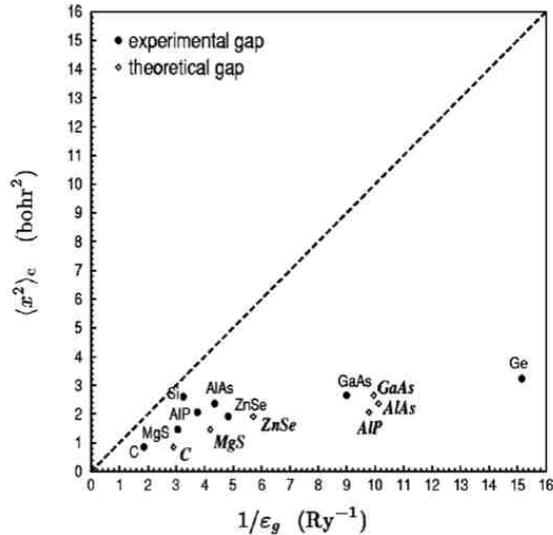
Figure 7.2: Diagonal element of the Kohn-Sham RS squared localization length λ^2—also indicated as $\langle x^2 \rangle_c$—vs. the inverse direct gap (theoretical and experimental), for several elemental and binary semiconductors. The points corresponding to Si and Ge with the theoretical gaps are out of scale. After Ref. [148].

7.5.2 Band insulators: tetrahedrally coordinated semiconductors

The first ab-initio implementation of the modern theory of the insulating state (in 2001) addressed several elemental and binary cubic semiconductors at the Kohn-Sham level [148]. The tensor is real and isotropic. The computed λ^2 (Fig. 7.2) is smaller than 3 bohr^2 in all the materials studied: the ground many-body wavefunction is therefore very localized in this class of materials. The SWM inequality, Eq. (7.77), was also checked, and found to be well verified using both the theoretical Kohn-Sham gap and the experimental one (the latter is typically larger).

7.5.3 Anderson insulator: model $1d$ system

According to the famous "gang of four" paper by Abrahams, Anderson, Licciardello, and Ramakrishnan [172] there is no metal-insulator transition in $1d$: any amount of disorder induces Anderson localization, yielding an insulating ground state. Matters are quite different in $3d$: see Sect. 7.5.4 below. It is well known both from analytical arguments and actual simulations that the spectrum of this paradigmatic model is gapless [144, 173]: as discovered by Anderson, and later emphasized by Kohn, the insulating state does not require an energy gap.

We rewrite the tight-binding Hamiltonian of Eq. (7.108) in the more general

125
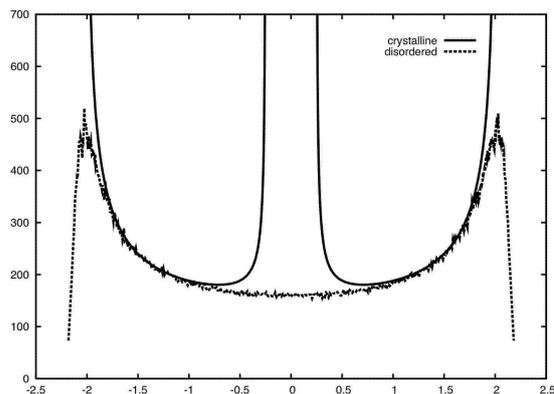
Figure 7.3: Density of states (arbitrary units) for a model binary alloy in $1d$. The crystalline (band) case corresponds to the Hamiltonian of Eq. (7.108) with $\Delta = 0.25$ and $t = 1$. The disordered (Anderson) case corresponds to a random choice of the anion/cation distribution.

form of

$$H = \sum_{j\sigma} [\, \epsilon_j \, c_{j\sigma}^\dagger c_{j\sigma} - t(c_{j\sigma}^\dagger c_{j+1\sigma} + \text{H.c.})\,]. \tag{7.113}$$

The site energies in Eq. (7.113) used in our simulations are therefore $\epsilon_j = (-1)^j \Delta$ in the crystalline case, while in the disordered case the string $(-1)^j$ is replaced by a random string of ± 1, chosen with equal (and uncorrelated) probability. The simulations of Ref. [167] are for samples of up to 2000 sites, whereas in the disordered case they are typically averaged over 1000 configurations. The density of states for both the ordered and disordered systems are shown in Fig. 7.3, and confirm the expected features. The disordered system is gapless, yet it is known to be insulating at any filling.

The conventional theory of transport focusses on the nature of the one-particle orbitals at the Fermi level; in Anderson insulators these are localized, thus forbidding steady state currents [142]. Sixty years of literature have been devoted to investigate Anderson insulators under the most diverse aspects [144, 173, 174, 175].

At variance with such wisdom, in Ref. [167] we have addressed this paradigmatic Anderson insulator from the alternative, and most fundamental, viewpoint of the modern theory of the insulating state. In the spirit of Kohn's theory the individual Hamiltonian eigenstates become apparently irrelevant, while the focus is on the many-electron ground state as a whole. We also emphasize that numerical simulations in this field are limited since the very beginning to lattice models, given that they adopt recursion methods and the like [173]. The modern theory of the insulating state is instead formulated in general terms: the lattice-model

implementation presented here is just a special case of it.

The OBC metric has been computed from Eq. (7.110), at both half- and quarter-filling: in both cases it is found to be finite in the large-N limit, as expected. Nonetheless its value is about 20 times larger than the one for the band insulator, at the same value of the parameters (i.e. $\Delta = 0.25, t = 1$). This reflects the fact that the scattering mechanisms are profoundly different: incoherent (Anderson) versus coherent (band). In the latter case, the Hamiltonian eigenstates are individually conducting but "locked" by the Pauli principle if the Fermi level lies in the gap.

7.5.4 Anderson metal-insulator transition in a model $3d$ solid

According to the scaling theory of localization (the famous "gang of four" paper [172]) a genuine metal-insulator transition—as a function of the disorder strength—is possible in $3d$. The paradigmatic model previously adopted in the literature to realize the transition is a simple tight-binding Hamiltonian on a cubic lattice, with random onsite matrix elements, at half filling:

$$ H = -t \sum_{<ij>} c_i^\dagger c_j + \text{H.c.} + W \sum_i \epsilon_i c_i^\dagger c_i, \qquad (7.114) $$

where i, j denote sites on a simple cubic lattice, $< ij >$ are pairs of nearest neighbor sites and the onsite energies ϵ_i are randomly picked from the interval $[-1, 1]$. W is the disorder strength and the model has previously been shown to exhibit an Anderson transition at $W_c/t = 8.25$ [176, 177, 178, 179].

The current computational methods to address the Anderson transition focus on the single-particle eigenstates at the Fermi level, and are often peculiar to lattice models (recursive methods and the like) [173, 180], while instead here we regard this case as an application of the general theory of the insulating state to the special case of a $3d$ tight-binding Hamiltonian.

The OBC metric has been calculated for this model in Ref. [181]; for the sake of clarity, we stress that the present localization length λ bears *no relationship* to the Anderson localization length [173]: the former is a geometrical property of the many-body ground state, while the latter is a property of the one-body eigenstates in an independent-electron system.

Simulating the Anderson transition requires notoriously rather large systems; in Ref. [181] we have calculated λ^2, within OBCs, for various values of W using rods of size $L \times d \times d$ where $L = 100$ and $d = 3, 5, 7$ (we set the cubic lattice constant $a = 1$). To obtain the configurational average we used 100 configurations and for each configuration the component of the tight-binding version of Eq. (7.88) along the rod was obtained by averaging over the two short dimensions.
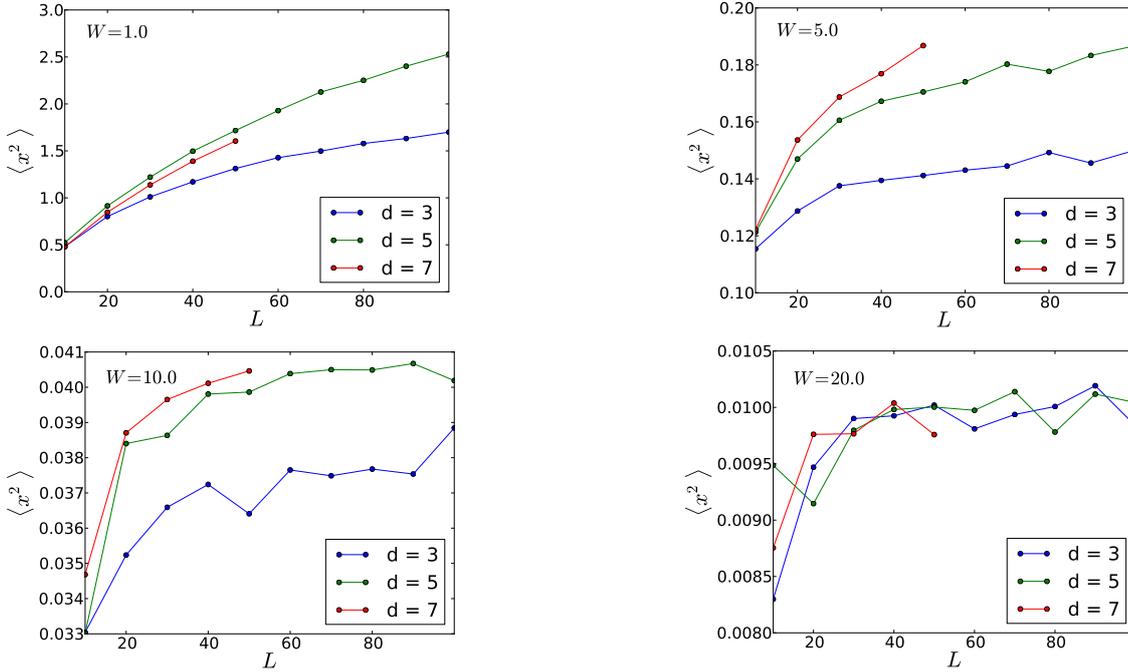
Figure 7.4: RS squared localization length λ^2 as a function of rod length L. λ^2 diverges for small values of W and saturates to a finite value for large values of W. After ref. [181].

In a disordered system the tight-binding matrix elements of $|\langle\mathbf{r}|\mathcal{P}|\mathbf{r}'\rangle|^2$ in Eq. (7.88) must be replaced by their configurational average $\langle\,|\langle\mathbf{r}|\mathcal{P}|\mathbf{r}'\rangle|^2\,\rangle_{\mathrm{conf}}$. A very crucial point is that $\langle\,|\langle\mathbf{r}|\mathcal{P}|\mathbf{r}'\rangle|^2\,\rangle_{\mathrm{conf}}$ is in general *different* from the squared modulus of the configurational average of $\langle\mathbf{r}|\mathcal{P}|\mathbf{r}'\rangle$. This is reminiscent of—and related to— the so-called vertex corrections in the well established transport theories based on Green's functions [153, 182]; the point is thoroughly discussed in our original paper, Ref. [181].

The results for various values of W are shown in Fig. 7.4 for different rod widths d. We clearly observe a tendency for λ^2 to saturate when W becomes large. For small W, instead, λ^2 appears to be increasing monotonically with the rod length L. Within the modern theory of the insulating state the Anderson transition would emerge as a transition from a divergent to a finite λ^2 in the limit of large L. While it seems plausible that this may happen around $W_{\mathrm{c}} = 8.25$, it is very difficult to extract a quantitative estimate of W_{c} from λ^2 alone. For example, for $W = 10$, the localization length appears to be saturated at a finite value for $L \sim 100$, but it is hard to verify if this is really the case or if λ^2 is merely increasing too slowly to be observable at the size of our simulations.

Given the above difficulty, a different and more complex strategy has been devised in Ref. [181]; we refer to the original paper for the details. The convergence/divergence of λ^2 has been assessed by means of an ad-hoc indicator which analyzes the asymptotic behavior of $\langle\,|\langle\mathbf{r}|\mathcal{P}|\mathbf{r}'\rangle|^2\,\rangle_{\mathrm{conf}}$ for large $|\mathbf{r}-\mathbf{r}'|$. Notice that λ^2 stays obviously finite if the behavior is exponential, but it is also finite if $\langle\,|\langle\mathbf{r}|\mathcal{P}|\mathbf{r}'\rangle|^2\,\rangle_{\mathrm{conf}} \sim |\mathbf{r}-\mathbf{r}'|^{-\beta}$ with $\beta > 5$. The transition is very sharp using our indicator, which switches in a narrow W interval; The approach yields a critical disorder parameter $W_{\mathrm{c}} \approx 8.5$, not far from the best value of about 8.25 from the previous literature.

7.5.5 Two-band model insulator in $1d$: topological nature of the Mott-like transition

Arguably the simplest possible $1d$ highly correlated insulator obtains from the noninteracting Hamiltonian of Eq. (7.108) and augmenting it with an on-site repulsive term. We thus get the two-band Hubbard model (at half filling):

$$H = \sum_{j\sigma}[(-1)^j\Delta\, c_{j\sigma}^\dagger c_{j\sigma} - t(c_{j\sigma}^\dagger c_{j+1\sigma} + \mathrm{H.c.})] + U\sum_j n_{j\uparrow}n_{j\downarrow}. \qquad (7.115)$$

We assume $\Delta > 0$, and neutralizing classical charges equal to $+1$ on all sites; the system is clearly inversion-symmetric at any U.
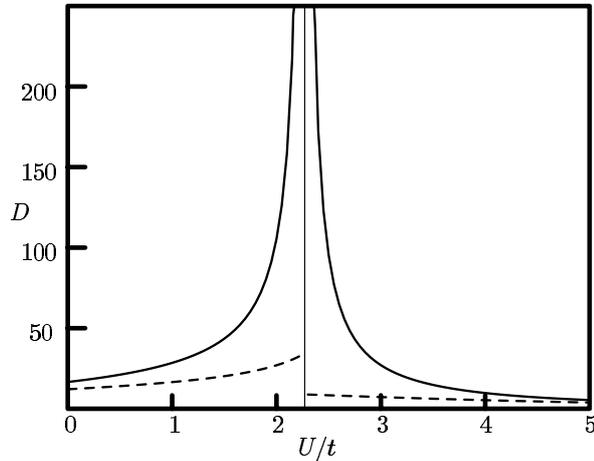

Figure 7.5: Squared localization length for the Hamiltonian in Eq. (7.115) at half filling for $t/\Delta = 1.75$: the plot shows the dimensionless quantity $D = (2\pi N/L)^2\lambda^2$. The system undergoes a quantum phase transition from band-like insulator ($\mathbb{Z}_2$-odd) to Mott-like insulator ($\mathbb{Z}_2$-even) at $U/t = 2.27$. After Ref. [146].

Preliminarly, it is expedient to investigate the trivial $t = 0$ case. At small U the anion site (odd j) is doubly occupied, and the energy per cell is $-2\Delta + U$; at $U > 2\Delta$ single occupancy of each site is instead energetically favored. As for polarization, it is easily realized that the system is $\mathbb{Z}_2$-odd in the former case and $\mathbb{Z}_2$-even in the latter. At the transition point $U_c = 2\Delta$ the ground state is degenerate and the spectrum is gapless, ergo the system is "metallic". If the hopping t is then switched on adiabatically, the $\mathbb{Z}_2$ invariant in each of the two topological phases cannot flip unless a metallic state is crossed.

The explicitly correlated ground-state wavefunction has been found by exact diagonalization in Ref. [146], and the corresponding λ^2 has been computed as a function of U for fixed $t/\Delta = 1.75$. The results are shown in Fig. 7.5 in dimensionless units; it turns out that there is only one singular point $U = 2.27t$, where λ^2 diverges. Indeed, it has been verified that at such value the ground-state becomes degenerate with the first excited singlet state, i.e. the system is metallic. The singular point is the fingerprint of a quantum phase transition: on the left we have a band-like insulator, and on the right a Mott-like insulator. There is no metal-insulator transition, only an insulator-insulator transition, while the system is metallic at the transition point.

Other studies of the RS localization within the same Hubbard model can be found in Ref. [183]. Other studies of the Mott transition with $1d$ model Hamiltonians, and based on the localization length λ^2, have appeared [184, 185].

The two insulating states are *qualitatively* different; by adopting the modern jargon, nowadays we could say that they are *topologically* distinct. The static ionic charges (on anion and cation) are continuous across the transition, while it was shown that dynamical (Born) effective charge on a given site changes sign [186]. The topological nature of the transition was neither clear nor made explicit in the original Refs. [186] and [146]. Here we reformulate their major findings in a different language, with some hindsight. We have discussed above, Sect. 5.7, the topological nature of $1d$ polarization in centrosymmetric systems. In topological jargon the system is $\mathbb{Z}_2$-odd for $U < 2.27t$ and $\mathbb{Z}_2$-even for $U > 2.27t$.

It is expedient to start from the pure band insulator at $U = 0$. At half filling there is a single occupied band and a doubly occupied Wannier function: the Wannier center, defined as in Eq. (5.9), sits at the anion site (i.e. the site with $\Delta < 0$). When the contribution of the classic charges is added, it is easy to verify that $P = e/2 \mod e$. Suppose now we switch on the Hubbard U continuously: the Wannier function is no longer defined, while polarization P is well defined at any U value (the transition U-value excepted). According to the most general tenet of topology when applied to electronic structure, the topological invariant cannot change insofar as the system remains insulating. At the transition point $U = 2.27t$ the system crosses a metallic state, and its polarization switches to $0 \mod e$ at large

U values: it becomes $\mathbb{Z}_2$-even.

7.5.6 Mott metal-insulator transition in a linear hydrogen chain

A "crystalline" linear array of equally spaced H atoms has long been considered as the paradigm for the Mott metal-insulator transition. At the independent-particle level, the system has an half-filled band at any interatomic distance R, and is therefore a metal. As first pointed out by Mott in 1949 [141, 143], for large interatomic distances the system must behave as an assembly of independent H atoms, and therefore it is expected to be an insulator, with antiferromagnetic correlation. Roughly speaking, the large-R wavefunction is expected to be a kind of generalization on the two-electron Heitler-London wavefunction; it is also expected that the correlated ground state $|\Psi_0\rangle$ switches from metallic to insulating at some critical R value.

A naive model for this many electron system is the Hubbard one-band model Hamiltonian, which proves to be grossly inadequate. The exact solution is known since long time [187], and does not show a metal-insulator transition.

Despite the apparent simplicity of this system, the problem is very challenging: largely inaccurate results were accreditated until recently [189, 190]. At the time of writing (2021) the state-of-the-art results are those reported in Ref. [188]. It could be naively guessed that the system can be schematized as a single-band model, where atomic orbitals different from 1s could be neglected. Instead, close to the transition, the correlated wavefunction displays a multi-band character, where bands deriving from 2s and 2p atomic orbitals play a non negligible role. We reproduce here the $|\mathfrak{z}_N|$ calculations from Ref. [188] in Fig. 7.6, which indicates the state-of-the-art transition value at $R = 1.70$.

7.5.7 Quantum Hall insulator

An electron fluid in the quantum Hall regime has zero longitudinal conductivity; we are going to prove next that the modern theory of the insulating state successfully
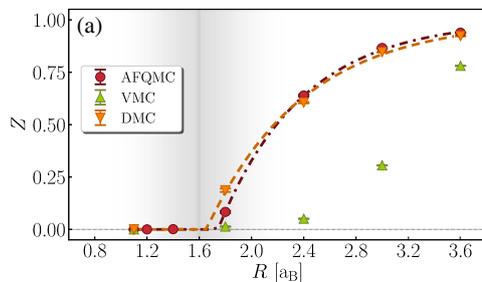
Figure 7.6: Metal-insulator transition in H-chains: $|\mathfrak{z}_N|$ as a function of the interatomic distance R for different correlated wavefunctions at $N = 40$. The plot indicates a transition R value of 1.70 bohr. After Ref. [188].

131

addresses even this exotic kind of insulator.

Soon after the experimental discovery [44], Laughlin [46] gave evidence that the electrons are kept in the quantum Hall regime by a certain amount of substrate disorder. Here we consider instead the academic case of $2d$ noninteracting electrons in a flat substrate potential, where all formulæ can be evaluated analytically.

The system is insulating whenever the Fermi level μ is is in a gap between Landau levels, i.e. at integer fillings; we consider here only the lowest Landau level (LLL). The electrons are spin-polarized because of Zeeman splitting, larger than the spacing between Landau levels. We define the magnetic length

$$\ell = \left(\frac{\hbar c}{eB} \right)^{\frac{1}{2}}; \tag{7.116}$$

we set $z = (x - iy)/\ell$ and we adopt the symmetric gauge $\mathbf{A}(\mathbf{r}) = \mathbf{B} \times \mathbf{r}/2$. In this gauge the LLL orbitals and the corresponding projector are [48]:

$$\varphi_m(z) = \frac{1}{\sqrt{2\pi 2^m m!}\, \ell} z^m e^{-|z|^2/4}, \qquad \langle z|\mathcal{P}|z'\rangle = \sum_m \varphi_m(z)\varphi_m^*(z'). \tag{7.117}$$

At complete filling the sum of the series yields

$$\langle z|\mathcal{P}|z'\rangle = \frac{1}{2\pi\ell^2} e^{-|z|^2/4} e^{-|z'|^2/4} \sum_{m=0}^{\infty} \frac{1}{m!} \left(\frac{zz'^*}{2} \right)^m \tag{7.118}$$

$$= \frac{1}{2\pi\ell^2} e^{-|z-z'|^2/4} e^{(zz'^* - z^*z')/4} \tag{7.119}$$

$$|\langle \mathbf{r}|\mathcal{P}|\mathbf{r}'\rangle| = \frac{1}{2\pi\ell^2} e^{-|\mathbf{r}-\mathbf{r}'|^2/(4\ell^2)}. \tag{7.120}$$

The electron density is therefore uniform, with $n_0 = 1/(2\pi\ell^2)$, and the decay of $|\langle \mathbf{r}|\mathcal{P}|\mathbf{r}'\rangle|$ in the relative coordinate is Gaussian, much faster than in a band insulator (where it is quasi-exponential): therefore the metric is convergent for sure.

It is instructive to evaluate the RS localization length: for single occupancy Eq. (7.88) yields

$$\lambda^2 = \mathfrak{g}_{xx} = \frac{1}{2N} \int d\mathbf{r}\, d\mathbf{r}'\, (x - x')^2 \, |\langle \mathbf{r}|\mathcal{P}|\mathbf{r}'\rangle|^2 \tag{7.121}$$

$$= \frac{1}{4n_0} \int d\mathbf{r}\, r^2 |\langle \mathbf{0}|\mathcal{P}|\mathbf{r}\rangle|^2 = \frac{1}{2}\ell^2. \tag{7.122}$$

For $B \to 0$ the system becomes a free-electron gas in $2d$, hence metallic. In fact after Eq. (7.116) the magnetic length ℓ diverges, and the RS localization length λ diverges as well.

132

We may also prove that—according to the modern theory—a partially filled LLL yields a metallic behavior. For instance, a possible state at half filling obtains by occupying the odd-m orbitals only:

$$\langle z|\mathcal{P}|z'\rangle = \sum_{m=0}^{\infty} \varphi_{2m+1}(z)\varphi_{2m+1}^*(z') = \frac{1}{2\pi\ell^2}\mathrm{e}^{-|z|^2/4}\mathrm{e}^{-|z'|^2/4}\sinh(zz'^*/2), \qquad (7.123)$$

which does not go to zero at large relative coordinates (e.g. for $z' = -z$). Hence Eq. (7.121) diverges.

Chapter 8

Anomalous Hall conductivity

8.1 Generalities

Edwin Hall discovered the eponymous effect in 1879; two years later he discovered the anomalous Hall effect in ferromagnetic metals. The latter is, by definition, the Hall effect in absence of a macroscopic **B** field. Nonvanishing transverse conductivity requires breaking of T-symmetry: in the normal Hall effect the symmetry is broken by the applied **B** field; in the anomalous one it is spontaneously broken, for instance by the development of ferromagnetic order. The theory of anomalous Hall conductivity in metals has been controversial for many years; since the early 2000s it became clear that, besides extrinsic effects, there is also an intrinsic contribution, which can be expressed as a geometrical property of the occupied Bloch manifold in the pristine crystal. The classical review on the topic is Ref. [12].

The other epochal revolution in this field is the discovery—100 years later—of the quantum Hall effect, due to von Klitzing and collaborators in 1980 [44], Nobel prize 1985. As discussed above in Sect. 4.2.4, the celebrated TKNN (Thouless, Kohmoto, Nightingale, and den Nijs) paper [57], marks the debut of topology (and geometry as well) in electronic structure: when expressed in natural units, the quantized Hall conductivity is a Chern number (a.k.a. $\mathbb{Z}$ invariant).

In presence of a macroscopic magnetic field the Hamiltonian cannot be periodical—neither lattice periodical nor Born-von-Kàrmàn periodical—in any gauge; other unusual features were well known before 1980, like the occurrence of Landau levels, and the famous Hofstadter butterfly [191].

For many years it was tacitly assumed that a macroscopic **B** field were necessary in order to make the quantum Hall effect possible. The breakthrough came from Duncan Haldane in 1988 [192]: by means of a model Hamiltonian in $2d$ he showed that a topologically non trivial ground state—and quantized Hall conductivity— may exist even in absence of a field. In this model material (which I baptized "Haldanium" in some papers of mine) there are no Landau levels: the band structure

and density of states are "plain vanilla". Haldanium is the archetypical topological material i.e. (by definition) a material having a nontrivial topology in absence of an external field. Thouless and Haldane were among the recipients of the 2016 Nobel prize.

The synthesis of a $2d$ material which realizes Haldane's model was only achieved since 2013 onwards [193, 194]. Much before that date, topological insulators of a different class were synthesized: TI (T-invariant) insulators, first in $2d$, and later in $3d$ [10, 11]. The theoretical paradigm for TI insulators in $2d$ is a famous 2005 paper by Kane and Mele [195], where a $\mathbb{Z}_2$ invariant is introduced. We are not going to discuss further TI topological insulators (some features have been outlined in Sect. 6.1.1, albeit for the $3d$ case only).

8.2 Many-body theory

8.2.1 Kubo formula

We adopt once more the Hamiltonian of Eq. (7.56), reproduced here for the sake of clarity:

$$\hat{H}_{\boldsymbol{\kappa}} = \frac{1}{2m} \sum_{i=1}^{N} |\mathbf{p}_i + \frac{e}{c} \mathbf{A}(\mathbf{r}_i) + \hbar \boldsymbol{\kappa}|^2 + \hat{V}, \tag{8.1}$$

where both $\hat{V}$ and $\mathbf{A}(\mathbf{r})$ are Born-von-Kàrmàn periodic: this means that both fields $\boldsymbol{\mathcal{E}}$ and $\mathbf{B}$ average to zero over the sample. Therefore $\mathbf{A}(\mathbf{r})$ may be due to spontaneous development of ferromagnetic order, or to spin-orbit coupling to a background of ordered local moments. The Kubo formulae, Eqs. (7.21) and (7.22), are expressed in terms of the velocity operator at $\boldsymbol{\kappa} = 0$:

$$\hat{\mathbf{v}}_{\boldsymbol{\kappa}} = \frac{1}{m} \sum_{i=1}^{N} \left[\mathbf{p}_i + \frac{e}{c} \mathbf{A}(\mathbf{r}_i) + \hbar \boldsymbol{\kappa} \right]; \tag{8.2}$$

we focus here in the antisymmetric contributions to conductivity:

$$\mathrm{Re}\ \sigma_{\alpha\beta}^{(-)}(\omega) = \frac{2e^2}{\hbar L^d} \sum_{n \neq 0} \frac{\mathcal{I}_{n,\alpha\beta}}{\omega_{0n}^2 - \omega^2} \tag{8.3}$$

$$\mathrm{Im}\ \sigma_{\alpha\beta}^{(-)}(\omega) = \frac{\pi e^2}{\hbar L^d} \sum_{n \neq 0} \frac{\mathcal{I}_{n,\alpha\beta}}{\omega_{0n}} [\delta(\omega - \omega_{0n}) - \delta(\omega + \omega_{0n})]. \tag{8.4}$$

Using again Eq. (7.70) the dc transverse conductivity is easily recast in terms of the many-body Berry curvature, Eq. (7.58), at $\boldsymbol{\kappa} = 0$:

$$\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(\boldsymbol{\kappa}) = -2\,\mathrm{Im}\ \langle \partial_{\kappa_\alpha} \Psi_{0\boldsymbol{\kappa}} | \partial_{\kappa_\beta} \Psi_{0\boldsymbol{\kappa}} \rangle. \tag{8.5}$$

$$\text{Re } \sigma_{\alpha\beta}^{(-)}(0) = -\frac{e^2}{\hbar L^d}\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(0); \tag{8.6}$$

the expression holds for metals and insulators, in either $2d$ or $3d$; thy can be nozero only if T-symmetry is spontaneously broken in the given material (at $\boldsymbol{\kappa} = 0$).

8.2.2 Transverse dc conductivity

At finite ω, conductivity requires time-dependent perturbation theory, ergo Kubo formulæ. Insofar as we address dc conductivity only, Kubo formulæ are not strictly needed and Eq. (8.6) can be derived in an alternative, very meaniingful way, as I am going to show next. Suppose that $\boldsymbol{\kappa}$ is adiabatically varied in time: the instantaneous current density is then the sum of two terms: the expectation value of the current operator, and the Niu-Thouless adiabatic current [196, 13]. The latter is the many-body generalization of Eq. (4.77); the two terms therefore yield

$$
\begin{aligned}
j_\alpha &= -\frac{e}{\hbar L^d}\langle \Psi_{0\boldsymbol{\kappa}}|\,\partial_{\kappa_\alpha}\hat{H}_{\boldsymbol{\kappa}}\,|\Psi_{0\boldsymbol{\kappa}}\rangle \\
&+ \frac{ie}{L^d}(\langle\partial_{\kappa_\alpha}\Psi_{0\boldsymbol{\kappa}}|\dot{\Psi}_{0\boldsymbol{\kappa}}\rangle - \langle\dot{\Psi}_{0\boldsymbol{\kappa}}|\partial_{\kappa_\alpha}\Psi_{0\boldsymbol{\kappa}}\rangle) \\
&= -\frac{e}{L^d}\left(\frac{1}{\hbar}\partial_{\kappa_\alpha}E_{0\boldsymbol{\kappa}} - \tilde{\boldsymbol{\Omega}}_{\alpha\beta}(\boldsymbol{\kappa})\dot{\kappa}_\beta\right),
\end{aligned}
\tag{8.7}
$$

the extensive quantity $\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(\boldsymbol{\kappa})\dot{\kappa}_\beta$ is the many-electron anomalous velocity. In the static case ($\dot{\boldsymbol{\kappa}} = 0$) no dc current may flow trough an insulating sample, ergo the ground-state energy $E_{0\boldsymbol{\kappa}} = E_0$ is—as already observed above—$\boldsymbol{\kappa}$-independent; the opposite is true in metals [128].

The linear conductivity is by definition

$$\sigma_{\alpha\beta}(\omega) = \frac{\partial j_\alpha(\omega)}{\partial\mathcal{E}_\beta(\omega)} = \frac{\partial j_\alpha(\omega)}{\partial A_\beta(\omega)}\frac{dA(\omega)}{d\mathcal{E}(\omega)}; \tag{8.8}$$

since $\mathcal{E}(\omega) = i\omega A(\omega)/c$, causal inversion yields the last factor as [152]

$$\frac{dA(\omega)}{d\mathcal{E}(\omega)} = -\frac{ic}{\omega + i\eta} = -c\left[\pi\delta(\omega) + \frac{i}{\omega}\right]. \tag{8.9}$$

Here we evaluate $\partial j_\alpha(\omega)/\partial A_\beta(\omega)$ in the adiabatic limit only, hence we set in Eq. (5.11)

$$\boldsymbol{\kappa} = \frac{e}{\hbar c}\mathbf{A}(\omega), \qquad \dot{\boldsymbol{\kappa}} = -\frac{ie\omega}{\hbar c}\mathbf{A}(\omega) = -\frac{e}{\hbar}\boldsymbol{\mathcal{E}}(\omega), \tag{8.10}$$

and we keep the leading terms in ω. It follows that

$$\frac{\partial j_\alpha(\omega)}{\partial A_\beta(\omega)} \doteq -\frac{e^2}{\hbar c L^d}\left(\frac{1}{\hbar}\frac{\partial^2 E_0}{\partial\kappa_\alpha\partial\kappa_\beta} - i\omega\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(0)\right), \tag{8.11}$$

where the symbol "$\doteq$" means "equal in the dc limit". The product of Eq. (8.11) times Eq. (8.9) yields the real parts of symmetric (longitudinal) and antisymmetric (transverse) dc conductivities as:

$$\mathrm{Re}\ \sigma_{\alpha\beta}^{(+)}(\omega) \quad \doteq \quad \frac{\pi e^2}{\hbar^2 L^d} \frac{\partial^2 E_0}{\partial \kappa_\alpha \partial \kappa_\beta} \delta(\omega) = D_{\alpha\beta}\delta(\omega); \tag{8.12}$$

$$\mathrm{Re}\ \sigma_{\alpha\beta}^{(-)}(\omega) \quad \doteq \quad \mathrm{Re}\ \sigma_{\alpha\beta}^{(-)}(0) = -\frac{e^2}{\hbar L^d}\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(0). \tag{8.13}$$

We have thus retrieved Eqs. (7.11) and (8.6); the present unconventional derivation has the virtue of being easily generalizable to nonlinear dc conductivity, as it will be shown below, Sect. 8.6.

8.2.3 Chern number and quantum anomalous Hall effect

The case of a two-dimensional insulator deserves a separate discussion. The famous TKNN result, Eq. (4.24), has been later generalized—beyond the independent-electron framework—by Niu, Thouless, and Wu (hereafter quoted as NTW) [197]:

$$\sigma_{xy}^{(-)}(0) = -\frac{e^2}{h}C_1, \tag{8.14}$$

where C_1 is the many-body Chern number. In the present formulation we have assumed PBCs at any $\boldsymbol{\kappa}$, and—following Kohn [128]—we have twisted the Hamiltonian. The reverse is actually done by NTW: the Hamiltonian is kept fixed, and the boundary conditions are twisted. In order to proceed, preliminarly we need to prove that

$$C_1 = \frac{1}{2\pi} \int_0^{\frac{2\pi}{L}} d\kappa_x \int_0^{\frac{2\pi}{L}} d\kappa_x\ \tilde{\boldsymbol{\Omega}}_{xy}(\boldsymbol{\kappa}) \tag{8.15}$$

is a topological integer. In fact Eq. (8.15) is the alternative definition of the NTW Chern number within our formalism: Eq. (8.15) is quantized because it is equivalent to the integral over a torus [13]. In order to show this, we remind that in insulators the ground-state energy $E_{0\boldsymbol{\kappa}}$ is $\boldsymbol{\kappa}$-independent, and we observe that whenever the components of $\boldsymbol{\kappa} - \boldsymbol{\kappa}'$ are integer multiples of $2\pi/L$, then the state $\mathrm{e}^{i(\boldsymbol{\kappa}-\boldsymbol{\kappa}')\cdot\hat{\mathbf{r}}}|\Psi_{0\boldsymbol{\kappa}}\rangle$ is eigenstate of $\hat{H}_{\boldsymbol{\kappa}'}$ with the same eigenvalue as $|\Psi_{0\boldsymbol{\kappa}}\rangle$. The eigenstates which define $\tilde{\boldsymbol{\Omega}}_{xy}(\boldsymbol{\kappa})$ have therefore the required toroidal periodicity:

$$|\Psi_{0\boldsymbol{\kappa}'}\rangle = \mathrm{e}^{i(\boldsymbol{\kappa}-\boldsymbol{\kappa}')\cdot\hat{\mathbf{r}}}|\Psi_{0\boldsymbol{\kappa}}\rangle. \tag{8.16}$$

Since $\tilde{\boldsymbol{\Omega}}_{xy}(\boldsymbol{\kappa})$ is gauge-invariant, an arbitrary $\boldsymbol{\kappa}$-dependent phase factor may relate the two members of Eq. (5.12). It is worth stressing that in the topological case a globally smooth periodic gauge does not exist; in other words we can enforce

Eq. (5.12) as it stands (with no extra phase factor) only locally, not globally; we also notice that Eq. (8.16) may be regarded as the many-body analogue of the periodic gauge in band-structure theory [24].

Eq. (3.18) is independent of the L value, and its integrand is extensive: therefore in the large-L limit the integration domain contracts to a point:

$$C_1 = \frac{1}{2\pi} \left(\frac{2\pi}{L}\right)^2 \tilde{\boldsymbol{\Omega}}_{xy}(0). \tag{8.17}$$

By comparing this to Eq. (8.13) for $d = 2$, Eq. (8.14) is immediately retrieved.

8.2.4 Extrinsic effects

First of all I stress the quite different role of the impurities between the AHC in metals and the quantized AHC in $2d$ insulators: in the former case there must necessarily be extrinsic effects, while in the latter case extrinsic effects are ruled out. In fact—as a basic tenet of topology—any impurity has no effect on linear Hall conductivity insofar as the system remains insulating.

In a pristine metal the dc longitudinal conductivity is infinite: the Drude term is proportional to $\delta(\omega)$. Extrinsic mechanisms are necessary to warrant Ohm's law, and are accounted for by relaxation time(s) τ; in absence of T-symmetry, extrinsic effects contribute to AHC as well. Two distinct mechanisms have been identified: they go under the name of "side jump" and "skew scattering" [12]. The side-jump term is nondissipative (independent of τ). Since a crystal with impurities actually is a (very) dilute alloy, it was previously argued [198] that the sum of the intrinsic and side-jump terms can be regarded as the intrinsic (geometrical) term of the dirty sample, whose AHC is given by Eq. (8.13) as it stands, provided that the potential $\hat{V}$ includes the effect of the impurities. At the independent-electron level, the same effect can in principle be retrieved from the complementary real-space formulation of AHC [199], discussed below. The other extrinsic term (skew scattering) is dissipative, proportional to τ in the single-relaxation-time approximation, and presumably cannot be explained by means of geometrical concepts.

8.3 Independent electrons

8.3.1 Transverse dc conductivity

The above Kubo formula for transverse dc conductivity has been anticipated above, Sect. 4.2.4, for the insulating case in $2d$; but indeed the intrinsic term in metals is no different. The many-band independent-electron curvature is (for both metals

and insulators)

$$\mathbf{\Omega}_{\alpha\beta}(\mathbf{k}) = i \sum_j \theta(\mu - \epsilon_{j\mathbf{k}}) [\, \langle \partial_\alpha u_{j\mathbf{k}} | \partial_\beta u_{j\mathbf{k}} \rangle - \langle \partial_\beta u_{j\mathbf{k}} | \partial_\alpha u_{j\mathbf{k}} \rangle \,], \qquad (8.18)$$

and the independent-electron AHC of a pristine crystal is

$$\mathrm{Re}\, \sigma^{(-)}_{\alpha\beta}(0) = -\frac{e^2}{\hbar} \int_{\mathrm{BZ}} \frac{d\mathbf{k}}{(2\pi)^d} \mathbf{\Omega}_{\alpha\beta}(\mathbf{k}). \qquad (8.19)$$

This clearly generalizes the formulæ of Sect. 4.2.4.

At the independent-electron level (either Hartree-Fock or Kohn-Sham) the many-electron wavefunction is a Slater determinant of Bloch orbitals; the relationship between the many-body curvature per unit volume and the independent-particle curvature is therefore

$$\frac{1}{L^d} \tilde{\mathbf{\Omega}}_{\alpha\beta}(0) = \int_{\mathrm{BZ}} \frac{d\mathbf{k}}{(2\pi)^d} \, \mathbf{\Omega}_{\alpha\beta}(\mathbf{k}),. \qquad (8.20)$$

The equality holds in the $L \to \infty$ limit. The convergence of Eq. (8.20) with $1/L$ has been indeed investigated by means of tight-binding simulations in the simple case of a Chern insulator, where the r.h.s. is quantized: see Fig. 8.3 below. The $L \to \infty$ is implicitly understood in the l.h.s.; it is instead explicit in the r.h.s., given that the Bloch vector therein is a continuous variable.

When $\boldsymbol{\kappa} \neq 0$ is set in Kohn's Hamiltonian $\hat{H}_{\boldsymbol{\kappa}}$, the corresponding Kohn-Sham periodic orbitals $|u_{j\mathbf{k}}\rangle$ are eigenstates of the single-particle Hamiltonian

$$\mathrm{e}^{-i\mathbf{k}\cdot\mathbf{r}} H_{\boldsymbol{\kappa}} \mathrm{e}^{i\mathbf{k}\cdot\mathbf{r}} = \frac{1}{2m} \left[\mathbf{p} + \frac{e}{c}\mathbf{A}(\mathbf{r}) + \hbar\mathbf{k} + \hbar\boldsymbol{\kappa} \right]^2 + V_{\mathrm{KS}}, \qquad (8.21)$$

where V_{KS} is the Kohn-Sham potential, hence

$$\frac{1}{L^d} \tilde{\mathbf{\Omega}}_{\alpha\beta}(\boldsymbol{\kappa}) = \sum_j \int_{\mathrm{BZ}} \frac{d\mathbf{k}}{(2\pi)^d} \mathbf{\Omega}_{\alpha\beta}(\mathbf{k} + \boldsymbol{\kappa}). \qquad (8.22)$$

8.3.2 AHC in gauge-invariant form

The Berry curvature of the occupied manifold, defined as in Eq. (8.18), allows to write the AHC in form of a BZ integral, which applies to both insulators and metals: Eq. (8.26). The curvature is gauge-invariant in the generalized sense, i.e. is invariant for unitary transformations of the occupied $|u_{j\mathbf{k}}\rangle$ at any given $\mathbf{k}$. The expression in Eq. (8.18), however, requires the so-called "Hamiltonian gauge".

It is possible to write the curvature in a manifestly gauge-invariant trace form, whose entries are—instead of the orbitals $|u_{j\mathbf{k}}\rangle$—the Bloch projector $\mathcal{P}_{\mathbf{k}}$, Eq. (7.3), and its $\mathbf{k}$-derivatives, Eq. (7.4). The expression is:

$$\boldsymbol{\Omega}_{\alpha\beta}(\mathbf{k}) = i\mathrm{Tr}\ \{\mathcal{P}_{\mathbf{k}}\,[\,\partial_\alpha\mathcal{P}_{\mathbf{k}}, \partial_\beta\mathcal{P}_{\mathbf{k}}]\}, \tag{8.23}$$

and holds as it stands for both insulators and metals. The singularity in the first line of Eq. (7.4) does not affect Eq. (8.23), owing to antisymmetrization; the proof of Eq. (8.23) is straightforward by direct calculation.

We have shown already this expression in Eq. (3.48); a similar form holds for the metric as well. Taking them altogether the metric-curvature tensor is cast in compact forma as

$$\mathcal{F}_{\alpha\beta}(\mathbf{k}) = \mathrm{Tr}\ \{\mathcal{P}_{\mathbf{k}}\,(\partial_\alpha\mathcal{P}_{\mathbf{k}})(\partial_\beta\mathcal{P}_{\mathbf{k}})\}. \tag{8.24}$$

8.3.3 Metals

The theory of the AHC in metals has been controversial for many years, and important developments are quite recent [198, 199, 200]. A key turning point occurred in the early 2000s [201, 202], when it became clear that, besides the extrinsic mechanisms, there is an important intrinsic contribution which is geometrical in nature. We have previously defined the Berry curvature for band insulators, Eqs. (3.38) and (3.48); it is expedient to generalize it for band metals as

$$\boldsymbol{\Omega}_{\alpha\beta}(\mathbf{k}) = i\sum_j \theta(\mu - \epsilon_{j\mathbf{k}})[\,\langle\partial_\alpha u_{j\mathbf{k}}|\partial_\beta u_{j\mathbf{k}}\rangle - \langle\partial_\beta u_{j\mathbf{k}}|\partial_\alpha u_{j\mathbf{k}}\rangle\,], \tag{8.25}$$

and we observe that $\boldsymbol{\Omega}_{\alpha\beta}(\mathbf{k})$ is smooth in insulators and piecewise continuous (hence integrable) in metals.

We address solely the geometrical (a.k.a. intrinsic) contrubution here; its expression is

$$\mathrm{Re}\ \sigma_{\alpha\beta}^{(-)}(0) = -\frac{e^2}{\hbar}\int_{\mathrm{BZ}}\frac{d\mathbf{k}}{(2\pi)^d}\ \boldsymbol{\Omega}_{\alpha\beta}(\mathbf{k}), \tag{8.26}$$

where $\boldsymbol{\Omega}_{\alpha\beta}(\mathbf{k})$ is the Berry curvature of the occupied manifold, Eq. (8.18); this is in fact the nonquantized version of the famous TKNN formula, and it holds as it stands in either $2d$ or $3d$. It is easily verified that for a $2d$ insulator Eq. (8.26) coincides indeed with Eq. (4.24).

Given that the Fermi surface is symmetrical under $\mathbf{k} \to -\mathbf{k}$, the symmetry considerations of Sect. 3.8 show that Eq. (8.26) can be nonzero only T-symmetry is absent, while inversion symmetry is irrelevant. The typical case studies are the ferromagnetic metals, whose ground state breaks indeed T-symmetry in *absence* of a macroscopic $\mathbf{B}$ field.

First-principle calculations were performed for Ni, Cu, and Fe, as well as or for some oxides. The intrinsic geometric contribution appears to be the dominant one. These calculations also pointed out the crucial role played by avoided crossings of the bands near the Fermi surface, which induce a very spiky behavior of the Berry curvature in the BZ. More than 10^6 **k** points where used in Ref. [203] in order to perform the integration in Eq. (8.26); a more efficient strategy was devised later [204].

A noninteracting (e.g. KS) many-electron system is a trivial example of a Fermi liquid. Haldane [205] pointed out that the very basic tenet of Landau's Fermi-liquid theory is that charge transport involves only quasiparticles with energies within $k_{\rm B}T$ of the Fermi level. This is apparently at odds with Eq. (8.26), which is an integration over the whole occupied Fermi sea. The two viewpoints can be reconciled, essentially via an integration by parts [205]. Even this alternative form has been implemented in first-principle calculations [161].

8.3.4   Chern invariant in band insulators

Band insulators where the Chern invariant is nonzero are called "Chern insulators" (normal insulators otherwise); they go under the equivalent name of QAHE (quantum anomalous Hall effect) insulators. We emphasize that the nonvanishing of the Chern number prevents the existence of a smooth periodic gauge, Eq. (3.34), across the whole reciprocal cell (or BZ): if, for instance, Eq. (3.34) is enforced on the reciprocal cell boundary, an "obstruction" will show up at some point inside [24]. Because of this same reason, exponentially localized Wannier functions do not exist [206, 207]; this is at variance with normal insulators where localized Wannier functions exist, and can even be chosen as exponentially localized [19, 56, 208].

Digression: Bulk-boundary correspondence

As discussed throughout these Notes, geometrical and topological properties of a given system manifest themselves in reciprocal space. The concepts of reciprocal space and of **k** vector are based on PBCs: the sample has no boundaries by construction. However, for the same physical system, one could consider a bounded sample within OBCs. In this case there is no **k** vector to speak of: the geometrical properties manifest themselves as surface (in $3d$) or edge (in $2d$) properties. Bulk-boundary correspondence is the hallmark of geometry and topology in electronic structure.

The archetype of bulk-boundary correspondence is the IQHE. When a toroidal geometry is adopted there is no boundary, and the fingerprint of nontrivial topology is the Chern number. If instead the $2d$ electron fluid in the quantum Hall regime has the shape of a ribbon, then there are edge states which carry a chiral current.

However, these edge states are a bulk property and are "topologically protected". Similar considerations apply to Chern insulators in $2d$.

For $2d$ T-invariant (TI) topological insulators the edge states cannot carry any net current; instead there are counterpropagating spin currents which, once more, are topologically protected: one speaks then of "quantum spin Hall effect". In the $3d$ case, there are topologically protected surface states, whose nature is nontrivial.

There is a subtle difference between Chern insulators and TI insulators. In the former case the quantization of the (electrical) edge current is exact; in the latter case the quantization of the spin edge current is only approximate, because spin is not conserved in these materials [209].

The bulk-boundary correspondence also occurs in the case of polarization, where the topologically protected boundary quantity is not current but charge. Here one has to pay attention to the fact that polarization is the sum of an electronic and a ionic term: only the former term is expressed as a $\mathbf{k}$ integral. However, the bulk polarization determines the surface charge modulo the "quantum" (Sec. 5.6.1). In the simple case where the bulk is centrosymmetric, topology requires the charge per surface cell to be an integer or an half integer (see also Sec. 5.7). We stress that, quite often, a polar surface is metallic and topology does not apply.

Computing the Chern number

The computation of a Chern number proceeds similarly to what presented in Sec. 4.1.4 for a toy-model Hamiltonian. Suppose we discretize the reciprocal cell with a regular mesh, as in Fig. 8.1. We start enforcing the periodic gauge $|u_{j\mathbf{k}+\mathbf{G}}\rangle = \mathrm{e}^{-i\mathbf{G}\cdot\mathbf{r}}|u_{j\mathbf{k}}\rangle$ on the boundary, i.e. for all the couples of boundary points which are related by a reciprocal lattice vector. At all the interior points the gauge is chosen by the diagonalization routine and is therefore erratic. The mesh actually shown in Fig. 8.1 would actually require $8 \times 8 = 64$ Hamiltonian diagonalizations.
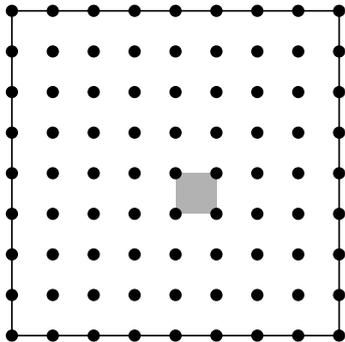

Figure 8.1: Discretization of the reciprocal cell. The periodic gauge is enforced on the boundary; otherwise the gauge is unspecified and possibly erratic. The Chern number is the sum of the Berry phases computed on the small squares.

We recall that the curvature is the Berry phase per unit reciprocal area, and we compute it on each of the small squares using the four-point discrete formula

$$\gamma = -\text{Im } \ln \langle u_{\mathbf{k}_1}|u_{\mathbf{k}_2}\rangle\langle u_{\mathbf{k}_2}|u_{\mathbf{k}_3}\rangle\langle u_{\mathbf{k}_3}|u_{\mathbf{k}_4}\rangle\langle u_{\mathbf{k}_4}|u_{\mathbf{k}_1}\rangle, \tag{8.27}$$

which applies to the single-band case. In the many-band case this is replaced by

$$\gamma = -\text{Im } \ln \det S(\mathbf{k}_1,\mathbf{k}_2)S(\mathbf{k}_2,\mathbf{k}_3)S(\mathbf{k}_3,\mathbf{k}_4)S(\mathbf{k}_4,\mathbf{k}_1), \tag{8.28}$$

where S is the overlap matrix

$$S_{nn'}(\mathbf{k}_s,\mathbf{k}_{s'}) = \langle u_{n\mathbf{k}_s}|u_{n'\mathbf{k}_{s'}}\rangle; \tag{8.29}$$

see Eq. (3.55). One has to choose the "Im log" branch with γ in $[-\pi,\pi]$.

The Chern number C_1 is the sum of all the γs, covering the whole reciprocal cell (64 in Fig. 8.1). In our discontinuous approach, the obstruction is not located at any particular $\mathbf{k}$ point, although its effect becomes apparent after all γs are summed. we remind that—as discussed in Sec. 4.1.4—the discretised Chern number is always an exact integer, and *not* an approximate integer, even for a coarse mesh. This feature has been apparently first pointed out in 2005 [210].

Chern number as the curvature for unit area

We have seen above that for an insulating $2d$ system the Chern number is proportional to the many-body Berry curvature per unit area, Eq. (8.17),

$$C_1 = 2\pi\frac{\hat{\mathbf{\Omega}}(0)}{A}, \tag{8.30}$$

where $A = L^2$ is the system area and the thermodynamic limit is implicit. In the independent-electron case the wavefunction is a Slater determinant, and the many-body Berry curvature is the BZ integral of the band-structure curvature: see Eq. (8.20).

Here we investigate what happens when the BZ integral is discretized, which corresponds to evaluating $\hat{\mathbf{\Omega}}(0)/A$ at finite size. Eventually, the Chern number can be obtained from a *single* Hamiltonian diagonalization for a large supercell—either crystalline or disordered—and no derivative is actually involved.

This bears some relationship to the so-called single-point Berry phase, discussed in Sec. 5.6.2. Here we follow the approach of Ref. [211]: to give a flavor of how it works, we start from Fig. 8.1, where one needs 64 independent diagonalizations, as said above. One may look at the same system by doubling the elementary cell in each direction, in which case the number of occupied bands increases fourfold, and the new reciprocal cell has one quarter of the original area. Now the calculation
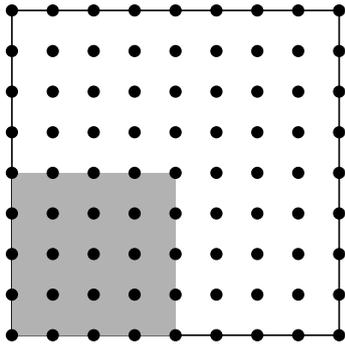
Figure 8.2: Upon doubling the elementary cell, the reciprocal-cell area is divided by 4, and the number of occupied bands is multiplied by 4.
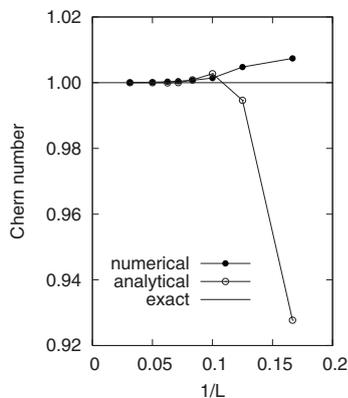

Figure 8.3: Convergence of the Chern number for topological crystalline Haldanium, as computed from a single Hamiltonian diagonalization in a supercell geometry; the **k**-derivatives are discretized in two alternative ways. After Ref. [211].

of C_1 requires $64/4 = 16$ independent diagonalizations; the algorithm of Eq. (8.28) yields a C_1 value identical to previous one (up to computer accuracy), because we are folding *the same orbitals* in a smaller reciprocal cell (or BZ). We can go on with the folding, diagonalizing in 64, 16, 4, 1 points. If the supercell size L is large enough, the C_1 value computed via a single Hamiltonian diagonalization and (super)lattice-periodical orbitals is very close to the thermodynamic limit, as perspicuously shown in simulations on a simple test case: Fig. 8.3. Once the large supercell is adopted, disorder can be introduced at no extra cost; since C_1 is a topological invariant, disorder cannot change its topological value insofar as the system remains insulating.

8.3.5 Hermaphrodite orbitals

Hermaphrodite orbitals were first introduced by us in Ref. [148]. They are nowadays a common tool to address TI topological insulators. Unfortunately, they have been rebaptized "hybrid Wannier functions" [212].

It is expedient to start with a finite crystallite within OBCs, with N electrons

144

in a singlet ground state:

$$\rho = 2\,P = 2\sum_{i=1}^{N/2} |\varphi_i\rangle\langle\varphi_i|. \tag{8.31}$$

The average quadratic spread in the x direction is by definition

$$\lambda_{xx}^2 = \frac{2}{N}\sum_{i=1}^{N/2}(\langle\varphi_i|x^2|\varphi_i\rangle - \langle\varphi_i|x|\varphi_i\rangle^2). \tag{8.32}$$

We recast this identically as:

$$\lambda_{xx}^2 = \frac{2}{N}\sum_i \langle\varphi_i|\,x\,(\,1 - \sum_j |\varphi_j\rangle\langle\varphi_j|\,)\,x\,|\varphi_i\rangle + \frac{2}{N}\sum_{i\neq j}|\langle\varphi_i|x|\varphi_j\rangle|^2. \tag{8.33}$$

The first term in Eq. (8.33) is gauge invariant, since we can identically write:

$$\lambda_{xx}^2 = \frac{2}{N}\mathrm{Tr}\ xPx(1-P) + \frac{2}{N}\sum_{i\neq j}|\langle\varphi_i|x|\varphi_j\rangle|^2. \tag{8.34}$$

The gauge–invariant term in Eq. (8.34) coincides with the Resta-Sorella squared localization length, for the appropriate geometry.

If we look for the orbitals which minimize the average spread in the x direction, the solution, after Eq. (8.34), is provided by those orbitals which diagonalize the position operator x, projected over the occupied manifold, i.e. the operator PxP. Obviously, a set of orthonormal orbitals which diagonalize it can always be found, since PxP is a Hermitian operator. The quadratic spread of these orbitals is the minimum and equals then the gauge-invariant spread.

If we now switch to the thermodynamic limit for a crystalline system, translational symmetry implies $P(\mathbf{r},\mathbf{r}') = P(\mathbf{r}+\mathbf{R},\mathbf{r}'+\mathbf{R})$. For the sake of simplicity we limit ourselves to a rectangular lattice, where $\mathbf{R} = (X,Y,Z)$, and each of the X,Y,Z are one-dimensional lattices. It is then obvious that the eigenstates of PxP can be labelled with a Bloch vector in the yz directions. If we use the notation $\chi_{s\,k_y k_z}(\mathbf{r})$ for a generic eigenstate of PxP, then it obeys the Bloch theorem in the form

$$\chi_{s\,k_y k_z}(x, y + Y, z) = \mathrm{e}^{ik_y Y}\chi_{s\,k_y k_z}(x, y, z), \tag{8.35}$$

and analogous in the z direction.

We address the simple case of a single occupied band. If $\chi_{0\,k_y k_z}(x,y,z)$ is eigenstate of PxP with eigenvalue x_0, then its lattice-translate $\chi_{0\,k_y k_z}(x - X, y, z)$ is also eigenstate of PxP, with eigenvalue $x_0 + X$: this is reminiscent of WFs, see

Figure 8.4: Two-masted vessels. Brig: both masts are square rigged. Schooner: both masts are fore-and-aft rigged (gaff sails and topsails here). Hermaphrodite brig: the foremast is square rigged, the mainmast is fore-and-aft rigged.

Eq. (5.7). The proof is very simple: $P(x - X)P = PxP - XP$, and XP on the occupied manifold is just a constant multiplication by X.

In the many-band case we need—as for WFs—a band index j and a $1d$ cell index X. All hermaphrodite orbitals are obtained by lattice translation from the central cell ones:

$$\langle x, y, z | \chi_{j\,X\,k_y k_z} \rangle = \langle x - X, y, z | \chi_{j\,0\,k_y k_z} \rangle; \qquad (8.36)$$

a glance at Eq. (5.7) explains the notations and the meaning. Notice that the hermaphrodite-orbital center x_0 *does depend* on the (k_y, k_z) Bloch vector, as well as on the band index j.

The name "hermaphrodite orbitals" is therefore pretty clear (see Fig 8.4): they are Bloch-like in the yz direction, and Wannier-like in the x direction. The present construction guarantees that they are maximally localized in the x direction and their quadratic spread is equal to the xx component of the (gauge-invariant) localization tensor. Similarly to standard Bloch and Wannier orbitals, our hermaphrodite orbitals are an orthonormal basis for the occupied manifold.

Chern number and hermaphrodite orbitals

Here we address the simple case of a $2d$ square lattice of constant a and a single occupied band. We simplify notations by indicating with $\chi_{k_y}(x, y)$ the central-cell (k_y-dependent) orbital; the complete set in the occupied manifold is, after Eq. (8.36), $\chi_{m\,k_y}(x, y) = \chi_{k_y}(x - ma, y)$, where $m \in \mathbb{Z}$ and $k_y \in [0, 2\pi/a)$. The inverse transformation yields the Bloch orbitals as

$$
\begin{aligned}
\psi_{\mathbf{k}}(\mathbf{r}) &= \sum_m e^{imk_x a} \chi_{k_y}(x - ma, y) & (8.37) \\
u_{\mathbf{k}}(\mathbf{r}) &= \sum_m e^{-ik_x(x-ma)} \tilde{\chi}_{k_y}(x - ma, y), \quad \tilde{\chi}_{k_y}(x, y) = e^{-ik_y y} \chi_{k_y}(x, y).
\end{aligned}
$$

146

The $u_{\mathbf{k}}$ orbitals are clearly lattice-periodical and normalized over the unit cell; their $\mathbf{k}$ derivative yields

$$\partial_1 u_{\mathbf{k}}(\mathbf{r}) = -i \sum_m e^{-ik_x(x-ma)}(x-ma)\,\tilde{\chi}_{k_y}(x-ma, y),$$

$$\partial_2 u_{\mathbf{k}}(\mathbf{r}) = \sum_{m'} e^{-ik_x(x-m'a)}\frac{\partial}{\partial k_y}\tilde{\chi}_{k_y}(x-m'a, y),$$

$$\langle \partial_1 u_{\mathbf{k}}|\partial_2 u_{\mathbf{k}}\rangle = i \sum_m e^{-ik_x ma}\langle \tilde{\chi}_{m\,k_y}|(x-ma)|\frac{\partial}{\partial k_y}\tilde{\chi}_{0\,k_y}\rangle, \tag{8.38}$$

where "$\langle\ \rangle$" means $\mathbf{r}$ integral in $(-\infty, \infty) \times (0, a)$. The Berry curvature and the Chern number are therefore

$$\begin{aligned}
\mathbf{\Omega}(\mathbf{k}) &= -2\,\mathrm{Im}\,\langle \partial_1 u_{\mathbf{k}}|\partial_2 u_{\mathbf{k}}\rangle \\
&= -2\,\mathrm{Re}\,\sum_m e^{-ik_x ma}\langle \tilde{\chi}_{m\,k_y}|(x-ma)|\frac{\partial}{\partial k_y}\tilde{\chi}_{0\,k_y}\rangle,
\end{aligned}$$

$$\begin{aligned}
C_1 &= -\frac{1}{\pi}\mathrm{Re}\,\int_0^{2\pi/a}dk_x \int_0^{2\pi/a}dk_y \sum_m e^{-ik_x ma}\langle \tilde{\chi}_{m\,k_y}|(x-ma)|\frac{\partial}{\partial k_y}\tilde{\chi}_{0\,k_y}\rangle \\
&= -\frac{2}{a}\int_0^{2\pi/a}dk_y\,\mathrm{Re}\,\langle \tilde{\chi}_{k_y}|x|\frac{\partial}{\partial k_y}\tilde{\chi}_{k_y}\rangle, \tag{8.39}
\end{aligned}$$

where $|\tilde{\chi}_{k_y}\rangle \equiv |\tilde{\chi}_{0\,k_y}\rangle$ is the central-cell orbital. If we define x_0 as its (k_y dependent) center, then

$$\begin{aligned}
x_0(k_y) &= \langle \tilde{\chi}_{k_y}|x|\tilde{\chi}_{k_y}\rangle, \\
\frac{dx_0}{dk_y} &= 2\,\mathrm{Re}\,\langle \tilde{\chi}_{k_y}|x|\frac{\partial}{\partial k_y}\tilde{\chi}_{k_y}\rangle \tag{8.40}
\end{aligned}$$

$$C_1 = -\frac{1}{a}\int_0^{2\pi/a}dk_y\,\frac{dx_0}{dk_y} = -\frac{1}{a}[\,x_0(2\pi/a) - x_0(0)\,]. \tag{8.41}$$

The hermaphrodite orbitals carry therefore a very clear topological signature. In a topologically trivial $2d$ insulator their center is periodical in k_y; while instead in the nontrivial case their center shifts by $-C_1 a$ when $k_y \to k_y + 2\pi/a$. More generally, not only the center but even the function $\tilde{\chi}_{m\,k_y}(x, y)$ enjoys an analogous property.

As for the $\mathbf{k}$ space periodicity of the Bloch functions, Eqs. (8.37) and (8.41) yield

$$\begin{aligned}
\psi_{k_x+2\pi/a,k_y}(\mathbf{r}) &= \psi_{k_x,k_y}(\mathbf{r}) \\
\psi_{k_x,k_y+2\pi/a}(\mathbf{r}) &= e^{iC_1 k_x a}\,\psi_{k_x,k_y}(\mathbf{r}). \tag{8.42}
\end{aligned}$$

this confirms that in the Chern case the gauge cannot be periodical over the reciprocal cell. The gauge implicit in Eq. (8.37) can be called "cylindrical": periodic in k_x but not in k_y.

Lowest Landau level

We consider a system of noninteracting electrons in a flat substrate potential and in a normal magnetic field. Although the system is translationally invariant in the (xy) plane, both the Hamiltonian and the density matrix P are *not* invariant. Despite this, the orbitals of e.g. the lowest Landau level (LLL) share many of the properties of the hermaphrodite orbitals.

If the LLL is fully occupied (i.e. at filling $\nu = 1$) the electron density is uniform and equal to $n_0 = 1/(2\pi\ell^2)$, where $\ell = (\hbar c/eB)^{1/2}$ is the magnetic length; the *modulus* of the density matrix has the translationally invariant expression (for single occupancy)

$$|P(\mathbf{r}, \mathbf{r}')| = n_0 e^{-|\mathbf{r}-\mathbf{r}'|^2/(4\ell^2)}, \tag{8.43}$$

and the cumulant second moment is clearly finite. The trace $\langle x^2 \rangle_c + \langle y^2 \rangle_c$ of the localization tensor is in fact equal to ℓ^2, the squared magnetic length, and the minimum quadratic spread in the x direction is $\ell^2/2$.

The LLL orbitals in the Landau gauge are labeled by a one-dimensional wave vector $q \in (-\infty, \infty)$ and have the form

$$\psi_q(\mathbf{r}) = e^{-iqy}\varphi(x - q\ell^2), \tag{8.44}$$

where $\varphi(x)$ is the normalized ground eigenfunction of a $1d$ harmonic oscillator with frequency $\omega_c = eB/m_e c$

$$\varphi(x) = \left(\frac{1}{\pi\ell^2}\right)^{1/4} e^{-x^2/(2\ell^2)}. \tag{8.45}$$

The quadratic spread of this function is indeed $\ell^2/2$; the $\psi_q(\mathbf{r})$ are therefore eigenstates of PxP, and are the magnetic analogue of the maximally localized hermaphrodite orbitals. Notice that in Eq. (8.44) $\psi_q(\mathbf{r})$ has a plane-wave-like normalization over y.

To make contact with our notations for hermaphrodite orbitals it is enough to define $a = \sqrt{2\pi}\,\ell$ and

$$q = k_y + 2\pi m/a, \quad k_y \in [0, 2\pi/a), \quad m \in \mathbb{Z}$$
$$\tilde{\chi}_{mk_y}(x, y) = \frac{1}{a}e^{-i2\pi my/a}\varphi(x - k_y\ell^2 - 2\pi m\ell^2/a)$$
$$= \frac{1}{a}e^{-i2\pi my/a}\varphi(x - k_y a^2/(2\pi) - ma); \tag{8.46}$$

as in the previous Section, the $\tilde{\chi}$ orbitals are normalized over $(-\infty, \infty) \times (0, a)$.

Finally, we verify the behavior of $\tilde{\chi}$ vis-a-vis the Chern number. The center of the central cell orbital is $x_0 = k_y a^2/(2\pi)$; if k_y is increased by $2\pi/a$, then x_0 increases by a, as it must be: we remind that $C_1 = -1$ in the LLL (sign conventions are not uniform across the literature).
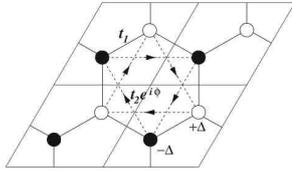
Figure 8.5: Four unit cells of crystalline Haldanium [192]. Filled (open) circles denote sites with $E_0 = -\Delta$ $(+\Delta)$. Solid lines connecting nearest neighbors indicate a real hopping amplitude t_1; dashed arrows pointing to a second-neighbor site indicates a complex hopping amplitude $t_2 e^{i\phi}$. Arrows indicate sign of the phase ϕ for second-neighbor hopping.

8.4 Haldanium

The Haldane model Hamiltonian is comprised of a 2d honeycomb lattice with two tight-binding sites per primitive cell with site energies $\pm\Delta$, real first-neighbor hopping t_1, and complex second-neighbor hopping $t_2 e^{\pm i\phi}$, as shown in Fig. 8.5. Within this two-band model, one deals with insulators by taking the lowest band as occupied; many years after its first occurrence, the same Hamiltonian has been used also to model a 2d metal where T-invariance is absent (it is enough to set the Fermi level μ across a band).

The appeal of the model is that the vector potential and the Hamiltonian are lattice periodical and the single-particle orbitals have the usual Bloch form. Essentially, the microscopic magnetic field can be thought as staggered—i.e. up and down in different regions of the cell—but its cell average vanishes. The T-breaking can be pictorially understood as induced by magnetic point-dipoles (normal to the plane) at the center of each hexagon; according to Fig. 9.1 below, such dipoles do not generate a microscopic **B** field.

As a consequence, the band structure and the density of states do not show any exotic feature (at variance with the case where the T-breaking is due to a macroscopic **B** field). The Haldane Hamiltonian has been used as a workhorse in many simulations, providing invaluable insight into topological features of the electronic wavefunction [192, 211, 206, 213, 214, 215], as well as into features of orbital magnetization [216, 217, 211, 218].
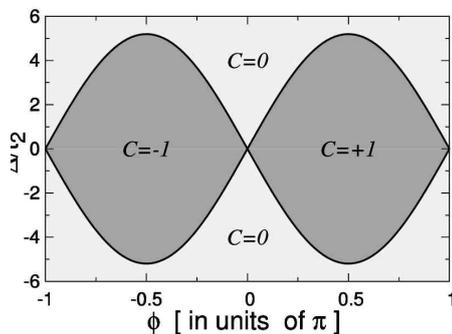
Figure 8.6: Chern number C_1 of the Haldane model at half filling as a function of the parameters ϕ and Δ/t_2; the plot is drawn for $t_1 = 1$ and $t_2 = 1/3$. At $\phi = 0$ Haldanium is a model for graphene and hexagonal boron nitride.

As a function of the flux parameter ϕ, the system—at half filling—undergoes a transition from a normal insulator ($C_1 = 0$) to Chern insulator ($C_1 = \pm 1$). Its phase diagram is shown in Fig. 8.6. On the vertical $\phi = 0$ axis the Hamiltonian is T-invariant and Haldanium becomes a model for hexagonal boron nitride; the center of the plot ($\Delta = 0, \phi = 0$) corresponds to graphene. Notice also that on the horizontal $\Delta = 0$ axis Haldanium is a nonpolar and centrosymmetric insulator (except at $\phi = 0$).

According to the theory of the insulating state, the Resta-Sorella λ^2—or equivalently the gauge-invariant quadratic spread Ω_I—is finite at half filling. This is confirmed by the simulations of Ref. [206], where the actual value of Ω_I is computed. We remind that, despite Ω_I being finite, localized Wannier functions (with finite quadratic spread) do not exist in Chern insulators.

Both insulating Haldanium and quantum Hall insulators display quantized transverse conductivity; and both are localized in the above sense. It is worth pointing out, though, that the decay of the density matrix is qualitatively different: exponential in Haldanium [206] vs. Gaussian in the (noninteracting) quantum Hall case, Eq. (8.43).

Chern insulators remained a curiosity of academic interest only for many years; no actual material belonging to this class was actually synthesized until 2013. The hallmark of a 2d Chern insulator is quantum anomalous Hall effect (QAHE). The effect was first observed in 2013 in Cr-doped $(B,Sb)_2Te_3$ thin films [193]; the quantization was later observed with much higher precision in V-doped $(B,Sb)_2Te_3$-doped $(B,Sb)_2Te_3$ [194].

8.4.1 Exact diagonalization; skyrmion-like invariant

.

The most general 2×2 tight-binding Hamiltonian $H_\mathbf{k}$ is a $\mathbf{k}$-dependent linear combination of the identity and the Pauli matrices:

$$H_\mathbf{k} = \eta(\mathbf{k}) + \mathbf{d}(\mathbf{k}) \cdot \vec{\sigma}, \tag{8.47}$$

and the lowest band is

$$\epsilon_\mathbf{k} = \eta(\mathbf{k}) - |\mathbf{d}(\mathbf{k})|. \tag{8.48}$$

The ground state projector at half filling has a very simple expression in terms of $\hat{\mathbf{d}}(\mathbf{k}) = \mathbf{d}(\mathbf{k})/|\mathbf{d}(\mathbf{k})|$:

$$P_\mathbf{k} = \frac{1}{2}[1 - \hat{\mathbf{d}}(\mathbf{k}) \cdot \vec{\sigma}], \qquad \partial_{k_\alpha} P_\mathbf{k} = -\frac{1}{2}\partial_{k_\alpha}\hat{d}_\gamma(\mathbf{k})\sigma_\gamma. \tag{8.49}$$

$$[\partial_{k_x} P_\mathbf{k}, \partial_{k_y} P_\mathbf{k}] = \frac{1}{4}(\partial_{k_x}\hat{d}_\gamma)(\partial_{k_y}\hat{d}_\eta)[\sigma_\gamma, \sigma_\eta]$$

$$= \frac{i}{2}\varepsilon_{\alpha\gamma\eta}(\partial_{k_x}\hat{d}_\gamma)(\partial_{k_y}\hat{d}_\eta)\sigma_\alpha = \frac{i}{2}\partial_{k_x}\hat{\mathbf{d}}(\mathbf{k}) \times \partial_{k_y}\hat{\mathbf{d}}(\mathbf{k}) \cdot \vec{\sigma}. \quad (8.50)$$

The Berry curvature is therefore

$$\boldsymbol{\Omega}_{xy}(\mathbf{k}) = i\,\mathrm{Tr}\,\{P_\mathbf{k}\,[\,\partial_{k_x}P_\mathbf{k}, \partial_{k_y}P_\mathbf{k}\,]\} = -\frac{1}{2}\partial_{k_x}\hat{\mathbf{d}}(\mathbf{k}) \times \partial_{k_y}\hat{\mathbf{d}}(\mathbf{k}) \cdot \mathrm{Tr}\,\{P_\mathbf{k}\vec{\sigma}\}. \quad (8.51)$$

We then insert the $P_\mathbf{k}$ expression, Eq. (8.49), and we notice that $\mathrm{Tr}\,\vec{\sigma} = 0$, $\mathrm{Tr}\,\sigma_\gamma\sigma_\eta = 2\,\delta_{\gamma\eta}$; ergo

$$\boldsymbol{\Omega}_{xy}(\mathbf{k}) = \frac{1}{2}\partial_{k_x}\hat{\mathbf{d}}(\mathbf{k}) \times \partial_{k_y}\hat{\mathbf{d}}(\mathbf{k}) \cdot \hat{\mathbf{d}}(\mathbf{k}). \quad (8.52)$$

$$C_1 = \frac{1}{4\pi}\int_{\mathrm{BZ}} d\mathbf{k}\,\hat{\mathbf{d}}(\mathbf{k}) \cdot \partial_{k_x}\hat{\mathbf{d}}(\mathbf{k}) \times \partial_{k_y}\hat{\mathbf{d}}(\mathbf{k}), \quad (8.53)$$

which indeed maps exactly into the well known formula for a magnetic skyrmion (in $\mathbf{r}$-space, Fig. 8.7).

8.5 Geometry and topology in r-space

We abandon here lattice periodicity and we address instead a bounded sample (possibly noncrystalline), where the single-particle orbitals $|\varphi_j\rangle$ are square-integrable: so-called "open" boundary conditions (OBCs). The mean-field Hamiltonian is written as

$$\mathcal{H}_{\boldsymbol{\kappa}} = \frac{1}{2m}(\mathbf{p} + \hbar\boldsymbol{\kappa})^2 + V(\mathbf{r}), \quad (8.54)$$

where setting $\boldsymbol{\kappa} \neq 0$ amounts to a trivial gauge transformation, easily "gauged away" within OBCs. The $\boldsymbol{\kappa}$-dependent orbitals are in fact $|\varphi_{j\boldsymbol{\kappa}}\rangle = \mathrm{e}^{-i\boldsymbol{\kappa}\cdot\mathbf{r}}|\varphi_j\rangle$, and the ground state projector in Schrödinger representation is

$$\langle\mathbf{r}|\,\mathcal{P}_{\boldsymbol{\kappa}}\,|\mathbf{r}'\rangle = \mathrm{e}^{i\boldsymbol{\kappa}\cdot(\mathbf{r}'-\mathbf{r})}\langle\mathbf{r}|\,\mathcal{P}\,|\mathbf{r}'\rangle, \quad (8.55)$$
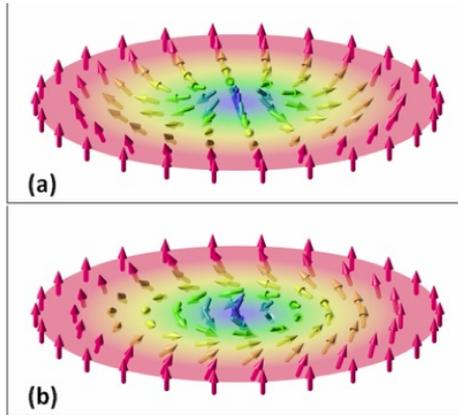

Figure 8.7: Magnetic skyrmion. The topological index (a.k.a. skyrmion number) is the integer n defined as $n = \frac{1}{4\pi}\int dxdy\,\mathbf{M}\cdot\partial_x\mathbf{M} \times \partial_y\mathbf{M}$, where $\mathbf{M}$ is the unit vector along the local magnetization of a $2d$ system. After Wikipedia.

where $\mathcal{P}$ is the zero-$\boldsymbol{\kappa}$ ground-state projector:

$$\mathcal{P} = \sum_{\epsilon_j \leq \mu} |\varphi_j\rangle\langle\varphi_j|. \tag{8.56}$$

Gauge-invariant geometric quantities within OBCs obtain from $\boldsymbol{\kappa}$-derivatives of $\mathcal{P}_{\boldsymbol{\kappa}}$ evaluated at $\boldsymbol{\kappa} = 0$. i.e

$$\langle\mathbf{r}| \partial_{\kappa_\alpha}\mathcal{P}_{\boldsymbol{\kappa}} |\mathbf{r}'\rangle = i(r'_\alpha - r_\alpha)\langle\mathbf{r}| \mathcal{P} |\mathbf{r}'\rangle, \tag{8.57}$$

or in operator notation $\partial_{\kappa_\alpha}\mathcal{P}_{\boldsymbol{\kappa}} = -i[r_\alpha, \mathcal{P}]$. This operator encodes the linear response of the ground state to an infinitesimal gauge potential.

The generic operator product is then

$$(\partial_{\kappa_\alpha}\mathcal{P}_{\boldsymbol{\kappa}})(\partial_{\kappa_\beta}\mathcal{P}_{\boldsymbol{\kappa}}) = -[r_\alpha, \mathcal{P}] [r_\beta, \mathcal{P}], \tag{8.58}$$

and the analogue of the integrated metric-curvature tensor is

$$\begin{aligned}
\mathcal{F}_{\alpha\beta} &= -\frac{1}{V}\mathrm{Tr} \{\mathcal{P}[r_\alpha, \mathcal{P}] [r_\beta, \mathcal{P}]\} \\
&= \frac{1}{V}\mathrm{Tr} \{\mathcal{P}r_\alpha r_\beta \mathcal{P}\} - \frac{1}{V}\mathrm{Tr} \{\mathcal{P}\alpha\mathcal{P}r_\beta\mathcal{P}\},
\end{aligned} \tag{8.59}$$

i.e. the second cumulant moment of the electron distribution (or quantum fluctuation of the dipole). I have divided by the volume (area for $d = 2$) in order to define an intensive quantity. This trace is obviously real symmetric even in absence of T-symmetry; it provides the OBCs analogue of $\Omega_\mathrm{I}/V_\mathrm{cell}$, as discussed e.g. in Ref. [148]. In the large-V limit this quantity converges to a finite limit in insulators, and diverges in metals [149].

We may write the trace of $\mathcal{F}_{\alpha\beta}$ in the Schrödinger representation, i.e.

$$\mathcal{F}_{\alpha\beta} = -\frac{1}{V} \int d\mathbf{r} \, \langle\mathbf{r}| \mathcal{P}[r_\alpha, \mathcal{P}] [r_\beta, \mathcal{P}] |\mathbf{r}\rangle; \tag{8.60}$$

it has been shown in Ref. [215] that the function $\mathrm{Im} \, \langle\mathbf{r}| \mathcal{P}[r_\alpha, \mathcal{P}] [r_\beta, \mathcal{P}] |\mathbf{r}\rangle$ carries the information needed to extract the value of the AHC even from an OBCs calculation. It is enough to evaluate the trace per unit volume by integrating over an inner region of the sample (and *not* over the whole sample). For instance for a bounded crystallite

$$\sigma_{\alpha\beta}^{(-)} = -\frac{2e^2}{\hbar V_\mathrm{cell}} \int_\mathrm{cell} d\mathbf{r} \, \mathrm{Im} \, \langle\mathbf{r}| \mathcal{P}[r_\alpha, \mathcal{P}] [r_\beta, \mathcal{P}] |\mathbf{r}\rangle, \tag{8.61}$$

where the cell is in the center of the crystallite, and the large-crystallite limit is taken. This is demonstrated in Ref. [215] for insulators and in Ref. [199] for metals via

simulations for Haldanium samples. We refer to the very perspicuous figures therein (not reproduced here) for an illustration of the present outstanding message.

We remind that in the $2d$ insulating case $\sigma_{xy}^{(-)}$, when expressed in natural units, coincides with (minus) the Chern number C_1. Therefore, as shown in Ref. [215], our local theory amounts to a space-resolved definition of the Chern number. We have dubbed this "topological marker", and shown its power including in disordered and inhomogenous (heterojuction) cases. We have called its metallic analogue "geometrical marker" and shown that it enjoys similar properties [199].

8.6 Nonlinear Hall conductivity

As seen above, nonzero linear Hall conductivity requires T-symmetry to be absent. An important development, due to Sodemann and Fu, occurred in 2015 [219]: they realized that second-order transverse dc conductivity can be nonzero even in T-symmetric materials, provided that inversion (I) symmetry is absent: the quadratic dc response is then called nonlinear Hall conductivity (NHC); the theory was originally developed at the independent-electron level for crystalline systems, and the relevant expressions are obtained semiclassically [219, 220, 221]. The discovery prompted a large interest into the phenomenon, both experimentally [222] and theoretically; the presentation given here follows Ref. [200].

We start again from Eq. (8.7), and we expand the anomalous velocity as

$$\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(\boldsymbol{\kappa})\dot{\kappa}_\beta \simeq \tilde{\boldsymbol{\Omega}}_{\alpha\beta}(0)\dot{\kappa}_\beta + \partial_{\kappa_\gamma}\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(0)\ \dot{\kappa}_\beta\kappa_\gamma. \tag{8.62}$$

The first term yields the AHC, Eq. (8.13); we focus on the second term in the following:

$$\sigma_{\alpha\beta\gamma}^{(-)}(-\omega,\omega) = \frac{2\,e}{L^d}\partial_{\kappa_\gamma}\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(0)\frac{\partial\dot{\kappa}_\beta}{\partial\,\mathcal{E}_\beta(-\omega)}\frac{\partial\kappa_\gamma}{\partial\,\mathcal{E}_\gamma(\omega)}, \tag{8.63}$$

where the factor of 2 owes to the fact that $(\mathrm{Re}\ \mathcal{E}_\beta \mathrm{e}^{-i\omega t})(\mathrm{Re}\ \mathcal{E}_\gamma \mathrm{e}^{i\omega t}) = \mathcal{E}_\beta\mathcal{E}_\gamma/2$ when averaged over the period. We then set

$$\dot{\kappa}_\beta = -\frac{e}{\hbar}\mathcal{E}_\beta(-\omega), \qquad \kappa_\gamma = \frac{e}{\hbar c}A_\gamma(\omega), \tag{8.64}$$

and we get to lowest order in ω:

$$\begin{aligned} \frac{1}{2}\sigma_{\alpha\beta\gamma}^{(-)}(-\omega,\omega) &\doteq -\frac{e^3}{\hbar^2 c L^d}\partial_{\kappa_\gamma}\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(0)\frac{dA(\omega)}{d\mathcal{E}(\omega)} \\ &\doteq \frac{e^3}{\hbar^2 L^d}\partial_{\kappa_\gamma}\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(0)\ \frac{i}{\omega + i\eta}. \end{aligned} \tag{8.65}$$

The dc current induced to second order by the constant field is thus:

$$j_\alpha = \frac{i}{\omega + i\eta}\chi_{\alpha\beta\gamma}\mathcal{E}_\beta\mathcal{E}_\gamma, \quad \chi_{\alpha\beta\gamma} = \frac{e^3}{\hbar^2 L^d}\partial_{\kappa_\gamma}\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(0). \tag{8.66}$$

The real part of the ω-dependent factor in Eq. (8.66) equals $\pi\delta(\omega)$: the many-electron system undergoes a transverse free acceleration. A dc current obtains upon replacement of the infinitesimal η with an inverse relaxation time $1/\tau$. This is in stark contrast with AHC, Eq. (8.13), accounting for a τ-independent dc current.

As for the symmetry properties of Eq. (8.66), we remind that in presence of T-symmetry $\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(\boldsymbol{\kappa}) = -\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(-\boldsymbol{\kappa})$, while in presence of I-symmetry $\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(\boldsymbol{\kappa}) = \tilde{\boldsymbol{\Omega}}_{\alpha\beta}(-\boldsymbol{\kappa})$ [13]: therefore in a T-symmetric system $\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(0) = 0$ and the AHC vanishes. In the case of NHC the parity is swapped: the gradient of $\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(\boldsymbol{\kappa})$ is even in T-symmetric systems, and odd in I-symmetric systems. Therefore the NHC requires breaking of I-symmetry; in the special case of a T-symmetric and I-breaking system, nonzero transverse conductivity appears at second order only.

Since the responses to $\mathcal{E}_\beta\mathcal{E}_\gamma$ and to $\mathcal{E}_\gamma\mathcal{E}_\beta$ coincide, $\chi_{\alpha\beta\gamma}$ is symmetrical in the β, γ indices, while instead it is antisymmetrical in the α, β and α, γ indices. Therefore the current is always orthogonal to the field: if—for an arbitrary $\mathcal{E}$ orientation—we set the x-axis along $\mathcal{E}$, then $j_x \propto \chi_{xxx} = 0$, while $j_y \propto \chi_{yxx}$ and $j_z \propto \chi_{zxx}$ are not constrained to be zero by (this) symmetry.

At the independent-electron level we exploit Eq. (8.22) to obtain

$$\frac{1}{L^d}\partial_{\kappa_\alpha}\tilde{\boldsymbol{\Omega}}_{\alpha\beta}(0) = \int_{\text{BZ}}\frac{d\mathbf{k}}{(2\pi)^d}\partial_{k_\alpha}\boldsymbol{\Omega}_{\alpha\beta}(\mathbf{k}), \tag{8.67}$$

$$\chi_{\alpha\beta\gamma} = \frac{e^3}{\hbar^2}\int_{\text{BZ}}\frac{d\mathbf{k}}{(2\pi)^d}\,\partial_{k_\gamma}\boldsymbol{\Omega}_{\alpha\beta}(\mathbf{k}). \tag{8.68}$$

This is equivalent—in the single-band case—to the semiclassical expression which first appeared in the founding NHC 2015 paper by Sodemann and Fu [219].

Chapter 9

Magnetization

Like polarization $\mathbf{P}$, even macroscopic magnetization $\mathbf{M}$ is a fundamental concept that all undergraduates learn about in elementary courses [95, 96]. Macroscopic magnetization $\mathbf{M}$ is the sum of two terms, which are unambiguously defined in nonrelativistic (and semirelativistic) quantum mechanics: spin magnetization $\mathbf{M}^{(\mathrm{spin})}$ and orbital magnetization $\mathbf{M}^{(\mathrm{orb})}$. Experimentally, magneto-mechanical measurements, based on the Einstein-de Haas effect, provide the two terms separately. For instance, the values of $\mathbf{M}^{(\mathrm{spin})}$ and $\mathbf{M}^{(\mathrm{orb})}$ for the three ferromagnetic metals (Fe, Co, and Ni) are accurately known since half a century [223]. From the viewpoint of the present review, $\mathbf{M}^{(\mathrm{spin})}$ is a dull quantity. Electronic-structure codes routinely compute the spin density, which is a simple lattice-periodical function. Its cell average (times a trivial factor) coincides with $\mathbf{M}^{(\mathrm{spin})}$. In other words, a "dipolar density" is unambiguously identified, while the same does not happen in the orbital case. Throughout this Chapter we address orbital magnetization, using the symbol $\mathbf{M}$ to identify $\mathbf{M}^{(\mathrm{orb})}$ only.

Polarization $\mathbf{P}$ and orbital magnetization $\mathbf{M}$ share some analogy, but also some significant differences. For the time being, we emphasize the main analogy: textbooks define both $\mathbf{P}$ and $\mathbf{M}$ as the dipole moment of a sample, divided by the volume V:

$$\mathbf{P} = \frac{\mathbf{d}}{V} = \frac{1}{V} \int d\mathbf{r} \, \mathbf{r} \rho^{(\mathrm{micro})}(\mathbf{r}), \quad \mathbf{M} = \frac{\mathbf{m}}{V} = \frac{1}{2cV} \int d\mathbf{r} \, \mathbf{r} \times \mathbf{j}^{(\mathrm{micro})}(\mathbf{r}), \qquad (9.1)$$

where $\rho^{(\mathrm{micro})}(\mathbf{r})$ and $\mathbf{j}^{(\mathrm{micro})}(\mathbf{r})$ are the microscopic charge and (orbital) current densities. Contrary to what most textbooks pretend, there is no such thing as a "dipolar density": the basic microscopic quantities are $\rho^{(\mathrm{micro})}(\mathbf{r})$ and $\mathbf{j}^{(\mathrm{micro})}(\mathbf{r})$ [224]. If the sample is *uniformly* polarized/magnetized, then the microscopic charge/current averages to zero in the bulk of the sample. Owing to the presence of the unbound operator $\mathbf{r}$, both $\mathbf{P}$ and $\mathbf{M}$ Eq. (9.1) are dominated by surface contributions, while instead phenomenologically they are bulk properties.

The modern theory of magnetization dates from 2005-06 [217, 225], and is still (2021) partly work on progress [225, 226, 227, 211, 228, 229, 218]. A few reviews have appeared [8, 18, 22], while only a few first-principle implementations appeared so far [230, 231, 232]. The theory is strictly at the independent-electron level; a formula providing $\mathbf{M}$ in terms of an explicitly correlated ground state has not yet (as of 2021) been proposed.

9.1 Magnetization and magnetic field

The modern theory of magnetization only addresses the polarization $\mathbf{M}$ in a null macroscopic $\mathbf{B}$ field; in this case $\mathbf{M}$ can be nonzero only if T-symmetry is broken in the spatial wavefunction. For instance, in a ferromagnet the spin-orbit interaction transmits the symmetry breaking from the spin degrees of freedom to the spatial (orbital) ones.

It must be realized that, insofar as we address an infinite system with no boundaries, the $\mathbf{B}$ field is quite arbitrary, in full analogy with the case of $\mathbf{E}$ discussed in Sec. 5.1. The usual choice invariably performed within all electronic-structure codes is to impose a lattice-periodical vector potential, i.e. $\mathbf{B} = 0$. Notice that, instead, spontaneous magnetization is customarily defined at $\mathbf{H} = 0$ (see Sect. A.2)

When addressing a finite sample with boundaries, the $\mathbf{B}$ field is in principle measurable inside the material, without reference to what happens at the sample boundary: in fact, $\mathbf{B}$ obtains by averaging over a macroscopic length scale the microscopic magnetic field $\mathbf{B}^{(\mathrm{micro})}(\mathbf{r})$, which fluctuates at the atomic scale [96]. In a macroscopically homogeneous system the macroscopic field $\mathbf{B}$ is constant, and in crystalline materials it coincides with the cell average of $\mathbf{B}^{(\mathrm{micro})}(\mathbf{r})$.

As for the electrical case, discussed in Sec. 5.1, there is no need of addressing finite samples and external vs. internal fields from a theoretician's viewpoint. Nonetheless a brief digression is in order, given that experiments *are* performed over finite samples, often in external fields. Suppose a finite macroscopic sample is inserted in a constant external field $\mathbf{B}^{(\mathrm{ext})}$: the microscopic field $\mathbf{B}^{(\mathrm{micro})}(\mathbf{r})$ coincides with $\mathbf{B}^{(\mathrm{ext})}$ far away from the sample, while it is different inside because of screening effects. If we choose an homogeneous sample of *ellipsoidal shape*, then the macroscopic average of $\mathbf{B}^{(\mathrm{micro})}(\mathbf{r})$, i.e. the macroscopic screened field $\mathbf{B}$, is constant in the bulk of the sample. The shape effects are embedded in the demagnetization coefficients [95]: the simplest case is the extremely oblate ellipsoid, i.e. a slab of a macroscopically homogeneous material; more details are given in Ref. [8]. For the slab geometry in a vanishing external field $\mathbf{B}^{(\mathrm{ext})}$ the internal field $\mathbf{B}$ vanishes when $\mathbf{M}$ is normal to the slab (longitudinal polarization), while $\mathbf{B} = -4\pi\mathbf{M}$ is the demagnetization field when $\mathbf{M}$ is parallel to the slab (transverse polarization): see Fig. 9.1. Notice that this is *the opposite* of what happens in the electrical case (Fig.
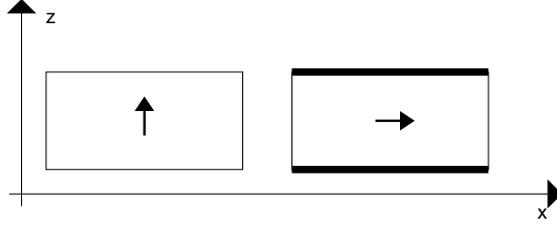
Figure 9.1: Macroscopic magnetization $\mathbf{M}$ in a slab normal to z, for a vanishing external field $\mathbf{B}^{(\text{ext})}$. Left: When $\mathbf{M}$ is normal to the slab, no demagnetizing field and no surface current is present. Right: When $\mathbf{M}$ is parallel to the slab, a demagnetizing field $\mathbf{B} = 4\pi\mathbf{M}$ is present inside the slab, and dissipationless currents $\mathbf{K}_{\text{surface}} = c\,\mathbf{M} \times \mathbf{n}$ flow at the surfaces.

5.1).

9.2 Orbital magnetization

In Eq. (9.1) we have emphasized the main analogy between $\mathbf{P}$ and $\mathbf{M}$; here, instead, we anticipate the main difference. In the case of $\mathbf{P}$—as stressed in Sec. 5.2—the modern theory addresses the difference in polarization $\Delta\mathbf{P}$ between two states of the material that can be connected by an adiabatic switching process. As for $\mathbf{P}$ "itself", the modern theory of polarization states that the bulk electron distribution determines $\mathbf{P}$ only *modulo a "quantum"*, whose value depends on the boundary of the sample [101, 117] (Sec. 5.3.2, and also Figs. 5.4 and 5.5). The modern theory of magnetization, instead, addresses $\mathbf{M}$ "itself" directly, and is not affected by any quantum indeterminacy. We may express this outstanding difference by saying that by tinkering with the boundary one can change the polarization of a sample (by a "quantum"); while instead tinkering with the boundary one cannot change the value of $\mathbf{M}$.

There is a profound reason for the striking difference between $\mathbf{M}$ and $\mathbf{P}$ from the viewpoint of quantum geometry. While $\mathbf{P}$ is basically a $1d$ phenomenon (it is defined even for a polymer), $\mathbf{M}$ in its simplest manifestation is a $2d$ phenomenon: orbital currents in a plane and $\mathbf{M}$ normal to such plane. The qualitative difference can be traced back to the discussion in Sect. 6.2.2, about odd- and even-dimension differential forms.

Another key difference is that, while $\mathbf{P}$ makes sense only for insulators at zero temperature, $\mathbf{M}$ is well defined even for metals and even for finite temperature. In the present Notes we limit ourselves to discuss the zero-T case, for normal insulators, Chern insulators, and metals. We refer to the original papers [225, 226], as well as to the review of Ref. [8] for the finite-T case.

157

9.3 Orbital magnetization of a bounded sample

For a bounded sample with square-integrable orbitals we may safely adopt the definition of Eq. (9.1):

$$\mathbf{M} = -\frac{1}{2cV} \int d\mathbf{r}\, \mathbf{r} \times \mathbf{j}^{(\text{micro})}(\mathbf{r}) = -\frac{e}{2cV} \sum_{\epsilon_j < \mu} \langle \varphi_j | \mathbf{r} \times \mathbf{v} | \varphi_j \rangle, \tag{9.2}$$

which applies to both the insulating and metallic cases. In Eq. (9.2) $|\varphi_j\rangle$ are the spinless orbitals, ϵ_j their energies, and $\mathbf{v} = i[\mathcal{H}, \mathbf{r}]/\hbar$; the formula is given per spin channel (or for "spinless electrons"). It is easily recast in terms of the ground-state projector, Eq. (7.84):

$$\mathcal{P} = \sum_{\epsilon_j < \mu} |\varphi_j\rangle\langle\varphi_j|, ; \tag{9.3}$$

$$\mathbf{M} = \frac{e}{2cV} \text{Tr}\, \{\mathcal{P}\mathbf{r} \times \mathbf{v}\}. \tag{9.4}$$

It is expedient to write the velocity operator in the equivalent form $\mathbf{v} = i[(\mathcal{H} - \mu), \mathbf{r}]/\hbar$, whence

$$\mathbf{M} = -\frac{ie}{2\hbar cV} \text{Tr}\, \{\mathcal{P}\mathbf{r} \times (\mathcal{H} - \mu)\mathbf{r}\}, \tag{9.5}$$

since $\mathbf{r} \times \mathbf{r} = 0$. Insofar as the system remains finite, the appearance of μ in Eq. (9.5) looks irrelevant. But our choice has the virtue that $\mathbf{M}$ is *explicitly* a function of $\mathcal{H} - \mu$: this is essential in the thermodynamic limit for metals, where the spectrum becomes dense around μ.

So far, we have identically transformed the textbook definition of Eq. (9.1) into the equivalent Eq. (9.5): this is still plagued by the same original drawback: the $\mathbf{r}$-space integral includes an extensive contribution from the boundary. Next we are going to tame the position operator by performing a transformation similar in spirit to an integration by parts, where the same integrated value obtains from two very different *integrands*. Eventually, the trace will be boundary-insensitive in the large-V limit.

Using the cyclic properties of the trace, the Cartesian components of $\mathbf{M}$ are

$$M_\gamma = \frac{ie}{2\hbar cV} \varepsilon_{\gamma\alpha\beta} \text{Tr}\, \{(\mathcal{H} - \mu)r_\alpha \mathcal{P} r_\beta\}. \tag{9.6}$$

where $\varepsilon_{\gamma\alpha\beta}$ is the antisymmetric tensor and the sum over repeated indices is implicit (here and throughout). The following lemma is then very useful. Let O be any Hermitian operator commuting with $\mathcal{P}$. Then

$$\text{Tr}\, \{O\,[r_\alpha, \mathcal{P}]\,[r_\beta, \mathcal{P}]\} = -\text{Tr}\, \{O\mathcal{P}\, r_\alpha r_\beta\} - \text{Tr}\, \{O(2\mathcal{P} - I)r_\alpha \mathcal{P} r_\beta\}. \tag{9.7}$$
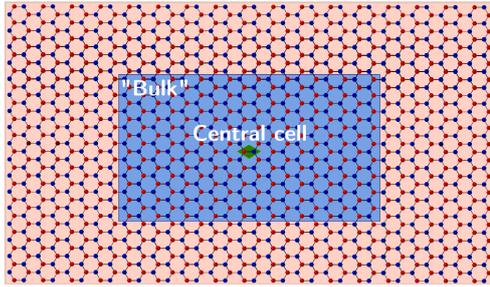
Figure 9.2: A bounded flake of "Haldanium": the one shown here has 1806 sites. In the text we consider three cases: the whole flake; the "bulk" region (1/4 of the sites) and the central cell (two sites). After Ref. [150].

We apply this lemma by identifying O with the operator $|\mathcal{H} - \mu| = (\mathcal{H} - \mu)(I - 2\mathcal{P})$ to get

$$M_\gamma = -\frac{ie}{2\hbar cV}\varepsilon_{\gamma\alpha\beta}\mathrm{Tr}\,\{|\mathcal{H} - \mu|\,[r_\alpha, \mathcal{P}]\,[r_\beta, \mathcal{P}]\}, \tag{9.8}$$

where we have used $(I - 2\mathcal{P})^2 = I$.

If we now define the vector field

$$\mathfrak{M}(\mathbf{r}) = -\frac{ie}{2\hbar c}\langle\mathbf{r}|\,|\mathcal{H} - \mu|\,[\mathbf{r}, \mathcal{P}] \times [\mathbf{r}, \mathcal{P}]\,|\mathbf{r}\rangle, \tag{9.9}$$

the previous main result can be cast in the very compact form

$$\mathbf{M} = \frac{1}{V}\int d\mathbf{r}\,\mathfrak{M}(\mathbf{r}), \tag{9.10}$$

where we remind that the formula as it stands applies to spinless electrons. As proved here, the integrated value in Eq. (9.10) is identical to Eqs. (9.1) and (9.5) at any sample size; the key difference is that the *integrands* are quite different: Eqs. (9.1) and (9.5) are dominated by boundary contributions, while Eq. (9.10) is free from such drawback. Indeed, Eq. (9.10) is a *local* definition of magnetization: if the bounded sample is a crystallite (cut from a bulk crystal), then a definition of **M** equivalent to Eq. (9.10) in the large-sample limit is

$$\mathbf{M} = \frac{1}{V_{\mathrm{cell}}}\int_{V_{\mathrm{cell}}} d\mathbf{r}\,\mathfrak{M}(\mathbf{r}), \tag{9.11}$$

where V_{cell} is a crystal cell in the center region of the crystallite. Obviously, a similar property *does not* hold for the original integral, Eq. (9.1). Owing to the locality property, the definition of Eq. (9.11) can also deal—with obvious modifications—with noncrystalline or inhomogeneous samples, like e.g. heterojunctions.

Our definition of spontaneous magnetization is $\mathbf{M} = -\partial\mathscr{F}/\partial\mathbf{B}$, where $\mathscr{F}$ is the free-energy density; the reason why we adopt **B** (and not **H**) as independent variable is explained in Appendix A.2 (the zero-subscript in $\mathbf{M}_0$ is dropped here). It is therefore clear that the macroscopic average of $\mathfrak{M}(\mathbf{r})$ coincides with $-\partial\mathscr{F}/\partial\mathbf{B}$.
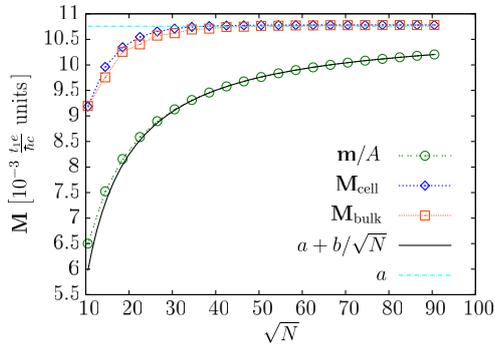
Figure 9.3: Green circles: orbital magnetization, evaluated from the textbook expression $\mathbf{m}/A$, as a function of the flake size, in the insulating case. The plots labeled as $\mathbf{M}_{\text{cell}}$ and $\mathbf{M}_{\text{bulk}}$ demonstrate the locality of orbital magnetization, as explained in the text. The horizontal dashed line is provided by the $\mathbf{k}$-space formula. After Ref. [233].

In all of the previous equations the large-sample limit is understood; it has been demonstrated [218, 233, 234] that the convergence of Eq. (9.11) is much faster than the convergence of Eqs. (9.1), (9.5), and (9.10). Two-dimensional simulations on bounded Haldanium flakes have proved that the convergence of Eqs. (9.1), (9.5), and (9.10) is of the order $1/L$, where L is the linear dimension of the sample, in both insulators and metals. Instead the convergence of Eq. (9.11) is exponential in insulators, and $L^{-\alpha}$, with $\alpha > 1$, in metals. The difference between the two owes to the different "nearsightedness" of $\mathcal{P}$ [235, 165].

A typical Haldanium flake is shown in Fig. 9.2; it is insulating at half-filling and metallic at any other filling. The magnetization as a function of the flake size is shown in Fig. 9.3 for a typical insulating case in the trivial (nontopological) regime. Diamonds and squares where obtained from integrating $\mathfrak{M}(\mathbf{r})$ over the central cell (two sites) and over the "bulk" region (1/4 of the sites), respectively, as displayed in Fig. 9.2. The conventional formula, Eqs. (9.1) and (9.5), converges like the inverse linear dimension of the flake (in both the insulating and the metallic case) to the asymptotic value in the large-flake limit. Other cases and other figures can be found in the original literature: Refs. [218, 233, 234] (not reproduced here).

9.3.1 Orbital magnetization of an unbounded crystalline sample

The definition of the vector field $\mathfrak{M}(\mathbf{r})$, Eq. (9.9), applies as it stands even to an unbounded sample within PBCs. In the crystalline case the $\mathcal{P}$ projector therein is expressed in terms of Bloch orbitals:

$$\mathcal{P} = \frac{V_{\text{cell}}}{(2\pi)^3} \sum_j \int_{\text{BZ}} d\mathbf{k} \; \theta(\mu - \epsilon_{j\mathbf{k}}) \, |\psi_{j\mathbf{k}}\rangle\langle\psi_{j\mathbf{k}}|. \tag{9.12}$$

160

It is easy to verify that the operator $[\mathbf{r}, \mathcal{P}]$ commutes with the lattice translations, and therefore $\mathfrak{M}(\mathbf{r})$ is a lattice-periodical function. In this sense the commutator with $\mathcal{P}$ effectively "tames" the nasty, unbounded, position operator $\mathbf{r}$. The expression for $\mathbf{M}$ as a trace per unit volume is therefore formally identical to Eq. (9.11), found above for a *bounded* sample. We show next how to express it as a BZ integral in terms of Bloch orbitals; in metals the BZ integral is actually a Fermi volume integral, owing to the θ function entering the $\mathcal{P}$ definition, Eq. (9.12).

We start rewriting Eq. (9.12) as

$$\langle \mathbf{r} | \mathcal{P} | \mathbf{r}' \rangle = \frac{V_{\text{cell}}}{(2\pi)^3} \int_{\text{BZ}} d\mathbf{k} \, e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')} \langle \mathbf{r} | \mathcal{P}_{\mathbf{k}} | \mathbf{r}' \rangle \tag{9.13}$$

$$\mathcal{P}_{\mathbf{k}} = \sum_j \theta(\mu - \epsilon_{j\mathbf{k}}) |u_{j\mathbf{k}}\rangle \langle u_{j\mathbf{k}}|. \tag{9.14}$$

The $\mathcal{P}_{\mathbf{k}}$ projector is gauge-invariant in the generalized sense, i.e. it is invariant for any unitary transformation of the occupied $|u_{j\mathbf{k}}\rangle$ at the given $\mathbf{k}$; it is also periodic in $\mathbf{k}$ (even when the Chern invariant is nonzero). Therefore taking the $\mathbf{k}$-gradient inside the integral we get

$$0 = i(\mathbf{r} - \mathbf{r}')\langle \mathbf{r} | \mathcal{P} | \mathbf{r}' \rangle + \frac{V_{\text{cell}}}{(2\pi)^3} \int_{\text{BZ}} d\mathbf{k} \, e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')} \langle \mathbf{r} | \nabla_{\mathbf{k}} \mathcal{P}_{\mathbf{k}} | \mathbf{r}' \rangle; \tag{9.15}$$

$$[\mathbf{r}, \mathcal{P}] = \frac{iV_{\text{cell}}}{(2\pi)^3} \int_{\text{BZ}} d\mathbf{k} \, e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')} \nabla_{\mathbf{k}} \mathcal{P}_{\mathbf{k}}. \tag{9.16}$$

The Hamiltonian $\mathcal{H}$ can also be written as a BZ integral, hence

$$|\mathcal{H} - \mu| = \frac{V_{\text{cell}}}{(2\pi)^3} \int_{\text{BZ}} d\mathbf{k} \, e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')} |\mathcal{H}_{\mathbf{k}} - \mu|, \tag{9.17}$$

where $\mathcal{H}_{\mathbf{k}} = e^{-i\mathbf{k}\cdot\mathbf{r}} \mathcal{H} e^{i\mathbf{k}\cdot\mathbf{r}}$.

We thus get the relevant operator identity:

$$|\mathcal{H} - \mu| [r_\alpha, \mathcal{P}] [r_\beta, \mathcal{P}] = -\frac{V_{\text{cell}}}{(2\pi)^3} \int_{\text{BZ}} d\mathbf{k} \, e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')} |\mathcal{H}_{\mathbf{k}} - \mu| (\partial_\alpha \mathcal{P}_{\mathbf{k}})(\partial_\beta \mathcal{P}_k). \tag{9.18}$$

Notice that the three BZ integrals entering the product on the l.h.s.—from Eqs. (9.16) and (9.17)—eventually contract to a single BZ integral. This owes to the fact that both $\langle \mathbf{r} | \mathcal{H}_{\mathbf{k}} | \mathbf{r}' \rangle$ and $\langle \mathbf{r} | \nabla \mathcal{P}_{\mathbf{k}} | \mathbf{r}' \rangle$ are lattice-periodical in $\mathbf{r}$ and $\mathbf{r}'$ *separately*.

The ultimate formula for the orbital magnetization of a crystalline system is then, after Eqs. (9.9) and (9.11)

$$\mathbf{M} = \frac{ie}{2\hbar c} \int_{\text{BZ}} \frac{d\mathbf{k}}{(2\pi)^3} \text{Tr} \left\{ |\mathcal{H}_{\mathbf{k}} - \mu| (\nabla_{\mathbf{k}} \mathcal{P}_{\mathbf{k}}) \times (\nabla_{\mathbf{k}} \mathcal{P}_{\mathbf{k}}) \right\}. \tag{9.19}$$

9.3.2 Insulators and metals

Our main magnetization formula, Eq. (9.19) holds for both insulators and metals: the latter case requires further clarification. We start from the $\mathcal{P}_\mathbf{k}$ definition, Eq. (9.14), and we further adopt a gauge where $|u_{j\mathbf{k}}\rangle$ is a differentiable function of $\mathbf{k}$ (this is always possible, even for topologically nontrivial solids): then

$$
\begin{aligned}
\nabla_\mathbf{k}\mathcal{P}_\mathbf{k} &= \sum_j \theta(\mu - \epsilon_{j\mathbf{k}})(\, |\nabla_\mathbf{k}u_{j\mathbf{k}}\rangle\langle u_{j\mathbf{k}}| + |\nabla_\mathbf{k}u_{j\mathbf{k}}\rangle\langle u_{j\mathbf{k}}| \,) \\
&- \sum_j \delta(\mu - \epsilon_{j\mathbf{k}})\nabla_\mathbf{k}\epsilon_{j\mathbf{k}}\, |u_{j\mathbf{k}}\rangle\langle u_{j\mathbf{k}}|.
\end{aligned}
\tag{9.20}
$$

The second line vanishes in insulators, but is singular in metals. Nonetheless the singularity does not affect Eq. (9.19), given that it cancels in the antisymmetrized product. We may therefore safely neglect the second line of Eq. (9.20) in the following. The integrand in Eq. (9.19) is a continuous function of $\mathbf{k}$ in insulators, and piecewise continuous in metals; Eq. (9.19) is indeed a well defined Fermi-volume integral in both cases. As for the first line of Eq. (9.20), we notice that, while $\nabla_\mathbf{k}\mathcal{P}_\mathbf{k}$ is gauge-invariant and Hermitian, the operators $|\nabla_\mathbf{k}u_{j\mathbf{k}}\rangle\langle u_{j\mathbf{k}}|$ and $|u_{j\mathbf{k}}\rangle\langle\nabla_\mathbf{k}u_{j\mathbf{k}}|$ are in general gauge-dependent and nonHermitian: in fact the trace of $|\nabla_\mathbf{k}u_{j\mathbf{k}}\rangle\langle u_{j\mathbf{k}}|$, times i, is nothing else than the Berry connection of band j, i.e. the essential ingredient of polarization theory.

At this point we wish to make contact with the original formula of magnetization theory [80, 216, 217], as reported in the previous reviews [8, 18], and implemented in a few first-principles calculations:

$$
M_\gamma = -\frac{ie}{2\hbar c}\varepsilon_{\gamma\alpha\beta} \sum_{\varepsilon_{j\mathbf{k}}<\mu} \int_{\text{BZ}} \frac{d\mathbf{k}}{(2\pi)^3} \langle\partial_\alpha u_{j\mathbf{k}}| \left(\mathcal{H}_\mathbf{k} + \epsilon_{j\mathbf{k}} - 2\mu\right) |\partial_\beta u_{j\mathbf{k}}\rangle.
\tag{9.21}
$$

It is important to observe that Eq. (9.21) requires the so-called "Hamiltonian gauge", i.e. the $|u_{j\mathbf{k}}\rangle$ are eigenstates of $\mathcal{H}_\mathbf{k}$: unitary mixing of them—à la Marzari-Vanderbilt [56, 19]—is *not* permitted. Our more general Eq. (9.19), instead, is fully gauge invariant.

In order to prove the equivalence, it is expedient to introduce a lemma. Let $O_\mathbf{k}$, with eigenvalues $o_{j\mathbf{k}}$, be any operator which commutes with $\mathcal{H}_\mathbf{k}$, i.e.

$$
O_\mathbf{k} = \sum_j |u_{j\mathbf{k}}\rangle o_{j\mathbf{k}}\langle u_{j\mathbf{k}}|.
\tag{9.22}
$$

Then the following identity holds:

$$
\begin{aligned}
\text{Tr}\,\{O_\mathbf{k}(\partial_\alpha\mathcal{P}_\mathbf{k})(\partial_\beta\mathcal{P}_\mathbf{k})\} &= \sum_j \theta(\mu - \epsilon_{j\mathbf{k}})o_{j\mathbf{k}}\langle\partial_\alpha u_{j\mathbf{k}}|\partial_\beta u_{j\mathbf{k}}\rangle \\
&+ \sum_j \theta(\mu - \epsilon_{j\mathbf{k}})\langle\partial_\beta u_{j\mathbf{k}}| O_\mathbf{k}(I - 2\mathcal{P}_\mathbf{k}) |\partial_\alpha u_{j\mathbf{k}}\rangle.
\end{aligned}
\tag{9.23}
$$

This identity is somewhat reminiscent of Eq. (9.7) and is proved via a straightforward although somewhat tedious calculation.

We now identify $O_{\mathbf{k}}$ with $|\mathcal{H}_{\mathbf{k}} - \mu| = (\mathcal{H}_{\mathbf{k}} - \mu)(I - 2\mathcal{P}_{\mathbf{k}})$ and we get

$$
\begin{aligned}
\mathrm{Tr}\left\{|\mathcal{H}_{\mathbf{k}} - \mu|(\partial_\alpha \mathcal{P}_{\mathbf{k}})(\partial_\beta \mathcal{P}_{\mathbf{k}})\right\} &= \sum_j \theta(\mu - \epsilon_{j\mathbf{k}})(\mu - \epsilon_{j\mathbf{k}})\langle \partial_\alpha u_{j\mathbf{k}} | \partial_\beta u_{j\mathbf{k}}\rangle \\
&+ \sum_j \theta(\mu - \epsilon_{j\mathbf{k}})\langle \partial_\beta u_{j\mathbf{k}} | (\mathcal{H}_{\mathbf{k}} - \mu) | \partial_\alpha u_{j\mathbf{k}}\rangle,
\end{aligned}
\tag{9.24}
$$

where we have exploited $(I - 2\mathcal{P}_{\mathbf{k}})^2 = I$. Antisymmetrization yields

$$
\begin{aligned}
&\varepsilon_{\gamma\alpha\beta}\mathrm{Tr}\left\{|\mathcal{H}_{\mathbf{k}} - \mu|(\partial_\alpha \mathcal{P}_{\mathbf{k}})(\partial_\beta \mathcal{P}_{\mathbf{k}})\right\} \\
&= \varepsilon_{\gamma\alpha\beta}\sum_j \theta(\mu - \epsilon_{j\mathbf{k}})\langle \partial_\alpha u_{j\mathbf{k}} | (2\mu - \epsilon_{j\mathbf{k}} - \mathcal{H}_{\mathbf{k}}) | \partial_\beta u_{j\mathbf{k}}\rangle,
\end{aligned}
\tag{9.25}
$$

and substitution into Eq. (9.19) concludes our proof.

Appendix A

Magnetoelectrics (basic features)

A.1 Generalities

Let us consider first a "normal", i.e. non magnetoelectric, material; we assume it to be in general anisotropic, but macroscopically homogeneous. The electromagnetic free-energy density of the continuous medium is

$$\mathscr{F}(\mathbf{E}, \mathbf{H}) = -\frac{1}{8\pi}\mathbf{E}\overset{\leftrightarrow}{\varepsilon}\mathbf{E} - \frac{1}{8\pi}\mathbf{H}\overset{\leftrightarrow}{\mu}\mathbf{H}, \qquad (A.1)$$

where $\overset{\leftrightarrow}{\varepsilon}$ is the macroscopic dielectric tensor (same as in Appendix B), and $\overset{\leftrightarrow}{\mu}$ is the magnetic permeability tensor. We draw attention to the *minus* signs, explained e.g. in Ref [95]: $\mathscr{F}$ is a free energy, *not* an energy.

Eq. (A.1) is nicely symmetric and—since we are adopting Gaussian units—the fields have the same dimensions. This however hides an important feature: electric and magnetic energies in condensed matter *are not* of the same order of magnitude. The ratio of the magnetic energy scale to the electric one is of the order of 10^{-4}. This ratio owes in fact to the actual value of the squared fine-structure constant: $(1/137)^2$.

As we will see in more detail below, magnetoelectric materials realize a coupling between electric and magnetic phenomena. The conversion of magnetic signals into electrical ones (and conversely) is obviously of the utmost technological interest. Nowadays, most devices exploit the phenomenon of the giant magnetoresistance (Nobel prize in 2007), which has generated an economy of billion of dollars. To date, the magnetoelectric effect is *not* commercially viable, due to the fact that in all known materials the coupling is very weak, owing again to the $(1/137)^2$ bottleneck. An active area of research concerns therefore mechanisms and materials where the effect would be enhanced.

A.2 B vs. H fields

In this Appendix we focus on the free energy $\mathscr{F}(\mathbf{E}, \mathbf{H})$, where $\mathbf{E}$ and $\mathbf{H}$ are chosen as the independent variables. As for the electric field, this is clearly the natural choice. We have stressed that $\mathbf{E}$ (not $\mathbf{D}$) is the *internal* field, which is measurable in principle inside the material. The macroscopic field $\mathbf{E}$ is also a control parameter in first-principle calculations: standard crystalline Hamiltonians adopt a lattice-periodical selfconsistent potential, hence $\mathbf{E} = 0$. Matters are different in the magnetic case: the *internal* field, measurable in principle inside the material is $\mathbf{B}$, not $\mathbf{H}$. Also, if a selfconsistent Hamiltonian is lattice-periodical, the macroscopic $\mathbf{B}$ field, not $\mathbf{H}$, is zero. And in fact the modern theory of magnetization, discussed here in Ch. 9, addresses magnetization in zero $\mathbf{B}$ field.

So, why instead we are adopting here $\mathbf{H}$, and not $\mathbf{B}$, as the independent variable? There are several reasons. Phenomenologically, the experimenter directly controls $\mathbf{E}$ (e.g. via capacitors) and $\mathbf{H}$ (via e.g. solenoids or generally currents). In the laboratory, you will hear people speaking of $\mathbf{E}$ and $\mathbf{H}$, more often than of $\mathbf{D}$ and $\mathbf{B}$, the reason being that $\mathbf{E}$ and $\mathbf{H}$ are directly read on the instruments. Several textbooks even call $\mathbf{H}$ the "magnetic field", which I find strongly misleading. I adopt the nomenclature of the good textbooks, such as Feynman [236] and Griffiths [237] popular textbooks, where $\mathbf{B}$ is called "magnetic field", and $\mathbf{H}$ is just "$\mathbf{H}$".

In the framework of magnetoelectric effects there are further reasons for using $\mathbf{H}$ as the independent variable. All formulæ–including Eq. (A.1)—have a pretty symmetric expression in terms of $\mathbf{E}$ and $\mathbf{H}$; Feynman warns however that "although the equations are analogous, the physics is not analogous" [236]. The equations would look asymmetrical and ugly using the genuine magnetic field $\mathbf{B}$ instead.

Custom dictates that spontaneous magnetization is defined as $\mathbf{M}_0 = -\partial \mathscr{F}/\partial \mathbf{H}$: while instead in Ch. 9 we addressed $\mathbf{M}_0 = -\partial \mathscr{F}/\partial \mathbf{B}$. However $\mathbf{B}$ and $\mathbf{H}$ are related by the magnetic permeability tensor as $\mathbf{B} = \overset{\leftrightarrow}{\mu}\, \mathbf{H}$, and in normal materials the permeability difference from one (in Gaussian units) is of the order of $(1/137)^2$. Caveat: matters are quite different in a superconducting material [154]. There is then little difference between $\mathbf{M}_0$ as addressed in Ch. 9 and $\mathbf{M}_0$ as customarily addressed. Whenever needed, the conversion of the response tensors from their $(\mathbf{E}, \mathbf{H})$ definition to the $(\mathbf{E}, \mathbf{B})$ counterpart requires only straightforward algebra. The $(\mathbf{E}, \mathbf{B})$ must be adopted in first-principle calculations, as indeed in Ch. 9; it is also adopted in Sec. A.6 below.

A.3 Multiferroics

The first order expansion of Eq. (A.1) reads

$$\mathscr{F}(\mathbf{E}, \mathbf{H}) = \mathscr{F}_0 - \mathbf{P}_0 \cdot \mathbf{E} - \mathbf{M}_0 \cdot \mathbf{H} + \dots , \tag{A.2}$$

where $\mathbf{P}_0$ is the spontaneous polarization and $\mathbf{M}_0$ is the spontaneous magnetization. The materials where both are nonzero are called multiferroics: ferroelectric and ferromagnetic at the same time. In such materials the electric-magnetic coupling is expected to be strong.

There are very few multiferroic materials: ferroelectricity and ferromagnetism seem to be almost mutually exclusive in nature. The reasons for this have been investigated [238].

A.4 Linear magnetoelectrics

Here we address only crystalline materials where both $\mathbf{P}_0$ and $\mathbf{M}_0$ vanish. The most general second order expansion of the free energy reads

$$\mathscr{F}(\mathbf{E}, \mathbf{H}) = \mathscr{F}_0 - \frac{1}{8\pi}\mathbf{E}\overset{\leftrightarrow}{\varepsilon}\mathbf{E} - \frac{1}{8\pi}\mathbf{H}\overset{\leftrightarrow}{\mu}\mathbf{H} - \frac{1}{4\pi}\mathbf{E}\overset{\leftrightarrow}{\alpha}\mathbf{H}, \qquad (A.3)$$

where textbooks ignore the last term here, and adopt Eq. (A.1) instead.

In high-energy physics a particle related to a $\mathbf{B} \cdot \mathbf{E}$ coupling term added to the electromagnetic Lagrangian has been postulated in the 1970s and dubbed "axion" (see Sec. A.6). However, it is was known since much earlier that magnetoelectric coupling does occur in condensed matter. In 1960 Dzyaloshinskii [239] realized that the magnetoelectric coupling tensor $\overset{\leftrightarrow}{\alpha}$ is in general nonzero in crystals where both inversion symmetry and T-symmetry are absent in the ground state. In fact $\mathbf{E}$ is odd under inversion and even under time-reversal, while the opposite happens for $\mathbf{B}$: therefore their combination in the last term of Eq. (A.3) is odd under each of the two transformation.

From Eq. (A.3) the conjugate variables are

$$\begin{aligned}\mathbf{D} &= \mathbf{E} + 4\pi\mathbf{P} = -4\pi\frac{\partial\mathscr{F}}{\partial\mathbf{E}} = \overset{\leftrightarrow}{\varepsilon}\mathbf{E} + \overset{\leftrightarrow}{\alpha}\mathbf{H}; \\ \mathbf{B} &= \mathbf{H} + 4\pi\mathbf{M} = -4\pi\frac{\partial\mathscr{F}}{\partial\mathbf{H}} = \overset{\leftrightarrow}{\alpha}{}^{\dagger}\mathbf{E} + \overset{\leftrightarrow}{\mu}\mathbf{H}.\end{aligned} \qquad (A.4)$$

Therefore in magnetoelectrics an $\mathbf{H}$ field induces an electrical polarization at zero $\mathbf{E}$ field, and conversely an $\mathbf{E}$ field induces magnetization at zero $\mathbf{H}$.

In its original paper, Dzyaloshinskii [239] proposed to look for the magnetoelectric effect in Cr_2O_3, which is non centrosymmetric and antiferromagnetic. Shortly afterwards, Soviet physicists indeed experimentally demonstrated both the direct and the converse magnetoelectric effect in Cr_2O_3 [240]. In this material the effect is by far too weak for being of technological use. There has been a resurgence of interest in the magnetoelectric effect; the search for materials where the effect could be stronger has been rather active in the last 15 years.

A.5 Parsing the magnetoelectric effect

In a non magnetoelectric cystal—where $\overset{\leftrightarrow}{\alpha}= 0$—the only field coupled to lattice coordinates is **E**: a magnetic field does not exert any force on the nuclei at rest. The effects of lattice-field coupling on the dielectric tensor $\overset{\leftrightarrow}{\varepsilon}$ and on the zone-center vibrational modes are analyzed in detail in Appendix D.

Whenever instead $\overset{\leftrightarrow}{\alpha}\neq 0$ both fields **E** and **H** are coupled to lattice displacements, formally on equal footing. As a consequence *all three* material constants $\overset{\leftrightarrow}{\varepsilon}$, $\overset{\leftrightarrow}{\mu}$, $\overset{\leftrightarrow}{\alpha}$ have an electronic contribution and a lattice contribution. The lattice-field coupling in magnetoelectrics has been investigated in the literature: the results presented in Appendix D for simple dielectrics have been generalised to linear magnetoelectrics. We do not present such generalisations here, and we refer to the original literature [241, 242, 243, 244].

Next, we consider the purely electronic—also called "clamped nuclei" response functions. Since the macroscopic magnetisation **M** is the sum of spin magnetization and orbital magnetization, Eq. (A.4) shows that both the clamped-nuclei $\overset{\leftrightarrow}{\mu}$ and $\overset{\leftrightarrow}{\alpha}$ have a spin and an orbital contributions. The full susceptibility tensor $\overset{\leftrightarrow}{\alpha}$ of the paradigmatic magnetoelectric material Cr_2O_3 has been computed in 2012 [245]: the spin and lattice contributions are separately addressed, and in both cases the response is decomposed into lattice and electronic parts.

A.6 Axion term

Here we adopt the $(\mathbf{E}, \mathbf{B})$ choice, according to the discussion at the end of Sec. A.2. Since second order energies depend on first order wavefunctions, one needs the linear response induced in the Bloch orbitals by either **E** or **B**, and then uses them to evaluate the induced magnetization or polarization, respectively: both approaches are feasible [63].

A very interesting geometrical feature contributes to the clamped-nuclei orbital term. Besides the genuine linear-response (a.k.a. Kubo) quantities, this response includes a *ground-state* term, to be evaluated with the Bloch states of the *unperturbed* crystal. This contribution to the magnetoelectic susceptibility $\overset{\leftrightarrow}{\alpha}$ is scalar (more precisely a pseudoscalar), hence it provides a contribution to the induced magnetization parallel to **E**, or equivalently a contribution to polarization parallel to **B**.

Within PBCs this contribution is a novel geometrical quantity, discussed above in Ch. 6, and given by the Brillouin-zone integral of a Chern-Simons 3-form, Eq. (6.5) [137, 136, 60, 61]. The peculiar feature is that this geometrical term is multivalued and thus reminiscent of the Berry-phase expression for the spontaneous polarization:

it remains ambiguous until the surface termination of the $3d$ sample has been specified. The multivalued nature of the Chern-Simons term is made explicit by writing it in terms of an angle θ as:

$$\alpha^{\mathrm{CS}} = \frac{1}{4\pi^2} \frac{e^2}{\hbar c} \theta.$$

(A.5)

The term "axion" comes from high-energy physics and was coined by F. Wilczek in 1977-78. It is based on writing the Lagrangian for the electromagnetic field as

$$\mathcal{L} = -\rho\,\Phi + \frac{1}{c}\mathbf{j}\cdot\mathbf{A} + \frac{1}{8\pi}E^2 - \frac{1}{8\pi}B^2 + \frac{\vartheta}{4\pi}\mathbf{E}\cdot\mathbf{B},$$

(A.6)

where the postulated axion term is the last one. The equation of motion is unaffected if ϑ is constant in space-time. Without the axion term the standard $\mathcal{L}$ is a relativistic scalar in 4d space-time; the axion term is pseudoscalar, and would account for the (very rare) processes which break charge-parity symmetry.

Appendix B

Fundamentals of piezoelectricity

B.1 Generalities

Piezoelectricity was first demonstrated in 1880 by Pierre and Jacques Curie. Nowadays it is the most common effect industrially used to convert electrical signals into mechanical ones, and conversely. Therefore the occurrence of the piezoelectric effect in everyday's life is ubiquitous. We make use of literally thousands of piezoelectric sensors and actuators in our cars, computers....

Piezoelectricity is symmetry forbidden in centrosymmetric crystals. The most symmetrical crystal structure where piezoelectricity is allowed is zincblende: therein, uniaxial strain along the (1,1,1) direction is coupled to a macroscopic $\mathbf{E}$ field along the same direction. Therefore not all piezoelectric materials are pyroelectric or ferroelectric, but the converse is true: all pyroelectric and ferroelectric materials are also piezoelectric. A clarification about semantics is in order: both pyroelectric and ferroelectric crystals have a preferred axis, along which the spontaneous polarization $\mathbf{P}_0$ is oriented. In ferroelectrics $\mathbf{P}_0$ is switchable (without crossing a metallic state), while in pyroelectrics switching is impossible: breaking and reforming of covalent bonds would be needed. The simplest pyroelectric symmetry class—i.e. where $\mathbf{P}_0 \neq 0$ is allowed—is wurtzite (see Fig. 5.2 and the discussion about it), while the prototypical ferroelectric is $BaTiO_3$, whose crystal structure is a cubic perovskite undergoing spontaneous inversion-symmetry breaking.

A wurtzite material much in fashion nowadays is GaN (see Nobel prize 2014). Although the industrial interest in this material is about light emission and not about the piezoelectric effect, its piezoelectricity has been thoroughly investigated [246].

Most piezoelectrics in industrial and commercial use are ferroelectric materials, but they are generally ceramics, not crystalline. A new class of useful materials has been discovered in 1997 [247]: they are single-crystal solid-solution ferroelectrics. Some of these new materials have found their way into industrial applications, e.g.

in medical echographs.

For reasons explained in Sec. 5.1 (and in the previous Appendix) the most convenient electrical variable in condensed matter is the field $\mathbf{E}$ (not $\mathbf{D}$). We remind that, when $\mathbf{E}$ is chosen as independent variable, the macroscopic free energy density in isotropic dielectric media is $\mathscr{F} = -\varepsilon E^2/(8\pi)$ in Gaussian units. As for Eq. (A.1), we draw attention to the *minus* sign [95]. The conjugate field obtains as $\mathbf{D} = -4\pi \partial \mathscr{F}/\partial \mathbf{E}$.

Here we are going to introduce macroscopic strain $\overset{\leftrightarrow}{\eta}$ as an additional independent variable besides $\mathbf{E}$, and we address anisotropic media. Given that we are going to address strain and volume changes, it is better to switch from the free energy density $\mathscr{F}$ to the free energy *per cell* $\mathcal{F} = V_{\text{cell}}\mathscr{F}$, where V_{cell} is the *equilibrium* volume.

It is expedient to use compact notations, where the Cartesian indices of the relevant tensors (of various ranks) are left implicit. The macroscopic material tensors discussed in this Appendix are genuinely static ones, i.e. include the lattice contribution. We anticipate that, instead, in Appendix C the lattice contribution to the dielectric constant (or tensor) will be separately investigated.

B.2   First order properties

The first order expansion of the free energy per cell is

$$
\begin{aligned}
\mathcal{F}(\overset{\leftrightarrow}{\eta}, \mathbf{E}) &= V_{\text{cell}}\mathscr{F}_0 + \left.\frac{\partial \mathcal{F}}{\partial \overset{\leftrightarrow}{\eta}}\right|_0 \cdot \overset{\leftrightarrow}{\eta} + \left.\frac{\partial \mathcal{F}}{\partial \mathbf{E}}\right|_0 \cdot \mathbf{E} + \quad \dots \\
&= \mathcal{F}_0 - V_{\text{cell}} \overset{\leftrightarrow}{\sigma}_0 \cdot \overset{\leftrightarrow}{\eta} - V_{\text{cell}}\mathbf{P}_0 \cdot \mathbf{E} + \quad \dots,
\end{aligned}
\tag{B.1}
$$

where $\mathbf{P}_0$ is the spontaneous polarization in zero field. Since we are by definition expanding around the equilibrium crystal structure, the macroscopic stress tensor $\overset{\leftrightarrow}{\sigma}_0$ in Eq. (B.1) vanishes.

However, the definition of "equilibrium" deserves a clarification. We remind that $\mathbf{E}$ is the *internal* (screened) field inside the material. According to our discussion in Sec. 5.1, whenever $\mathbf{P}_0 \neq 0$ the field $\mathbf{E}$ is in general nonzero inside a finite sample of arbitrary shape in zero *external* field. Therefore a slab, cut parallel to $\mathbf{P}_0$, and in zero external field, is in equilibrium and unstrained. Instead—according to our expansion in Eq. (B.1)—for a slab normal to $\mathbf{P}_0$ and in zero external field we may consider two cases: (1) If ideally the structure is kept unrelaxed the stress in nonzero (and proportional to $\mathbf{P}_0$); (2) In the relaxed structure the stress is zero, but the equilibrium strain is nonzero (and proportional to $\mathbf{P}_0$). This will appear more clearly when the second order expansion is considered.

B.3 Second order properties

Setting $\overset{\leftrightarrow}{\sigma}_0 = 0$ the expansion up to second order is

$$
\begin{aligned}
\mathcal{F}(\overset{\leftrightarrow}{\eta}, \mathbf{E}) \;\simeq\; & \mathcal{F}_0 \;-\; V_{\text{cell}} \mathbf{P}_0 \cdot \mathbf{E} \\
& + \; \frac{1}{2} \left. \frac{\partial^2 \mathcal{F}}{\partial \overset{\leftrightarrow}{\eta}\, \partial \overset{\leftrightarrow}{\eta'}} \right|_0 \overset{\leftrightarrow}{\eta}\, \overset{\leftrightarrow}{\eta} \;+\; \left. \frac{\partial^2 \mathcal{F}}{\partial \overset{\leftrightarrow}{\eta}\, \partial \mathbf{E}} \right|_0 \overset{\leftrightarrow}{\eta}\; \mathbf{E} \;+\; \frac{1}{2} \left. \frac{\partial^2 \mathcal{F}}{\partial \mathbf{E}\, \partial \mathbf{E'}} \right|_0 \mathbf{E}\, \mathbf{E},
\end{aligned}
\tag{B.2}
$$

where in the second derivatives we identify the elastic constants $\hat{C}$ (4th rank tensor), piezoelectric tensor $\hat{e}$ (3rd rank tensor), and dielectric tensor $\overset{\leftrightarrow}{\varepsilon}$ (2nd rank tensor), respectively. More precisely:

$$
\hat{C} = \frac{1}{V_{\text{cell}}} \left. \frac{\partial^2 \mathcal{F}}{\partial \overset{\leftrightarrow}{\eta}\, \partial \overset{\leftrightarrow}{\eta'}} \right|_0, \qquad \hat{e} = -\frac{1}{V_{\text{cell}}} \left. \frac{\partial^2 \mathcal{F}}{\partial \overset{\leftrightarrow}{\eta}\, \partial \mathbf{E}} \right|_0, \qquad \overset{\leftrightarrow}{\varepsilon} = -\frac{4\pi}{V_{\text{cell}}} \left. \frac{\partial^2 \mathcal{F}}{\partial \mathbf{E}\, \partial \mathbf{E'}} \right|_0. \tag{B.3}
$$

We therefore rewrite the second order expansion as

$$
\begin{aligned}
\mathcal{F}(\overset{\leftrightarrow}{\eta}, \mathbf{E}) \;\simeq\; & \mathcal{F}_0 \;-\; V_{\text{cell}} \mathbf{P}_0 \cdot \mathbf{E} \\
& + \; \frac{V_{\text{cell}}}{2} \overset{\leftrightarrow}{\eta}\, \hat{C}\, \overset{\leftrightarrow}{\eta} \;-\; V_{\text{cell}} \mathbf{E}\, \hat{e}\, \overset{\leftrightarrow}{\eta} \;-\; \frac{V_{\text{cell}}}{2} \mathbf{E}\, \overset{\leftrightarrow}{\varepsilon}\, \mathbf{E},
\end{aligned}
\tag{B.4}
$$

and we remind that—as stressed above—all material constants entering Eq. (B.4) are defined at $\overset{\leftrightarrow}{\eta} = 0$ and $\mathbf{E} = 0$.

The conjugate variables are

$$
\overset{\leftrightarrow}{\sigma} \;=\; -\hat{C}\, \overset{\leftrightarrow}{\eta} + \hat{e}^{\dagger}\, \mathbf{E} \tag{B.5}
$$

$$
\mathbf{D} \;=\; \hat{e}\, \overset{\leftrightarrow}{\eta} \;+\; \overset{\leftrightarrow}{\varepsilon}\, \mathbf{E} \;+\; 4\pi \mathbf{P}_0; \tag{B.6}
$$

here the dagger indicates the transpose, and $\hat{e}^{\dagger}$ is called the "converse" piezoelectric tensor. Eq. (B.6) can be recast as

$$
\mathbf{P} - \mathbf{P}_0 = \hat{e}\, \overset{\leftrightarrow}{\eta} + \frac{1}{4\pi}(\overset{\leftrightarrow}{\varepsilon} - \overset{\leftrightarrow}{1})\, \mathbf{E}. \tag{B.7}
$$

Therefore the piezoelectric tensor can be defined as either the stress linearly induced by a unit field at zero strain ("converse"), or the polarization linearly induced by unit strain at zero field ("direct").

B.4 Open circuit vs. closed circuit

In Fig. B.1 we show two different ideal measurements of a piezoelectric coefficient. The crystal is uniaxially strained along one of its piezoelectric axes.

In the left (short-circuit) figure one measures the time-integrated current flowing through the sample, and we remind that the integrated current is the polarization difference. The field $\mathbf{E}$ is zero at all times, hence according to Eq. (B.7) the measurement directly provides the vertical component of $\hat{e}$. We pause at this point to remind that, as stressed in Sec. 5.2, the modern theory of polarization directly evaluates the integrated current at zero field. Since currents are easer to measure than charges, the ideal experiment in the left figure is indeed close to the real experiments.

In the right (closed-circuit) figure one aims at measuring the dipole of the slab, or equivalently the strain-induced surface charge. In this setting we have $\mathbf{D} = 0$, not $\mathbf{E} = 0$, therefore for $\mathbf{P}$ normal to the slab $\mathbf{E} = -4\pi\mathbf{P}$. Replacing this into Eq. (B.7) one gets

$$\mathbf{P} = (\overset{\leftrightarrow}{\varepsilon})^{-1}(\hat{e}\overset{\leftrightarrow}{\eta} + \mathbf{P}_0). \tag{B.8}$$

In the simple case where $\mathbf{P}_0 = 0$ (e.g. in zincblende crystals) Eq. (B.8) shows that the open-circuit and closed-circuits measurements are simply related via the dielectric tensor: this is no small difference, since dielectric constants in common dielectrics are of the order $5-10$.

The case with $\mathbf{P}_0 \neq 0$, i.e. pyroelectric and ferroelectric materials, deserves further discussion. Let us assume throughout that even $\mathbf{P}_0$ is normal to the slab. At zero strain we set $\overset{\leftrightarrow}{\eta} = 0$ and we find that $\mathbf{P}$ is related to $\mathbf{P}_0$ by the inverse dielectric tensor. This is in full agreement with our discussion in Sec. 5.1 about longitudinal vs. transverse polarization; see also Fig. 5.1. For a more complete treatment see Ref. [8]. The open-circuit piezoelectric tensor is given therefore by

$$\Delta\mathbf{P} = (\overset{\leftrightarrow}{\varepsilon})^{-1}(\hat{e}\overset{\leftrightarrow}{\eta} + \mathbf{P}_0) - (\overset{\leftrightarrow}{\varepsilon})^{-1}\mathbf{P}_0 = (\overset{\leftrightarrow}{\varepsilon})^{-1}\hat{e}\overset{\leftrightarrow}{\eta}. \tag{B.9}$$

The formula looks therefore the same as for the $\mathbf{P}_0 = 0$ case, but a final subtlety is
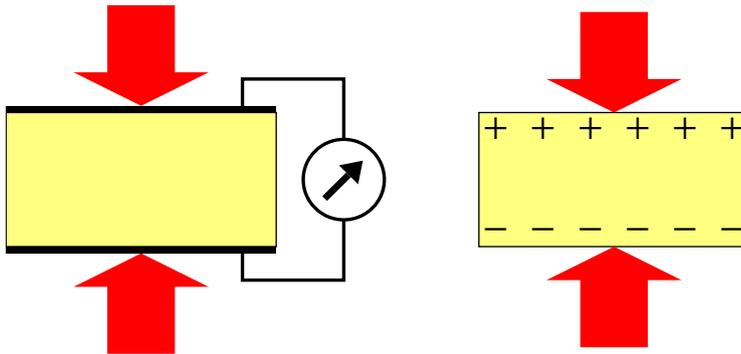

Figure B.1: Ideal measurements of the piezoelectric effect. Left: closed circuit. Right: open circuit.

worth pointing out.

The free energy expansion in Eq. (B.4) is around the equilibrium geometry in zero field, and in fact according to Eq. (B.5) both strain and stress are zero when $\mathbf{E} = 0$. But in a slab where $\mathbf{P}_0$ is normal to the slab $\mathbf{E}$ is nonzero, as thoroughly discussed in Sec. 5.1. Hence in a free standing slab in zero *external* field, and zero stress, the equilibrium strain is nonzero. In fact, from Eqs. (B.5) and (B.6), the equilibrium obtains by solving the linear system

$$0 \;=\; -\hat{C} \stackrel{\leftrightarrow}{\eta} + \hat{e}^{\dagger} \mathbf{E} \tag{B.10}$$

$$0 \;=\; \hat{e} \stackrel{\leftrightarrow}{\eta} + \stackrel{\leftrightarrow}{\varepsilon} \mathbf{E} + 4\pi \mathbf{P}_0 \;, \tag{B.11}$$

which provides the values of $\mathbf{E}$ (depolarisation field) and η (equilibrium strain in a longitudinal geometry). Therefore the open-circuit response in Eq. (B.9) is indeed the response to the *additional* strain besides the equilibrium strain.

B.5 Piezoelectricity from first principles

So far we have given a macroscopic phenomenological treatment of the linear piezoelectric effect; the underlying microscopic concepts are nowadays rooted in the modern theory of polarization. It is interesting, nonetheless, to address the theory of piezoelectricity within an historical perspective.

Piezoelectricity has been for years an intriguing problem. The *formal* proof that piezoelectricity is a well defined bulk property—independent of surface termination—is due to R. M. Martin [248], in a paper appeared in 1972 under the lapidary title of "Piezoelectricity" (no subtitle!). This proof has been challenged, and the debate lasted for about 20 years [249]. The very first quantum-mechanical calculation ever of a piezoelectric tensor was published in 1989 [250]. Therein, the "converse" definition of Eq. (B.5) was adopted, i.e. the computed quantity was the stress linearly induced by a unit field in III-V semiconductors. Since the calculation was in terms of Bloch orbitals within PBCs, Ref. [250] provided further evidence (if any was needed) that piezoelectricity is indeed a bulk effect.

After the advent of the modern theory of polarization in the early 1990s it became possible to implement even the "direct" definition of Eq. (B.7), i.e. to compute the strain-induced linear polarization as a numerical derivative, via the modern Berry-phase approach.

Appendix C

Linear-response theory

C.1 Definitions

To start with, we fix our conventions about Fourier transforms:

$$f(\omega) = \int_{-\infty}^{\infty} dt\, e^{i\omega t} f(t) \qquad f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega\, e^{-i\omega t} f(\omega). \qquad (C.1)$$

Suppose we have a general input signal $f_{\text{input}}(t)$ and the corresponding output $f_{\text{output}}(t)$, which is due to the response of a time-independent physical system. The most general linear response is given by a convolution

$$f_{\text{output}}(t) = \int_{-\infty}^{\infty} dt'\, \chi(t - t') f_{\text{input}}(t'). \qquad (C.2)$$

The response χ, also known as generalised susceptibility, can be equivalently written as the functional derivative

$$\chi(t - t') = \frac{\delta f_{\text{output}}(t)}{\delta f_{\text{input}}(t')}, \qquad (C.3)$$

evaluated at equilibrium. The response is therefore an equilibrium property of the system in absence of perturbation. The convolution theorem yields

$$f_{\text{output}}(\omega) = \chi(\omega)\, f_{\text{input}}(\omega). \qquad (C.4)$$

C.2 Causality

If the input signal is chosen as $f_{\text{input}}(t) = \delta(t)$, then we have $f_{\text{output}}(t) = \chi(t)$: in a causal system therefore $\chi(t) = 0$ for $t < 0$.

We pause to notice that not all linear systems must be causal: an obvious example of a noncausal system is an amplifier. An amplifier can start self oscillations

174

(Larsen effect), which quickly become anharmonic; however it needs to get energy from some power source.

We also stress that causality and dissipation are *not* synonymous. While a dissipative system is causal, one can address causal systems where no mechanism allows for dissipation: the paradigmatic example is the (undamped) harmonic oscillator in classical mechanics. A formal trick to account for causality is to consider a dissipative system in the nondissipative limit.

The Fourier transform of a causal $\chi(t)$ is

$$\chi(\omega) = \chi'(\omega) + i\chi''(\omega) = \int_0^\infty dt \, e^{i\omega t} \chi(t). \tag{C.5}$$

We only address a response $\chi(t)$ which is real in the time domain: therefore $\chi'(\omega)$ is even and $\chi''(\omega)$ is odd. Exploiting this fact, the antitransform is

$$\chi(t) = \frac{1}{2\pi} \int_{-\infty}^\infty d\omega \, [\, \chi'(\omega) \cos \omega t + \chi''(\omega) \sin \omega t \,]. \tag{C.6}$$

The two terms in the integrand cancel for $t < 0$, and provide an identical contribution for $t > 0$:

$$\chi(t) = \frac{1}{\pi} \int_{-\infty}^\infty d\omega \, \chi'(\omega) \, e^{-i\omega t}, \qquad t > 0. \tag{C.7}$$

In order to prove this, we exploit the identity $\chi(t) = [\chi(t) + \chi(-t)]\,\theta(t)$, hence

$$\chi(t) = \frac{1}{2\pi} \int_{-\infty}^\infty d\omega \, \chi(\omega)[\, e^{-i\omega t} + e^{i\omega t} \,], \qquad t > 0. \tag{C.8}$$

Since $\chi''(\omega)$ is odd, we immediately arrive at Eq. (C.7).

C.3   Kramers-Kronig relationships

The Kramers-Kronig relationships are most frequently proved by addressing the Fourier transform of the response function in the complex ω plane. Here we provide a less common—and to my taste more elegant—proof, found in the literature.

It is expedient to write

$$\chi(t) = \frac{\chi(t) - \chi(-t)}{2} + \mathrm{sgn}(t)\frac{\chi(t) - \chi(-t)}{2}, \tag{C.9}$$

where "sgn" is the signum function, and to extend the integral to $(-\infty, \infty)$:

$$i\chi''(\omega) = \frac{1}{2} \int_{-\infty}^\infty dt \, e^{i\omega t}[\chi(t) - \chi(-t)] \tag{C.10}$$

$$\chi'(\omega) = \frac{1}{2} \int_{-\infty}^\infty dt \, e^{i\omega t}\mathrm{sgn}(t)\,[\chi(t) - \chi(-t)]. \tag{C.11}$$

We notice then that the Fourier transform of $\text{sgn}(t)$ is the distribution $2i/\omega$, understood as a principal part. Using then the convolution theorem in Eq. (C.11), and exploiting Eq. (C.10) we have

$$\chi'(\omega) = \frac{2i}{2\pi} \int_{-\infty}^{\infty} d\omega' \, \frac{i\chi''(\omega')}{\omega - \omega'} = \frac{1}{\pi} \int_{-\infty}^{\infty} d\omega' \, \frac{\chi''(\omega')}{\omega' - \omega}. \tag{C.12}$$

The real and imaginary parts of $\chi(\omega)$ are related via an Hilbert transform. The inverse relationship reads

$$\chi''(\omega) = -\frac{1}{\pi} \int_{-\infty}^{\infty} d\omega' \, \frac{\chi'(\omega')}{\omega' - \omega}. \tag{C.13}$$

Finally, we may take advantage once more of the fact that $\chi'(\omega)$ is even and $\chi''(\omega)$ is odd, arriving at the Kramers-Kronig relationships in the equivalent form

$$\begin{aligned}
\chi'(\omega) &= \frac{2}{\pi} \int_{0}^{\infty} d\omega' \, \frac{\omega' \chi''(\omega')}{\omega'^2 - \omega^2} \\
\chi''(\omega) &= -\frac{2\omega}{\pi} \int_{0}^{\infty} d\omega' \, \frac{\chi'(\omega')}{\omega'^2 - \omega^2}.
\end{aligned} \tag{C.14}$$

Finally, we notice a subtle feature occurring when considering response functions which are related by a time derivative, i.e.

$$X'(t) = \chi(t). \tag{C.15}$$

If $X(t)$ is causal, then $\chi(t)$ is obviously causal. The example most relevant to the present notes concerns electrical polarizability χ and conductivity σ. In fact $\mathbf{P} = \chi \mathbf{E}$, $\mathbf{j} = \sigma \mathbf{E}$, and $\mathbf{j} = d\mathbf{P}/dt$.

If $X(\omega)$ obeys Kramers-Kronig, then even $\chi(\omega) = -i\omega X(\omega)$ also obeys. Instead, if $\chi(\omega)$ is causal, one would say that $X(\omega) = i\chi(\omega)/\omega$ is causal, but this is in general incorrect: one has to fix the integration constant in the time domain. The correct causal response function is

$$X(\omega) = \left[\frac{i}{\omega} + \pi\delta(\omega) \right] \chi(\omega), \tag{C.16}$$

where the two terms in parenthesis are easily recognised as the Fourier transform of the causal function $\theta(t) = [\text{sgn}(t) + 1]/2$; the second term can be neglected only when $\chi(\omega)$ vanishes at $\omega = 0$.

C.4 Energy dissipation

So far, we have been very general as for the nature of $f_{\text{input}}(t)$ and $f_{\text{output}}(t)$. In order to introduce energy considerations, we identify $f_{\text{input}}(t)$ with a generalized force $f(t)$, and $f_{\text{output}}(t)$ with a generalized coordinate $u(t)$:

$$u(t) = \int_{-\infty}^{\infty} dt' \, \chi(t-t')f(t'), \qquad u(\omega) = \chi(\omega)f(\omega). \qquad \text{(C.17)}$$

The power applied is $P(t) = f(t)\dot{u}(t)$; for a monochromatic pertubation at frequency ω one has

$$
\begin{aligned}
P(t) &= -[\text{Re } f(\omega)\text{e}^{-i\omega t}]\,[\text{Re } i\omega u(\omega)\text{e}^{-i\omega t}] && \text{(C.18)}\\
&= \omega[\text{Re } f(\omega)\text{e}^{-i\omega t}]\,[\text{Im } \chi(\omega)f(\omega)\text{e}^{-i\omega t}]\\
&= \frac{1}{4i}[f(\omega)\text{e}^{-i\omega t} + f^*(\omega)\text{e}^{i\omega t}]\,[\chi(\omega)f(\omega)\text{e}^{-i\omega t} - \chi^*(\omega)f^*(\omega)\text{e}^{i\omega t}],
\end{aligned}
$$

and average over the period yields:

$$P(t) = \frac{1}{4i}\omega|f(\omega)|^2[\chi(\omega) - \chi^*(\omega)] = \frac{1}{2}\omega|f(\omega)|^2\chi''(\omega). \qquad \text{(C.19)}$$

The imaginary part $\chi''(\omega)$ of the susceptibility—i.e. the out-of-phase response—is therefore the dissipative part; the in-phase term $\chi'(\omega)$ is called the reactive part. According to the second principle of thermodynamics $\chi''(\omega)$ is non negative whenever the responding system is not connected to an energy reservoir (e.g. as in the case of an amplifier). Furthermore, in order for dissipation to occur, a finite system must interact with a thermostat. It is remarkable that, instead, the thermostat is not needed in condensed matter theory, since an infinite system can be regarded as its own thermostat.

It is customary to extend the definition of $\chi(\omega)$ in the complex ω plane. A statement equivalent to the Kramers-Kronig relationships holds that the susceptibility $\chi(\omega)$ of a causal system is a regular function in the upper complex half plane, whereas in general it has poles in the lower half plane. To be more precise, for a nondissipative causal system (as e.g. for ideal undamped oscillators) the poles are infinitesimally close from below the the real ω-axis; for a dissipative system (e.g. damped oscillator) the poles are at a finite distance from the real axis. See also Sect. D.3.

A final caveat is in order. The location of the poles on the upper/lower half plane depends on our convention on Fourier transforms; the opposite choice may be found in the literature.

C.5   Fluctuation-dissipation theorem (classical)

We address the response of a classical system and the fluctuations of a classical variable at finite T; the quantum theorem is discussed below. The simplified presentation given in the following is inspired by Ch. 8 in Chandler's textbook [158]. We work in the canonical ensemble, hence we are addressing the *isothermal* response throughout the present Notes.

Let us assume that the *extensive* quantity of interest obtains as the trace (integral over the $3N$-dimensional phase space) of a classical dynamical variable $B(\mathbf{r}^N, \mathbf{p}^N)$. In order to simplify notations, and without loss of generality, we assume that the equilibrium value of B in the unperturbed system vanishes:

$$\langle B \rangle = \frac{\int d\mathbf{r}^N d\mathbf{p}^N \ B \, \mathrm{e}^{-\beta H}}{\int d\mathbf{r}^N d\mathbf{p}^N \ \mathrm{e}^{-\beta H}} = 0, \tag{C.20}$$

where H is the classical many-body Hamiltonian.

Suppose next that the system is perturbed at $t = -\infty$ with a time-independent perturbation ΔH. After relaxation, the new equilibrium value at $t = 0$ is

$$\bar{B}(0) = \frac{\int d\mathbf{r}^N d\mathbf{p}^N \ B \, \mathrm{e}^{-\beta(H+\Delta H)}}{\int d\mathbf{r}^N d\mathbf{p}^N \ \mathrm{e}^{-\beta(H+\Delta H)}} \neq 0. \tag{C.21}$$

Since we are addressing linear response, we neglect terms quadratic in ΔH and we get

$$\bar{B}(0) = -\beta \frac{\int d\mathbf{r}^N d\mathbf{p}^N \ \Delta H \, B \, \mathrm{e}^{-\beta H}}{\int d\mathbf{r}^N d\mathbf{p}^N \ \mathrm{e}^{-\beta H}} \ . \tag{C.22}$$

At time $t = 0$ we suddenly switch off the perturbation ΔH. The system is now out of equilibrium, and will evolve towards it, from $\bar{B}(0) \neq 0$ to $\bar{B}(\infty) = 0$.

We then define $B(t) = B(t; \mathbf{r}^N, \mathbf{p}^N)$, where each phase-space point in $B(t)$ evolves from the initial value $(\mathbf{r}^N, \mathbf{p}^N)$; the evolution is governed by the *unperturbed* Hamiltonian H. We assume that $\bar{B}(t)$ is a well behaved function of the $t = 0$ distribution in phase space. We thus have

$$\bar{B}(t) = -\beta \frac{\int d\mathbf{r}^N d\mathbf{p}^N \ \Delta H \, B(t) \, \mathrm{e}^{-\beta H}}{\int d\mathbf{r}^N d\mathbf{p}^N \ \mathrm{e}^{-\beta H}} \qquad t \geq 0. \tag{C.23}$$

Now we identify ΔH with *minus* the dynamical variable $A(\mathbf{r}^N, \mathbf{p}^N)$ and we assume for the sake of simplicity that even $\langle A \rangle = 0$. We then rewrite Eq. (C.23) as

$$\bar{B}(t) = \beta \ \frac{\int d\mathbf{r}^N d\mathbf{p}^N \ A(0) \, B(t) \, \mathrm{e}^{-\beta H}}{\int d\mathbf{r}^N d\mathbf{p}^N \ \mathrm{e}^{-\beta H}} = \beta \ \langle B(t) A(0) \rangle \qquad t \geq 0, \tag{C.24}$$

where we have introduced the simplified notation for the time-correlation function $\langle B(t)A(0)\rangle$. Notice that it is a measure of the *equilibrium fluctuations* of the dynamical variables A and B in the unperturbed system. Notice that $B(t)$ and $A(0)$ commute in the classical case.

We have previously seen that the generalized susceptibility $\chi(t)$ is the response to an instantaneous perturbation $\propto \delta(t)$, while instead the perturbation considered so far can be written as $-f_{\text{input}}(t)A$, where $f_{\text{input}}(t) = \theta(-t)$ is the externally applied disturbance coupled to the dynamical variable A. Given that $\delta(t) = d\theta(t)/dt$, clearly Eq. (C.24) determines the response. It is enough to identify $\bar{B}(t) = f_{\text{output}}(t)$; then from the basic Eq. (C.2) we have

$$f_{\text{output}}(t) = \beta \, \langle B(t)A(0)\rangle = \int_{-\infty}^{0} dt' \; \chi(t - t') = \int_{t}^{\infty} dt' \; \chi(t'). \qquad \text{(C.25)}$$

We thus arrive at the final statement of the (classical) fluctuation-dissipation theorem

$$
\begin{aligned}
\chi(t) &= 0, \quad t < 0 \\
\chi(t) &= -\beta \, \frac{d}{dt}\langle B(t)A(0)\rangle, \quad t > 0.
\end{aligned}
\qquad \text{(C.26)}
$$

In general, owing to dissipation, the correlation functions—and $\chi(t)$—are expected to vanish exponentially for $t \to \infty$; there are exceptions, though. Even in a simple system, like the hard-sphere fluid, the velocity-velocity autocorrelation function has a power-law tail, called "hydrodynamic tail". This important discovery came from computer simulations performed in in 1970 by Alder and Wainwright [251]. The analytical explanation came only *after* the computer experiment [252], as it happens in the most creative episodes of computational physics.

C.5.1 Green-Kubo formula for ionic conductivity

.

We consider here the ionic conductivity in insulating liquids (see Sec. 4.6.2), i.e. electrolytes, including molten salts, where—differently from the metallic case— charge transport is associated to mass transport:

$$\mathbf{j}(t) = \int dt' \; \sigma(t - t')\boldsymbol{\mathcal{E}}(t'). \qquad \text{(C.27)}$$

As shown in the following, $\sigma(\omega)$ obtains from the autocorrelation function of the equilibrium fluctuating charge current in the absence of an electric field [91, 92].

Steady-state currents require PBCs, hence the vector-potential gauge. If the ℓ-th ion has mass M_ℓ and charge Q_ℓ the perturbed Hamiltonian is

$$
\begin{aligned}
H + \Delta H &= \sum_\ell \frac{1}{2M_\ell} \left(\mathbf{p}_\ell - \frac{Q_\ell}{c} \mathbf{A} \right)^2 + U; \\
\Delta H &= -\mathbf{A} \cdot \sum_\ell \frac{Q_\ell}{M_\ell} \mathbf{p}_\ell = -\frac{V}{c} \mathbf{A} \cdot \mathbf{j},
\end{aligned}
\tag{C.28}
$$

where V is the sample volume end $\mathbf{j}$ is the macroscopic current density. Following the above prescription, at negative times we set a constant vector potential $\mathbf{A}$, and we switch it off suddenly at $t = 0$. If $B(t)$ is identified with the electrical current density, then Eq. (C.24) yields

$$
\mathbf{j}(t) = \frac{V\mathbf{A}}{3ck_{\mathrm{B}}T} \langle\, \mathbf{j}(t) \cdot \mathbf{j}(0) \,\rangle,
\tag{C.29}
$$

where we have exploited isotropy.

From the response to $\mathbf{A}$, Eq. (C.29), we easily find the response to $\boldsymbol{\mathcal{E}}$. In fact, since we have set $\mathbf{A}(t) = \mathbf{A}\theta(-t)$, then

$$
\boldsymbol{\mathcal{E}}(t) = -\frac{1}{c} \frac{d\mathbf{A}(t)}{dt} = \frac{1}{c} \mathbf{A}\delta(t),
\tag{C.30}
$$

hence the above prescription yields directly the response to a δ-like field:

$$
\sigma(t) = \frac{V}{3k_{\mathrm{B}}T} \langle\, \mathbf{j}(t) \cdot \mathbf{j}(0) \,\rangle \, \theta(t), .
\tag{C.31}
$$

Finally, we arrive at the famous Green-Kubo formula for dc ionic conductivity:

$$
\sigma = \frac{V}{3k_{\mathrm{B}}T} \int_0^\infty dt \, \langle\, \mathbf{j}(t) \cdot \mathbf{j}(0) \,\rangle.
\tag{C.32}
$$

C.5.2 Susceptibilities

In the simple case where A depends on the coordinates only (and not on the momenta), the instantaneous kick corresponds to apply a δ-like force at $t = 0$ on the system at equilibrium. It is then clear that at $t = 0$ the coordinates are continuous, while the momenta undergo a discontinuous jump. The system relaxes toward equilibrium, and $\chi(t)$ is the ensemble average of the dynamical variable B as a function of time.

Integrating by parts, the Fourier-transformed response is

$$
\chi(\omega) = \beta \langle A(0)B(0) \rangle + i\beta\omega \int_0^\infty dt \, \mathrm{e}^{i\omega t} \langle B(t)A(0) \rangle.
\tag{C.33}
$$

The generalised susceptibility has a real and an imaginary part, which obey by construction the Kramers-Kronig relationships:

$$\chi'(\omega) = \beta\langle A(0)B(0)\rangle - \beta\omega \int_0^\infty dt\, \langle B(t)A(0)\rangle \sin\omega t$$

$$\chi''(\omega) = \beta\omega \int_0^\infty dt\, \langle B(t)A(0)\rangle \cos\omega t. \tag{C.34}$$

The fluctuation-dissipation theorem applies unchanged to the case $A = B$, and the response is then given by the time autocorrelation function of the variable A. In the static limit the generalized susceptibility is

$$\chi_0 = \chi'(0) = \beta\langle A^2\rangle, \tag{C.35}$$

a result which can be directly proved in a simpler way.

The imaginary part is related to absorption, as shown above and also as it will become clear in the following: Secs. D.2 and D.11. For a system at equilibrium the zero of time is arbitrary, therefore $\langle A(t)A(0)\rangle = \langle A(0)A(-t)\rangle$. Since the (classical) dynamical variables are real and commute, the imaginary part of the response can be written as

$$\chi''(\omega) = \frac{\beta\omega}{2} \int_{-\infty}^\infty dt\, e^{i\omega t}\langle A(t)A(0)\rangle, \tag{C.36}$$

where the integral is often called "power spectrum".

In the present Notes we are going to consider in some detail the case where the extensive variable A is identified with the macroscopic dipole $\mathbf{d}$ of the system: see Sec. C.6; see also Eq. (D.96) below and the following text.

C.6 Finite-temperature polarizability

For pedagogical reasons we start addressing a classical model of a fluid (either liquid or gas) whose components are identical polar molecules. At the simplest level, all of the polarization is due to molecular orientation: the molecules are rigid and nonpolarizable. The paradigmatic system of this kind, used as a workhorse in many molecular dynamics simulation, is the Stockmayer fluid: therein the two-body interactions are Lennard-Jones supplemented with point dipoles.

There is no ambiguity about what the polarization $\mathbf{P}$ and the total dipole $\mathbf{d}$ are in a dipolar fluid: within either OBCs or PBCs the dipole is the sum of the point dipoles:

$$\mathbf{d} = V\mathbf{P} = \sum_j \mathbf{d}_j. \tag{C.37}$$

Instead, we have often stressed in the present Notes that in a general system (such e.g. a polar crystal) dealing with macroscopic polarization within PBCs is not so straightforward. The dipole $\mathbf{d} = V\mathbf{P}$ of the simulation cell *cannot* be decomposed in individual dipoles $\mathbf{d}_j$.

In presence of long range interactions the interpretation of the fluctuation-dissipation theorem requires some care. In order to see the problem, we start with a *bounded* and macroscopically homogeneous system within OBCs, and we apply an *external* field $\mathbf{E}_0$, uniform in space. We address the static case first. The field couples to the Hamiltonian via the term $-\mathbf{d} \cdot \mathbf{E}_0$. The dipole-dipole autocorrelation function provides, from Eq. (C.35), the response

$$\frac{\partial \mathbf{P}}{\partial \mathbf{E}_0} = \frac{1}{V} \frac{\partial \mathbf{d}}{\partial \mathbf{E}_0} = \frac{\beta}{3V} \langle \mathbf{d} \cdot \mathbf{d} \rangle. \tag{C.38}$$

We aim to relate this to the static dielectric constant of the fluid, whose definition is

$$\varepsilon_0 = 1 + 4\pi \frac{\partial \mathbf{P}}{\partial \mathbf{E}}, \tag{C.39}$$

where now $\mathbf{E}$ is the *screened* macroscopic field inside the polarized fluid.

The static dielectric constant ε_0 is a well defined bulk material property, while instead the relationship between $\mathbf{E}_0$ and $\mathbf{E}$ depend on the sample shape (see Sec. 5.1). A glance at Eq. (C.38) shows that $\langle d^2 \rangle$ must depend on shape as well: the physical origin of this is the depolarization field, which indeed affects the equilibrium quadratic fluctuation. Equivalently, the fluctuations depend on the boundary condition chosen for integrating Poisson equation in the selfconsistent Newton equations of motion.

C.6.1 Zero-field boundary conditions

The simplest choice is to evaluate the fluctuations at zero macroscopic field: this corresponds ideally to a finite sample coated with a metallic layer. In practice, one adopts instead PBCs and integrates Poisson equation by choosing the solution which is periodic over the simulation cell. In the molecular-dynamics jargon this choice goes under the name of Ewald-Kornfeld [253]. Within this choice—as thoroughly discussed in Sec. 5.1— there is no depolarization field, hence $\mathbf{E} = \mathbf{E}_0$.

Provided the equilibrium fluctuations are evaluated at zero macroscopic field, the dielectric constant has the simple expression

$$\varepsilon_0 = 1 + \frac{4\pi\beta}{3V} \langle d^2 \rangle. \tag{C.40}$$

We also display the expression for the imaginary part of the ω-dependent dielectric constant, which is the basic quantity to address infrared absorption.

When the dipole-dipole correlation functions are evaluated using periodic boundary conditions, the previous results and Eq. (C.36) yield

$$\varepsilon''(\omega) = \frac{2\pi\beta\omega}{3V} \int_{-\infty}^{\infty} dt \, e^{i\omega t} \langle \mathbf{d}(t) \cdot \mathbf{d}(0) \rangle. \tag{C.41}$$

The example of a fluid whose components are rigid molecules was used here for pedagogical purposes, but the classical fluctuation-dissipation theorem has a much wider scope. Since the pioneering work of Silvestrelli et al. in 1997 [93] first-principle infrared spectra are routinely evaluated from the dipole-dipole correlation function, where the polarization $\mathbf{P} = \mathbf{d}/V$ is provided by the modern theory of polarization, and its fluctuations are computed via a Car-Parrinello simulation. More about this will be said below, Sec. D.11. For the time being, we notice that $\mathbf{P}$ is due to both nuclei and electrons; in the adiabatic approximation the nuclear coordinates are purely classical, but their coupling to the field includes the electronic contribution in the form $-\mathbf{P} \cdot \mathbf{E}$. The case of a polar crystal in the harmonic regime, dealt with in detail in Appendix D, will make this point pretty clear.

C.6.2 Reaction field

In any quantum description of a condensed system within PBCs the adoption of the zero-field condition is almost mandatory. Only in this way, in fact, the electrostatic potential—and hence the Schrödinger Hamiltonian—is periodic over the simulation cell. In classical mechanics, instead, the potential plays a lesser role: the Newton equation of motions are directly formulated in terms of the electric field, not its potential. Other choices of boundary conditions may therefore be appealing.

Following Ch. 2 of the Landau-Lifshitz textbook [95], we consider an homogenous sphere of dielectric constant ε_0 embedded in a medium of dielectric constant $\varepsilon_{\mathrm{RF}}$ ("reaction field"). When an external constant field $\mathbf{E}_0$ is present in the surrounding dielectric the screened field $\mathbf{E}$ inside the sphere is [95]

$$\mathbf{E} = \frac{3\varepsilon_{\mathrm{RF}}}{2\varepsilon_{\mathrm{RF}} + \varepsilon_0} \mathbf{E}_0; \tag{C.42}$$

hence

$$\frac{\partial \mathbf{P}}{\partial \mathbf{E}_0} = \frac{3\varepsilon_{\mathrm{RF}}}{2\varepsilon_{\mathrm{RF}} + \varepsilon_0} \frac{\partial \mathbf{P}}{\partial \mathbf{E}}. \tag{C.43}$$

Next we need the bare field coupled to the molecular Hamiltonian, i.e. the field present in the empty spherical cavity. This easily obtains from the same Eq. (C.42) by replacing ε_0 with 1. The perturbation to the Hamiltonian is therefore

$$\Delta H = -\frac{3\varepsilon_{\mathrm{RF}}}{2\varepsilon_{\mathrm{RF}} + 1} \mathbf{d} \cdot \mathbf{E}_0. \tag{C.44}$$

The fluctuation-dissipation formula, Eq. (C.38), must be replaced with

$$\frac{\beta}{3V}\langle d^2 \rangle = \frac{2\varepsilon_{\mathrm{RF}}+1}{3\varepsilon_{\mathrm{RF}}}\frac{\partial \mathbf{P}}{\partial \mathbf{E}_0} = \frac{2\varepsilon_{\mathrm{RF}}+1}{2\varepsilon_{\mathrm{RF}}+\varepsilon_0}\frac{\partial \mathbf{P}}{\partial \mathbf{E}}. \tag{C.45}$$

Using then Eq. (C.39) we arrive at the final result

$$\frac{4\pi\beta}{3V}\langle d^2 \rangle = \frac{(2\varepsilon_{\mathrm{RF}}+1)(\varepsilon_0-1)}{2\varepsilon_{\mathrm{RF}}+\varepsilon_0}. \tag{C.46}$$

Setting $\varepsilon_{\mathrm{RF}}=\infty$ we retrieve Eq. (C.40). Another special case is $\varepsilon_{\mathrm{RF}}=\varepsilon_0$, leading to the (1939) Kirkwood-Fröhlich formula [254, 255]:

$$\frac{4\pi\beta}{3V}\langle d^2 \rangle = \frac{(2\varepsilon_0+1)(\varepsilon_0-1)}{3\varepsilon_0}. \tag{C.47}$$

The other interesting special case is $\varepsilon_{\mathrm{RF}}=1$, yielding the time-honored (1912) Debye formula

$$\frac{4\pi\beta}{3V}\langle d^2 \rangle = \frac{3(\varepsilon_0-1)}{2+\varepsilon_0}.. \tag{C.48}$$

This ideally corresponds to a spherical sample free-standing in vacuo: in practice, the formula apples to simulations performed within PBCs, but when the Coulomb interaction is truncated with a spherical cutoff. By contrast the zero-field formula, Eq. (C.40), requires the evaluation of conditionally convergent Ewald-like sums.

The dielectric constant is a well defined property of the bulk material: one gets the same ε_0 value by adopting different boundary conditions, and using the appropriate fluctuation formula for each of them. This has been first demonstrated by Neumann in 1983 by numerical simulations on the Stockmayer fluid [256].

It is pedagogically useful to provide the explicit expression for the depolarization field acting on the fluctuating unperturbed system. Starting from Eq. (C.42), we replace ε_0 with its definition, Eq. (C.39); the result can be recast as

$$\mathbf{E} = \frac{3\varepsilon_{\mathrm{RF}}}{2\varepsilon_{\mathrm{RF}}+1}\mathbf{E}_0 - \frac{4\pi}{2\varepsilon_{\mathrm{RF}}+1}\mathbf{P}. \tag{C.49}$$

The first term in the rhs is the bare field acting on the molecular system, as in Eq. (C.44), and the second term is the depolarization field. For the unperturbed system at thermal equilibrium $\mathbf{E}_0=0$ and the depolarization field is

$$\mathbf{E} = -\frac{4\pi}{2\varepsilon_{\mathrm{RF}}+1}\mathbf{P}. \tag{C.50}$$

The meaning of this is pretty clear: the fluctuating dipole of the molecular system induces a depolarization field $\mathbf{E}$ within the sphere. Two special cases are worth

184

mentioning: $\mathbf{E} = 0$ for $\varepsilon_{\mathrm{RF}} = \infty$ (metallic coating, no depolarization field), and $\mathbf{E} = -4\pi\mathbf{P}/3$ for $\varepsilon_{\mathrm{RF}} = 1$ (our sphere is free-standing in vacuo).

Some final words about the literature. The correct fluctuation formulæ–for dipolar systems as well as for ionic systems—first appeared in the 1980s [257, 256]. The original literature—as well as the more recent one—is plagued by clumsy and inelegant algebra. The simple proof provided here is, to the best of my knowledge, unpublished.

C.7 Quantum vs. classical

Before switching to linear-response theory and to fluctuation-dissipation in the quantum domain, it is expedient to emphasize analogies and differences.

The first difference concerns dissipation. The great majority of the systems dealt with in classical molecular dynamics is ergodic: there are exceptions, the most famous of them being the Fermi-Pasta-Ulam Hamiltonian [258, 259]. This was a great discovery made possible by computational physics; before Fermi-Pasta-Ulam (1953) it was postulated by everybody (Fermi included!) that all non separable Hamiltonians were ergodic in the thermodynamic limit. The discovery opened a new branch of physics: nonlinear systems and the physics of chaos [259].

Apart from such peculiar exceptions, "normal" systems are ergodic, evolve towards thermal equilibrium, fulfill the equipartition theorem, & the like. This is demonstrated since the birth of computational physics in the 1950s, where the first simulations on hard-sphere and Lennard-Jones fluids appeared. Dissipation is therefore present by itself in a canonical simulation, and the fluctuation-dissipation theorem directly provides the causal response functions. The same applies to Car-Parrinello simulations, where the nuclei are indeed classical particles. Matters are different within quantum mechanics, where the introduction of temperature and dissipation at the first-principle level is much more problematic.

In the present Notes we keep, as far as possible, the same notations adopted in the classical case. Now the dynamical variables A and B, and the Hamiltonian H, are Hermitian operators. They do not commute, while their classical counterparts do commute. The fluctuations addressed by the theorem are therefore fluctuations of the quantum variables A and B in the many-electron ground state of the unperturbed system.

C.8 Fluctuation-dissipation theorem (quantum)

In this section we give the zero-T formulation in detail. The finite-T quantum formula is presented at the end, Eq. (C.64), without proof.

We have seen above that $\chi(t)$ can be defined as the response in the observable B to an instantaneous δ-like "kick" at $t = 0$, where the perturbing probe is $\Delta H(t) = -\delta(t)A$. For $t < 0$ the system is in its (non degenerate) ground state; in order to simplify notations we assume—as we have done in the classical case— that the ground-state expectation value of both A and B vanish, i.e.

$$\langle A \rangle = \langle \Psi_0|A|\Psi_0 \rangle = 0, \quad \langle B \rangle = \langle \Psi_0|\hat{B}|\Psi_0 \rangle = 0, \quad t < 0. \tag{C.51}$$

The time-dependent Schrödinger equation is

$$H|\Psi(t)\rangle - \delta(t)A|\Psi(t)\rangle = i\hbar\frac{d}{dt}|\Psi(t)\rangle, \tag{C.52}$$

and the initial condition is $|\Psi(0^-)\rangle = |\Psi_0\rangle$. The temporal evolution of this state is

$$|\Psi(t)\rangle = \mathcal{T}\mathrm{e}^{-\frac{i}{\hbar}\int_{0^-}^{t} dt'[H-\delta(t')A]}\,|\Psi_0\rangle, \tag{C.53}$$

where $\mathcal{T}$ is the time-ordering operator; $|\Psi(t)\rangle$ has a discontinuous jump at $t = 0$, which to linear order in A is

$$|\Psi(0^+)\rangle = \left(1 + \frac{i}{\hbar}A\right)|\Psi_0\rangle = |\Psi_0\rangle + \frac{i}{\hbar}{\sum_{n\neq0}}'|\Psi_n\rangle\langle\Psi_n|A\,|\Psi_0\rangle. \tag{C.54}$$

This state evolves for $t > 0$ with the unperturbed Hamiltonian H:

$$|\Psi(t)\rangle = \mathrm{e}^{-iE_0t/\hbar}|\Psi_0\rangle + \frac{i}{\hbar}{\sum_{n\neq0}}'\mathrm{e}^{-iE_nt/\hbar}|\Psi_n\rangle\langle\Psi_n|A\,|\Psi_0\rangle, \quad t > 0. \tag{C.55}$$

Since we are at $T=0$, the response is causal but non dissipative: our $\chi(t)$ will have *undamped* oscillations.

Finally, the linear response is

$$\chi(t) = \langle B(t)\rangle = \theta(t)\,\langle\Psi(t)|B\,|\Psi(t)\rangle \tag{C.56}$$

$$= \frac{i}{\hbar}\theta(t){\sum_{n\neq0}}'(\mathrm{e}^{-i\omega_{0n}t}\langle\Psi_0|B\,|\Psi_n\rangle\langle\Psi_n|A\,|\Psi_0\rangle - \mathrm{e}^{i\omega_{0n}t}\langle\Psi_0|A\,|\Psi_n\rangle\langle\Psi_n|B\,|\Psi_0\rangle),$$

where $\omega_{0n} = (E_n - E_0)/\hbar$. Its Fourier transform obtains from

$$-i\int_{-\infty}^{\infty} dt\,\mathrm{e}^{i\omega t}\theta(t) = \mathcal{P}\frac{1}{\omega} - i\pi\delta(\omega) = \lim_{\eta\to0^+}\frac{1}{\omega + i\eta}, \tag{C.57}$$

and adopting the compact notation due to Zubarev [67, 156, 157]:

$$\langle\langle\hat{B}|\hat{A}\rangle\rangle_\omega = \frac{1}{\hbar}\lim_{\eta\to0^+}{\sum_{n\neq0}}'\left(\frac{\langle\Psi_0|\hat{B}|\Psi_n\rangle\langle\Psi_n|\hat{A}|\Psi_0\rangle}{\omega - \omega_{0n} + i\eta}\right.$$

$$\left. - \frac{\langle\Psi_0|\hat{A}|\Psi_n\rangle\langle\Psi_n|\hat{B}|\Psi_0\rangle}{\omega + \omega_{0n} + i\eta}\right). \tag{C.58}$$

The Kubo formula for the generalized susceptibility takes then the form

$$\chi(\omega) = -\langle\langle \hat{B}|\hat{A}\rangle\rangle_\omega. \tag{C.59}$$

We draw attention to the fact that the sign conventions adopted in these Notes agree with Zubarev [67, 156] and Chandler [158], but are opposite the the ones of McWeeny [157] and other textbooks. We have given here the many-body Kubo formula; for independent electrons Eq. (C.58) is easily transformed into a double sum over occupied and unoccupied orbitals.

The positive infinitesimal η in Eq. (C.58) enforces causality; dissipation and lifetime can be introduced phenomenologically by hand, e.g. by replacing η with a finite inverse relaxation time $1/\tau$. Insofar as η is infinitesimal, a closed system may only absorb at the discrete frequencies ω_{0n}, where $\chi''(\omega)$ is nonzero; it is actually δ-like, as from Eq. (C.57). No mechanism allows for energy dissipation into a thermostat (a.k.a. heath bath) at an arbitrary frequency insofar as the spectrum is discrete. In condensed matter one takes the thermodynamic limit, and the spectrum becomes continuous: in this case $\chi''(\omega)$ can be nonzero over a continuous range of frequencies even at infinitesimal η. In fact dissipation is an intrinsic feature of an extended system, which in a sense is its own heath bath (Sect. 3.2.5 in Ref. [166]).

The previous formulation is useful in order to explicitly provide the spectral representation of $\chi(\omega)$, Eq. (C.58) . Nonetheless a more compact and elegant formula obtains by adopting the Heisenberg representation. In fact Eq. (C.56) can be equivalently rewritten as

$$
\begin{aligned}
\chi(t) &= \theta(t)\,\langle\Psi(0^+)|\,e^{\frac{i}{\hbar}Ht}Be^{-\frac{i}{\hbar}Ht}\,|\Psi(0^+)\rangle = \langle\Psi(0^+)|\,B_H(t)\,|\Psi(0^+)\rangle \\
&= \frac{i}{\hbar}\theta(t)\,\langle\Psi_0|\,[B_H(t), A_H(0)]\,|\Psi_0\rangle.
\end{aligned} \tag{C.60}
$$

It is easy to prove that $\langle\Psi_0|\,[B_H(t-t'), A_H(0)]\,|\Psi_0\rangle = \langle\Psi_0|\,[B_H(t), A_H(t')]\,|\Psi_0\rangle$, hence we may also write

$$
\begin{aligned}
\chi(t-t') &= \frac{i}{\hbar}\theta(t-t')\,\langle\Psi_0|\,[B_H(t), A_H(t')]\,|\Psi_0\rangle \\
&= \frac{i}{\hbar}\theta(t-t')\,\mathrm{Tr}\,\{\hat{P}\,[B_H(t), A_H(t')]\,\},
\end{aligned} \tag{C.61}
$$

where $\hat{P} = |\Psi_0\rangle\langle\Psi_0|$ is the ground state projector. Zubarev defines his correlation function for complex "frequencies" z, i.e.

$$\langle\langle\hat{B}|\hat{A}\rangle\rangle_z = -\frac{i}{\hbar}\int_{-\infty}^{\infty} dt\,e^{izt}\,\theta(t)\,\mathrm{Tr}\,\{\hat{P}\,[B_H(t), A_H(t')]\,\}, \tag{C.62}$$

where the integral converges for z in upper half of the complex z plane. In fact the spectral representation in Eq. (C.58) immediately shows that the response

is analytical in the upper half, and has poles in the lower half, "infinitesimally close" to the real axis. This enforces causality, and suggests a simple picture. For $z = \omega + i\eta$, and finite positive η, the oscillations are (phenomenologically) damped with a relaxation time $\tau = 1/\eta$. The response is therefore dissipative, and the $\eta \to 0^+$ limit can be interpreted as the limit were the response is undamped (as it must be at $T = 0$), yet causal.

The finite-T Zubarev correlation formula appears as a simple modification of Eq. (C.62): it is enough to replace the ground-state projector therein with the finite-T equilibrium density matrix in the canonical ensemble, i.e.

$$\mathcal{P} \quad \to \quad \hat{\rho}_{\mathrm{eq}} = \frac{\mathrm{e}^{-\beta H}}{\mathrm{Tr}\ \mathrm{e}^{-\beta H}}; \tag{C.63}$$

$$\langle\langle \hat{B} | \hat{A} \rangle\rangle_z = -\frac{i}{\hbar} \int_{-\infty}^{\infty} dt\ \mathrm{e}^{izt}\ \theta(t)\ \mathrm{Tr}\ \{\hat{\rho}_{\mathrm{eq}} \left[B_H(t), A_H(t') \right]\}. \tag{C.64}$$

Despite the formal elegance and simplicity of this formula, its implementation at the first-principle level is a formidable task.

Appendix D

Advanced lattice dynamics

D.1 Huang's phenomenological theory

Huang's theory, published in 1950 [260, 261], is the simplest paradigm for lattice dynamics in presence of long-range forces. It addresses the zone-center optical modes of a cubic binary crystal. Zone-center modes are lattice-periodical, therefore there is only one microscopic dynamical variable: the relative sublattice displacement $\mathbf{u}$. The system is therefore a single oscillator, isotropic at the harmonic level, and where the electric field contributes to the restoring force.

We start from a free energy expression where the independent variable is the field $\mathbf{E}$, as in Appendices A and B, augmented here with the dynamical variable $\mathbf{u}$. The most general expansion to second order around equilibrium of the free energy *per cell* $\mathcal{F} = V_{\text{cell}}\mathscr{F}$ is

$$\mathcal{F}(E, \mathbf{u}) = \mathcal{F}_0 + \frac{1}{2}M\omega_{\text{TO}}^2\, u^2 - \frac{V_{\text{cell}}}{8\pi}\varepsilon_\infty E^2 - Z^*e\mathbf{u} \cdot \mathbf{E}. \tag{D.1}$$

The reasons for the symbols adopted will be clear in a moment. But we pause to stress that Eq. (D.1) is the most general second order expansion: we are making *no hypotheses* about the physical origin of the three expansion coefficients. Therefore all subsequent results are *exact* within the harmonic approximation, and must hold for lattice-dynamical models, for first-principle calculations, and for the experimental measurements as well.

If we identify M with the reduced mass of the two atoms, the three phenomenological parameters are ω_{TO}, ε_∞, and Z^* (the dimensionless Born charge). Notice that for $Z^* = 0$ the field and the lattice displacements are uncoupled. The conjugate variables are the force $\mathbf{f}$ and the field $\mathbf{D}$, i.e.

$$\mathbf{f} = -\frac{\partial \mathcal{F}}{\partial \mathbf{u}} = -M\omega_{\text{TO}}^2\mathbf{u} + Z^*e\mathbf{E} \tag{D.2}$$

$$\mathbf{D} = -\frac{4\pi}{V_{\text{cell}}}\frac{\partial \mathcal{F}}{\partial \mathbf{E}} = \varepsilon_\infty\mathbf{E} + \frac{4\pi}{V_{\text{cell}}}Z^*e\mathbf{u}. \tag{D.3}$$

From the second line it is clear that ε_∞ is the dielectric constant when the nuclei are kept "clamped" at their equilibrium position. This is the dielectric constant actually measured at frequencies much higher than the phonon frequencies, but lower than the electronic transition; it is sometimes named with the oxymoron "static high frequency" [97]. The physical origin of ε_∞ is clearly the dielectric response of the many-electron system: it is therefore also called electronic dielectric constant.

At equilibrium we set $\mathbf{f} = 0$, and we have

$$\begin{aligned}
\mathbf{u} &= \frac{Z^*e}{M\omega_{\mathrm{TO}}^2}\mathbf{E} \\
\mathbf{D} &= \left[\varepsilon_\infty + \frac{4\pi e^2(Z^*)^2}{V_{\mathrm{cell}}M\omega_{\mathrm{TO}}^2}\right]\mathbf{E}.
\end{aligned} \tag{D.4}$$

The first line yields the sublattice displacement, while the expression in parentheses in the second line is the genuinely static dielectric constant ε_0, including the lattice relaxation, with $\varepsilon_0 \geq \varepsilon_\infty$.

We now study forced oscillations: from Eqs. (D.2) and (D.3)

$$-M\omega^2\mathbf{u} = -M\omega_{\mathrm{TO}}^2\mathbf{u} + Z^*e\mathbf{E} \tag{D.5}$$

$$\mathbf{u} = \frac{Z^*e}{M(\omega_{\mathrm{TO}}^2 - \omega^2)}\mathbf{E} \tag{D.6}$$

$$\mathbf{D} = \left[\varepsilon_\infty + \frac{4\pi e^2(Z^*)^2}{V_{\mathrm{cell}}M(\omega_{\mathrm{TO}}^2 - \omega^2)}\right]\mathbf{E} = \varepsilon(\omega)\mathbf{E}. \tag{D.7}$$

The quantity in parentheses in the last expression is therefore the infrared dielectric constant $\varepsilon(\omega)$, with $\varepsilon(0) = \varepsilon_0$ and $\varepsilon(\infty) = \varepsilon_\infty$.

We have not identified the meaning of the other parameters yet. To this aim we notice that for a long-wavelength phonon of wave vector $\mathbf{q}$ the solid is macroscopically homogenous normal to $\mathbf{q}$, while its physical properties are modulated in the $\mathbf{q}$ direction. It follows that $\mathbf{E}$ vanishes normal to $\mathbf{q}$, while $\mathbf{D}$ vanishes parallel to $\mathbf{q}$. Setting $\mathbf{E} = 0$ in Eq. (D.6) confirms that the ω_{TO} parameter in the free-energy expansion is indeed the zone-center transverse optical frequency, while instead $\mathbf{D} = 0$ when $\varepsilon(\omega) = 0$. The (squared) longitudinal optical frequency ω_{LO} is therefore, from Eq. (D.7)

$$\omega_{\mathrm{LO}}^2 = \omega_{\mathrm{TO}}^2 + \frac{4\pi e^2(Z^*)^2}{\varepsilon_\infty V_{\mathrm{cell}}M}, \tag{D.8}$$

with $\omega_{\mathrm{LO}} \geq \omega_{\mathrm{TO}}$. We have already observed that the Born charge Z^* measures the coupling of the macroscopic field to the ionic motion. This coupling vanishes in nonpolar binary crystals, such as those having the diamond structure, where in fact no longitudinal-transverse (LT) splitting of zone-center optical phonons is observed.

The zone-center longitudinal oscillations in a polar material are driven by an extra restoring force, of electrical origin. This is the same mechanism which drives longitudinal plasma oscillations, and in fact if we ideally set $\omega_{\text{TO}} = 0$ in Eq. (D.8) we retrieve a squared plasma frequency, i.e. 4π times a squared charge density over a mass density; the interaction is screened with ε_∞. The relevant charge is the Born effective charge, and the relevant mass is the reduced mass.

From the previous equations it is straightforward to obtain the identity

$$\frac{\omega_{\text{LO}}^2}{\omega_{\text{TO}}^2} = \frac{\varepsilon_0}{\varepsilon_\infty},$$

(D.9)

i.e. the famous Lyddane-Sachs-Teller (1941) relationship [262]. It is remarkable that all microscopic parameters (force constants, masses, Born effective charges, cell volume) disappear from Eq. (D.9), which is *exact* in the harmonic regime. The identity relates quantities measured in quite different ways: inelastic neutron scattering—left hand side of Eq. (D.9)—vs. infrared spectroscopy—right hand side. A final comment about first-principle calculations: therein, the quantities directly provided by the code are ω_{TO}, ε_∞, and Z^* (and their generalization, see below) [104, 105, 106], while ω_{LO} and ε_0 are obtained from the phenomenological approach. Therefore Eq. (D.9) is fulfilled by construction.

D.2 Infrared absorption

The simple case of a cubic binary material is pedagogically useful to address dispersion and absorption in infrared spectroscopy. The ω-dependent dielectric constant found in Sec. D.1 is the real (dispersive) part of the complex response function $\varepsilon'(\omega) + i\varepsilon''(\omega)$, while absorption is provided by the imaginary part. The absorption coefficient per unit path length is related to $\varepsilon''(\omega)$ as [263]:

$$\alpha(\omega) = \frac{\omega}{c\,n(\omega)}\varepsilon''(\omega),$$

(D.10)

where $n(\omega)$ is the index of refraction.

We rewrite Eq. (D.7) as

$$\varepsilon'(\omega) = \varepsilon_\infty + \frac{4\pi e^2 (Z^*)^2}{V_{\text{cell}} M (\omega_{\text{TO}}^2 - \omega^2)}.$$

(D.11)

Causality requires that $\varepsilon'(\omega)$ and $\varepsilon''(\omega)$ are related by the Kramers-Kronig relations, Eq. (C.14), which in our case read

$$
\begin{aligned}
\varepsilon'(\omega) - \varepsilon_\infty &= \frac{2}{\pi} \int_0^\infty d\omega' \, \frac{\omega' \varepsilon''(\omega)}{\omega'^2 - \omega^2} \\
\varepsilon''(\omega) &= -\frac{2\omega}{\pi} \int_0^\infty d\omega' \, \frac{\varepsilon'(\omega)}{\omega'^2 - \omega^2}.
\end{aligned}
$$

(D.12)

It immediately follows that (for $\omega > 0$)

$$\varepsilon''(\omega) = \frac{4\pi^2 e^2 (Z^*)^2}{V_{\text{cell}} M} \delta(\omega_{\text{TO}}^2 - \omega^2) = \frac{2\pi^2 e^2 (Z^*)^2}{V_{\text{cell}} M \omega_{\text{TO}}} \delta(\omega_{\text{TO}} - \omega). \tag{D.13}$$

The meaning of this is pretty clear: we have an undamped oscillator at the frequency ω_{TO}, which only absorbs when the exciting field $\mathbf{E}$ is at resonance, and furthermore at resonance the response diverges. The coefficient of $\delta(\omega_{\text{TO}} - \omega)$ in Eq. (D.13) is the "oscillator strength" of the spectral line; this also explains why the Born effective charges are also called "infrared charges".

The simple example of a cubic binary crystal, having only a single dynamical variable $\mathbf{u}$, is pedagogically useful to address the response in time domain:

$$\mathbf{D}(t) = \int_{-\infty}^{\infty} dt' \, \varepsilon(t - t') \, \mathbf{E}(t'). \tag{D.14}$$

The response $\varepsilon(t)$ can be obviously found using Eqs. (D.11) and (D.13), and antitransforming $\varepsilon(\omega) = \varepsilon'(\omega) + i\varepsilon''(\omega)$. But we also remind that $\varepsilon(t)$ can be seen as the response to a δ-like input signal at $t = 0$, i.e. $\mathbf{E}(t) = \mathbf{E}_0 \, \delta(t)$. Using Eq. (D.2) we have:

$$\varepsilon(t) \, \mathbf{E}_0 = \mathbf{D}(t) = \varepsilon_\infty \mathbf{E}_0 \, \delta(t) + \frac{4\pi Z^* e}{V_{\text{cell}}} \mathbf{u}(t), \qquad t \geq 0, \tag{D.15}$$

where we consider the electronic response as instantaneous. The Newton equation for $\mathbf{u}(t)$, from Eq. (D.2), is

$$\ddot{\mathbf{u}}(t) = -\omega_{\text{TO}}^2 \mathbf{u}(t) + \frac{Z^* e}{M} \mathbf{E}_0 \, \delta(t). \tag{D.16}$$

Since we are looking for a causal response, $\mathbf{u}(t) = 0$ for $t < 0$, and the response to the instantaneous "kick" at $t = 0$ is the $t > 0$ solution of the homogenous equation with initial conditions:

$$\mathbf{u}(0) = 0; \qquad \dot{\mathbf{u}}(0) = \frac{Z^* e}{M} \mathbf{E}_0. \tag{D.17}$$

We thus get

$$\mathbf{u}(t) = \frac{Z^* e}{M \omega_{\text{TO}}} \mathbf{E}_0 \sin \omega_{\text{TO}} \, t. \tag{D.18}$$

Finally, the response function is, from Eq. (D.15):

$$\varepsilon(t) = 0, \quad t < 0; \qquad \varepsilon(t) = \varepsilon_\infty \, \delta(t) + \frac{(Z^*)^2 e^2}{V_{\text{cell}} M \omega_{\text{TO}}} \sin \omega_{\text{TO}} \, t, \quad t \geq 0, \tag{D.19}$$

where we retrieve the undamped oscillations of the responding system. More about infrared spectroscopy will be said below, Sec. D.10.

While $\varepsilon(\omega)$ is the response to an **E** field, its inverse $\varepsilon^{-1}(\omega)$ is the response to a **D** field. Even $\varepsilon^{-1}(\omega)$ must obey Kramers-Kronig relationships. We are not giving the details here, but the derivation is straightforward by exploiting the identity

$$\frac{\varepsilon'(\omega)}{\varepsilon_\infty} = \frac{\omega_{\mathrm{LO}}^2 - \omega^2}{\omega_{\mathrm{TO}}^2 - \omega^2}, \tag{D.20}$$

which clearly displays the symmetry. The real part of $\varepsilon(\omega)$ has a pole at ω_{TO}, and the corresponding imaginary part a delta peak at the same frequency; the real part of $\varepsilon^{-1}(\omega)$ has a pole at ω_{LO}, and the corresponding imaginary part a $\delta(\omega)$ peak at the same frequency.

D.3 Dissipation and lifetime

So far, our response function was causal but nondissipative. The single-mode prototypical example dealt with so far gives also the opportunity for introducing phenomenological dissipation (or equivalently lifetime) into our response function. We start addressing the issue in the time domain, and we include damping into Eq. (D.16):

$$\ddot{\mathbf{u}}(t) = -2\gamma\dot{\mathbf{u}}(t) - \omega_{\mathrm{TO}}^2\mathbf{u}(t) + \frac{Z^*e}{M}\mathbf{E}_0\,\delta(t). \tag{D.21}$$

The damping γ has the dimensions of an inverse time, and we have included a factor of 2 for convenience. We are interested in the case $\gamma \ll \omega_{\mathrm{TO}}$; the solution with the initial conditions of Eq. (D.17) is

$$\mathbf{u}(t) = \frac{Z^*e}{M\omega_1}\mathbf{E}_0\,\mathrm{e}^{-\gamma t}\,\sin\omega_1 t, \qquad \omega_1 = \sqrt{\omega_{\mathrm{TO}}^2 - \gamma^2}. \tag{D.22}$$

$$\varepsilon(t) = 0, \quad t < 0; \qquad \varepsilon(t) = \varepsilon_\infty\,\delta(t) + \frac{(Z^*)^2 e^2}{V_{\mathrm{cell}}M\omega_1}\,\mathrm{e}^{-\gamma t}\,\sin\omega_1 t, \quad t \geq 0, \tag{D.23}$$

The dielectric function in the ω-domain obtains by antitransforming this; γ is clearly the inverse lifetime.

It is simpler instead to address the forced oscillations at frequency ω. A dissipative term into Eq. (D.5) yields

$$-M\omega^2\mathbf{u} = 2iM\gamma\omega\mathbf{u} - M\omega_{\mathrm{TO}}^2\mathbf{u} + Z^*e\mathbf{E} \tag{D.24}$$

$$\mathbf{u} = \frac{Z^*e}{M(\omega_{\mathrm{TO}}^2 - 2i\gamma\omega - \omega^2)}\mathbf{E} \tag{D.25}$$

$$\varepsilon(\omega) = \varepsilon_\infty + \frac{4\pi e^2(Z^*)^2}{V_{\mathrm{cell}}M(\omega_{\mathrm{TO}}^2 - 2i\gamma\omega - \omega^2)}. \tag{D.26}$$

We notice that $\varepsilon(\omega)$ is now endowed with a real and an imaginary part, which obey Kramers-Kronig relationships by construction. In the complex ω plane $\varepsilon(\omega)$ is an analytic function, with poles at $\pm\omega_1 - i\gamma$. The previous expressions can be compared with the zero-T quantum response, and the location of its poles: see Eqs. (C.58) and (C.59) and the following discussion. In both cases the fact that the poles are in the lower half of the complex ω plane is a manifestation of causality, as previously emphasized in Sect. C.4 . Note that the location of the poles on the upper/lower half plane depends on our convention on Fourier transforms, Eq. (C.1); the opposite choice may be found in the literature.

Th most perspicuous difference between Eq. (D.26) and its quantum ($T = 0$) counterpart, Eqs. (C.58) and (C.59), is in the denominator: second order in ω vs. first order. This owes to the fact that the Newton equations of motion are second order in t, while Schrödinger equation is first order. If we are interested in the region around resonance we may replace

$$\omega_{\text{TO}}^2 - \omega^2 - 2i\gamma\omega \simeq 2\,\omega_{\text{TO}}\,(\omega_{\text{TO}} - \omega - i\gamma), \tag{D.27}$$

$$\varepsilon(\omega) \simeq \varepsilon_\infty + \frac{2\pi e^2 (Z^*)^2}{V_{\text{cell}} M \omega_{\text{TO}}} \frac{1}{\omega_{\text{TO}} - \omega - i\gamma}. \tag{D.28}$$

The non dissipative (although causal) response obtains by taking the $\gamma \to 0^+$ limit. Exploiting the well known identity

$$\lim_{\gamma \to 0^+} \frac{1}{x - i\gamma} = \mathcal{P}\frac{1}{x} + i\pi\delta(x) \tag{D.29}$$

we get $\varepsilon''(\omega)$ identical to Eq. (D.13). At finite positive γ the distribution $\delta(\omega_{\text{TO}} - \omega)$ is broadened into a normalized Lorentzian, and the oscillator strength of the spectral line is conserved. In general we have

$$\varepsilon'(\omega) \;\simeq\; \varepsilon_\infty + \frac{2\pi e^2 (Z^*)^2}{V_{\text{cell}} M \omega_{\text{TO}}} \frac{\omega_{\text{TO}} - \omega}{(\omega_{\text{TO}} - \omega)^2 + \gamma^2}, \tag{D.30}$$

$$\varepsilon''(\omega) \;\simeq\; \frac{2\pi e^2 (Z^*)^2}{V_{\text{cell}} M \omega_{\text{TO}}} \frac{\gamma}{(\omega_{\text{TO}} - \omega)^2 + \gamma^2}. \tag{D.31}$$

While $\varepsilon''(\omega)$ yields the absorptive part of the response, $\varepsilon'(\omega)$ yields the dispersive one. While crossing the resonance, $D(\omega)$ switches from in-phase to out-of-phase with $E(\omega)$.

D.4 Dynamical matrix

Let us consider a solid at zero temperature in the adiabatic approximation: the ionic positions at equilibrium are $\mathbf{R}_{ls} = \mathbf{R}_l + \mathbf{R}_s$, where l is a cell index and s is

(in nonprimitive lattices) a basis index in the unit cell of n atoms. A distorted configuration of the solid is described by the set of the ionic displacements $\{\mathbf{u}_{ls}\}$; as throughout the present Notes, Greek subscript are used for Cartesian components.

The total energy is expanded in the displacements around equilibrium as

$$E_{tot}(\{\mathbf{u}_{ls}\}) = E_{tot}^{(0)} + E_{tot}^{(2)}(\{\mathbf{u}_{ls}\}) + \dots \tag{D.32}$$

and the harmonic force constant are second derivatives, defined through

$$E_{tot}^{(2)}(\{\mathbf{u}_{ls}\}) = \frac{1}{2} \sum_{ll'ss',\alpha\beta} c_{ss',\alpha\beta}(l,l') u_{ls,\alpha} u_{l's',\beta}. \tag{D.33}$$

Because of lattice periodicity, the force constants depend only on $\mathbf{R}_{l'} - \mathbf{R}_l$ and hence $c_{ss',\alpha\beta}(0,l')$ contain all of the information. Besides this, there are further important constraints imposed by translational and rotational invariance [264]: we consider explicitly only the former. If the solid is translated as a whole, the energy is unchanged, and hence the energy expansion vanishes to all orders; this is easily shown to imply

$$c_{ss,\alpha\beta}(0,0) = -\sum_{l's'}{}' c_{ss',\alpha\beta}(0,l'), \tag{D.34}$$

with the usual meaning of the primed sum.

Owing to the adiabatic approximation [265, 266] the ionic motion is classical in the potential energy provided by E_{tot}. The harmonic oscillations around equilibrium are then governed by the equation of motion:

$$M_s \ddot{u}_{ls,\alpha} = -\sum_{l's',\beta} c_{ss',\alpha\beta}(l,l') u_{l's',\beta}, \tag{D.35}$$

for which normal-mode solutions (within periodic boundary conditions) are phonons of $\mathbf{q}$ wavevector:

$$u_{ls,\alpha} = u_{s,\alpha}(\mathbf{q}) e^{i\mathbf{q}\cdot(\mathbf{R}_l + \mathbf{R}_s)} e^{-i\omega(\mathbf{q})t}. \tag{D.36}$$

Owing to the transformation, the secular problem factorizes for different $\mathbf{q}$'s (chosen by convenience within the first Brillouin zone). Using then the reciprocal form of the force constants:

$$c_{ss',\alpha\beta}(\mathbf{q}) = e^{i\mathbf{q}\cdot(\mathbf{R}_{s'} - \mathbf{R}_s)} \sum_{l'} c_{ss',\alpha\beta}(0,l') e^{i\mathbf{q}\cdot\mathbf{R}_{l'}}, \tag{D.37}$$

and introducing the auxiliary quantities:

$$u_{s,\alpha}(\mathbf{q}) = \frac{1}{\sqrt{M_s}} e_{s,\alpha}(\mathbf{q}); \tag{D.38}$$

$$D_{ss',\alpha\beta}(\mathbf{q}) = (M_s M_{s'})^{-\frac{1}{2}} c_{ss',\alpha\beta}(\mathbf{q}), \tag{D.39}$$

where M_s are the ionic masses, the secular problem becomes:

$$\omega^2(\mathbf{q}) e_{s,\alpha}(\mathbf{q}) = \sum_{s',\beta} D_{ss',\alpha\beta}(\mathbf{q}) e_{s',\beta}(\mathbf{q}). \tag{D.40}$$

The dynamical matrix is Hermitian, and has (in stable systems) nonnegative eigenvalues: at every $\mathbf{q}$ there are $3n$ normal modes, whose frequencies are the square roots of the eigenvalues; the eigendisplacements are related to the eigenvectors by a trivial mass-dependent factor.

At this point I stress the importance—both in model and first-principle lattice dynamics—of a standard technical trick. It is convenient not to bother explicitly with the actual value of the on-site force constant $c_{ss,\alpha\beta}(0,0)$. To this effect one uses an auxiliary set of force constants, indicated with an overbar, which in general violate translational invariance and are therefore nonphysical. The physical force constants are recovered from

$$c_{ss',\alpha\beta}(\mathbf{q}) = \bar{c}_{ss',\alpha\beta}(\mathbf{q}) - \delta_{ss'} \sum_{s''} \bar{c}_{ss'',\alpha\beta}(0), \tag{D.41}$$

which ensures translational invariance in form. It is easy to verify that $\bar{c}_{ss,\alpha\beta}(0)$ cancels in this expression, and hence the value of the on-site force constant $c_{ss,\alpha\beta}(0,0)$ is irrelevant, whenever Eq. (D.41) is explicitly used.

D.5 Coulomb forces

We discuss in this section the contribution to lattice dynamics Coulomb interactions. This is an actual ingredient within the framework of the rigid-ion model, but is also used here for pedagogical purposes, in order to introduce nonanalytic features of the first-principle dynamical matrix in presence of Coulomb forces.

Suppose that the central two-body potential between ions of species s and s' is $\Phi_{ss'}(r)$; then its contribution to the lattice-dynamical force constants is

$$c_{ss',\alpha\beta}(l, l') = - \left. \frac{\partial^2 \Phi_{ss'}(|\mathbf{r}|)}{\partial r_\alpha \partial r_\beta} \right|_{\mathbf{r}=\mathbf{R}_{l'}+\mathbf{R}_{s'}-\mathbf{R}_l-\mathbf{R}_s}. \tag{D.42}$$

Explicit evaluation of the second derivatives of a radial function yields

$$\frac{\partial^2 \Phi_{ss'}(|\mathbf{r}|)}{\partial r_\alpha \partial r_\beta} = \frac{1}{r} \Phi'_{ss'}(r)(\delta_{\alpha,\beta} - \frac{r_\alpha r_\beta}{r^2}) + \Phi''_{ss'}(r)\frac{r_\alpha r_\beta}{r^2}, \tag{D.43}$$

and we only need to calculate $\Phi'(r)/r$ and $\Phi''(r)$ for each shell of neighbors. These quantities are often called "tangential" and "radial" force constants.

The above expressions only apply to $l \neq l'$ and $s \neq s'$; using an *arbitrary* value for the on-site force constant $c_{ss,\alpha\beta}(0,0)$, we get the analogous of Eq. (D.37) as

$$\bar{c}_{ss',\alpha\beta}(\mathbf{q}) = e^{i\mathbf{q}\cdot(\mathbf{R}_{s'}-\mathbf{R}_s)} \sum_{l'} c_{ss',\alpha\beta}(0,l') e^{i\mathbf{q}\cdot\mathbf{R}_{l'}}, \tag{D.44}$$

and the physical force constants are obtained from Eq. (D.41). The expression in Eq. (D.44) is usable as such only for short-range interactions, where the sum over l' is rapidly convergent; the resulting force constants are then analytic functions of $\mathbf{q}$.

It is now convenient to transform the force constants into an equivalent expression involving the Fourier transform $\Phi_{ss'}(k)$ of the pairwise interaction. Standard manipulations, starting from Eq. (D.44), give

$$\bar{c}_{ss',\alpha\beta}(\mathbf{q}) = \frac{1}{V_{\text{cell}}} \sum_{\mathbf{G}} (\mathbf{q}+\mathbf{G})_\alpha (\mathbf{q}+\mathbf{G})_\beta \Phi_{ss'}(|\mathbf{q}+\mathbf{G}|) e^{i\mathbf{G}\cdot(\mathbf{R}_{s'}-\mathbf{R}_s)}. \tag{D.45}$$

Specializing to a pure Coulomb interaction between point charges, its reciprocal space expression is $\Phi_{ss'}(k) = 4\pi Q_s Q_{s'}/k^2$. The force constants are nonanalytic, as a fingerprint of long-range Coulomb interactions: we cast them as

$$\bar{c}_{ss',\alpha\beta}^{(Coul)}(\mathbf{q}) = \frac{4\pi Q_s Q_{s'}}{V_{\text{cell}}} \left[\frac{q_\alpha q_\beta}{q^2} + \sum_{\mathbf{G}\neq 0} \frac{(\mathbf{q}+\mathbf{G})_\alpha (\mathbf{q}+\mathbf{G})_\beta}{|\mathbf{q}+\mathbf{G}|^2} e^{i\mathbf{G}\cdot(\mathbf{R}_{s'}-\mathbf{R}_s)} \right], \tag{D.46}$$

where the nonanalytic term has been expressly separated. This expression is valid at $\mathbf{q} \neq 0$: a major problem is the fact that $\bar{c}_{ss',\alpha\beta}(\mathbf{q}=0)$ does *not* cancel (for $s \neq s'$) in Eq. (D.41). What is the correct value to be used? There are several ways to reach a rigorous answer: the final result can however be summarized into a working recipe, first proposed by Pines in the 1950s. Under the hypothesis that the extended system is *overall* charge-neutral, one gets always correct results upon assumption that the Fourier transform of Coulomb interaction is $4\pi/k^2$ at $k \neq 0$ and vanishes at $k = 0$. Therefore the first term in square brackets must be taken as zero when $\mathbf{q}$ is exactly zero. When $\mathbf{q} \to 0$, *i.e.* in the long-wavelength limit, this term gives in general a finite contribution. In fact $q_\alpha q_\beta/q^2$ acts as a projector on the $\mathbf{q}$-direction. Owing to the presence of terms of this kind, the dynamical matrix of a polar dielectric is *non analytic* at the zone center.

In the simple case of a cubic binary crystal $q_\alpha q_\beta/q^2$ discriminates between longitudinal and transverse modes, and is therefore responsible for the LT splitting, already introduced within Huang's phenomenological theory.

Still, a minor problem remains: the reciprocal space sum over $\mathbf{G}$ in Eq. (D.46) does not converge. The solution is straightforward: one evaluates the Coulomb term giving a small Gaussian spread to the ionic charges, in order to ensure convergence. When the spread is much smaller than the typical interionic distances, the force

constants assume their physical (spread-independent) value. Alternatively, more efficient Ewald techniques can be used [264]. There is one simple case where the sum can be evaluated in closed form: let us consider $\bar{c}_{12,\alpha\beta}^{(Coul)}(0)$ in a cubic binary crystal. Because of symmetry, this must be diagonal in $\alpha\beta$ and equal to one third of the trace:

$$\bar{c}_{12,\alpha\beta}^{(Coul)}(0) = \delta_{\alpha\beta}\frac{4\pi Q_1 Q_2}{3V_{\text{cell}}}\sum_{\mathbf{G}\neq 0}e^{i\mathbf{G}\cdot(\mathbf{R}_2-\mathbf{R}_1)} . \tag{D.47}$$

Since we know that such a sum over all $\mathbf{G}$'s (including $\mathbf{G} = 0$) vanishes, we get the result

$$\bar{c}_{12,\alpha\beta}^{(Coul)}(0) = -\delta_{\alpha\beta}\frac{4\pi Q_1 Q_2}{3V_{\text{cell}}} . \tag{D.48}$$

D.6   A simple approach to alkali metals

Rather accurate phonon spectra in simple metals have been obtained during the 1960s from pseudopotential perturbation theory to second order. In a sense, this was the archetypical "total energy" approach to the electronic properties of solids. Although all this is a rather obsolete approach—superseded by first-principle calculations [134]—we include this Section for pedagogical purposes.

Within pseudopotential perturbation theory the total energy is the sum of two terms: the former depends only upon the average density, while the second is in fact a pairwise central interaction.

Phonon modes do not affect the average density, and therefore the full dynamical matrix is straightforwardly obtained from the formalism of the previous section, where the two body interaction is the screened interaction between the pseudoions. For a primitive lattice we omit the s subscripts; using Eqs. (D.41) and (D.45) we get the dynamical matrix in the form:

$$D_{\alpha\beta}(\mathbf{q}) = \frac{1}{MV_{\text{cell}}}\sum_{\mathbf{G}}[(\mathbf{q}+\mathbf{G})_\alpha(\mathbf{q}+\mathbf{G})_\beta\Phi(|\mathbf{q}+\mathbf{G}|) - G_\alpha G_\beta\Phi(G)]. \tag{D.49}$$

The crudest approximation for the two-body interaction obtains from the electron-gas dielectric function as

$$\Phi(k) \simeq \frac{4\pi Z^2 e^2}{\varepsilon(k)k^2}, \tag{D.50}$$

with $Z = 1$ for monovalent methods. This form for the dynamical matrix of a simple metal was first proposed in 1958 by Toya [267], and clearly approximates the pseudopotential with the bare Coulomb potential of the ion core, screened by the electron gas. It neglects the short-range repulsion which basically mimics orthonormalisation of the valence orbitals to the core ones.

Over the years, many forms have been proposed for the electron-gas dielectric function $\varepsilon(k)$. All of them are divergent for $k \to 0$, thus making the two-body interaction and the dynamical matrix analytic. The simplest of them is the Thomas-Fermi dielectric function yielding the screened interaction

$$\Phi(k) \simeq \frac{4\pi e^2}{k^2 + k_{\mathrm{TF}}^2}, \tag{D.51}$$

where k_{TF} is given in many textbooks [73]. If we single out the $\mathbf{G} = 0$ term in Eq. (D.49) we have

$$\begin{aligned} D_{\alpha\beta}(\mathbf{q}) &= \frac{4\pi e^2 q_\alpha q_\beta}{MV_{\mathrm{cell}}(q^2 + k_{\mathrm{TF}}^2)} \\ &+ \frac{1}{MV_{\mathrm{cell}}} {\sum_{\mathbf{G} \neq 0}}' [(\mathbf{q} + \mathbf{G})_\alpha (\mathbf{q} + \mathbf{G})_\beta \Phi(|\mathbf{q} + \mathbf{G}|) - G_\alpha G_\beta \Phi(G)]. \end{aligned} \tag{D.52}$$

If one further neglects the second line, i.e. neglects the discrete nature of the lattice, one gets the speed of sound as ω_p/k_{TF}, where ω_p is the electron-gas plasma frequency. We thus recover an important result found in 1950 upon macroscopic arguments by Bohm and Staver and derived in several textbooks [73]. The result is in semiquantitative agreement with the experiment; this proved for the first time that sound propagation in metals is dominated by electronic screening therein.

A much more accurate expression for the two-body interaction is provided by pseudopotential perturbation theory [268]. Without giving a derivation, we only show here the main result. One starts defining defining the dimensionless function:

$$\mathcal{G}(k) = \left(\frac{4\pi Ze^2}{k^2}\right)^{-2} v^2(k) \left[1 - \frac{1}{\varepsilon(k)}\right], \tag{D.53}$$

known as the "energy-wave-number characteristic", where $v(k)$ is the (local) pseudopotential. This accounts for the interaction between the ion cores, mediated by the electrons. When we add the bare repulsion between the ionic charges we get the two-body interaction as

$$\Phi(k) = \frac{4\pi Z^2 e^2}{k^2} \left[1 - \mathcal{G}(k)\right], \tag{D.54}$$

to be used in Eq. (D.49). As said above, this approach has provided in the 1960s remarkably accurate phonon spectra for the simple metals. The crude approximation of Eq. (D.50) is retrieved replacing $v(k)$ in Eq. (D.53) with the Coulomb interaction.

A more accurate expression for the electron-gas dielectric function $\varepsilon(k)$ has a logarithmic singularity for $k=2k_F$. The simplest $\varepsilon(k)$ displaying this feature is the

Lindhard dielectric function, discussed in many textbooks [73]. The singularity is due to the sharpness of the Fermi surface in metals: notice that even at room temperature the Fermi surface can be considered as sharp to the present purposes.

Therefore the dynamical matrix of Eq. (D.49) is singular at the $\mathbf{q}$ vectors (in the first Brillouin zone) which fullfill $|\mathbf{q} + \mathbf{G}| = 2k_F$ for some $\mathbf{G}$. The surfaces defined by this relationship can be found from a simple geometrical construction: when $\mathbf{q}$ crosses one of these surfaces, one expects a "kink" in the experimental phonon dispersion curves, as first proposed by W. Kohn [269]. Shortly after the theory, the effect was observed in the spectra of lead [270], and goes under the name of "Kohn anomaly" since then.

D.7 Ionic crystals: rigid-ion model

The lattice dynamics of ionic cristals has important qualitative features which are better illustrated—for pedagogical purposes—starting from the simple rigid-ion model, which was first applied to the lattice dynamics of alkali halides in 1940 [271]. The force model adopted here is exactly the same as used in the Kittel textbook [97] to discuss the structural stability of alkali halides: short-range Born-Mayer forces (which only act between first neighbours), and point-charge Coulomb forces. The parameters in the model are the ionic charges $\pm Q$ (where usually $|Q| = e$, i.e. full ionicity), plus other two parameters in the Born-Mayer forces. The latter are fitted to a pair of empirical data (typically the equilibrium lattice constant and the bulk modulus) [97]. One would naively guess that the Born-Mayer forces are entirely responsible for ω_{TO}, and the Coulomb forces are responsible for the longitudinal-transverse splitting: this is incorrect, as we will see below.

The dynamical matrix is therefore the sum of two contributions, which are separately evaluated. The Coulomb term is given by Eq. (D.46), using for the ionic charges the values $Q_s = (-1)^s Q$, where s=1 (2) labels the anion (cation); the short range term is evaluated directly in the form of Eq. (D.44), where the sum includes only nearest neighbors (six terms for the rocksalt structure). Therefore the short-range $\bar{c}_{ss',\alpha\beta}^{(s.r.)}(\mathbf{q})$ vanishes when s=s', while $\bar{c}_{12,\alpha\beta}^{(s.r.)}(\mathbf{q})$ depends on two parameters only, given in Eq. (D.43). A typical result, taken from the original literature, is shown in Fig. D.1.

It proves useful to study in detail the diagonalization at the zone center. Keeping only terms of order zero in $\mathbf{q}$, and using Eq. (D.41), we get:

$$c_{11,\alpha\beta}^{(s.r.)}(\mathbf{q}) \simeq c_{22,\alpha\beta}^{(s.r.)}(\mathbf{q}) \simeq -\bar{c}_{12,\alpha\beta}^{(s.r.)}(0) \ ; \ \ c_{12,\alpha\beta}^{(s.r.)}(\mathbf{q}) \simeq \bar{c}_{12,\alpha\beta}^{(s.r.)}(0); \qquad (D.55)$$

furthermore bulk Cartesian tensors are diagonal in a cubic material, hence $\bar{c}_{12,\alpha\beta}^{(s.r.)}(0) = -R_0\delta_{\alpha\beta}$, where the constant R_0 completely accounts for the short range

FIGURE 2. **p**: (0.0.0) to (5.5.5)
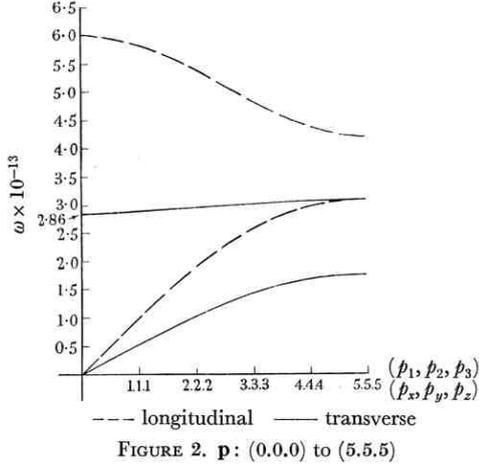
Figure D.1: The phonon spectrum of NaCl for **q** along the (1,1,1) direction, as computed in 1940 in Ref. [271].

forces at the zone center. Using the same path for the Coulomb term, Eq. (D.46), we find a nonanalytic term (homogeneous of degree zero in **q**), plus an analytic term which behaves in all respects like an effectively short-range additional interaction. From Eqs. (D.46) and (D.48) we cast the complete dynamical matrix at the zone center as

$$D_{ss',\alpha\beta}(\mathbf{q}) \simeq \left[\frac{4\pi Q^2}{V_{\text{cell}}} \frac{q_\alpha q_\beta}{q^2} + (R_0 - \frac{4\pi Q^2}{3V_{\text{cell}}})\delta_{\alpha\beta} \right] \mathcal{M}_{ss'}, \qquad (D.56)$$

where we define the 2×2 matrix $\mathcal{M}$ as :

$$\mathcal{M} = \begin{pmatrix} \frac{1}{M_1} & -\frac{1}{\sqrt{M_1 M_2}} \\ -\frac{1}{\sqrt{M_1 M_2}} & \frac{1}{M_2} \end{pmatrix}. \qquad (D.57)$$

One of the eigenvalues of $\mathcal{M}$ is zero, and the other is $1/M = 1/M_1 + 1/M_2$, *i.e.* the inverse reduced mass; the corresponding eigenvectors are easily recognized to be the acoustic and optic modes, respectively. Going back to the zone-center dynamical matrix, the acoustic mode is threefold degenerate with frequency zero, while the optic modes require further diagonalization over the Cartesian coordinates. There are two transverse modes—where the polarization is perpendicular to **q**—with

$$\omega_{TO}^2 = \frac{1}{M}(R_0 - \frac{4\pi Q^2}{3V_{\text{cell}}}), \qquad (D.58)$$

and one longitudinal mode—where the polarization is parallel to **q**—whose frequency is

$$\omega_{LO}^2 = \omega_{TO}^2 + \frac{4\pi Q^2}{MV_{\text{cell}}}. \qquad (D.59)$$

The LT splitting is substantially overestimated with respect to the experiment, and this drawback is simply due to the fact that no mechanism within the rigid-ion

201

model allows for electronic screening of the Coulomb interaction. In fact comparing Eq. (D.59) to the *exact* expression in Eq. (D.8), we easily realise that the rigid-ion model sets $Z^* = 1$ and $\varepsilon_\infty = 1$. While the former is a reasonable approximation to the real Z^* in ionic crystals, the latter is not: $\varepsilon_\infty \simeq 2 - 3$ in alkali halides.

Models correcting this problem are well known in the literature: they agree better with the experimental spectra, at the price of using more empirical parameters. A comprehensive atlas of phonon spectra, as computed by means of empirical models for many materials, was published in 1979 [272]. Nowadays such models have historical interest only, since phonon spectra are routinely computed by first principles, even for very complex crystal structures [104, 105, 106].

One final lesson can be drawn from the rigid-ion model, looking at Eq. (D.58), where it is perspicuous that ω_{TO} is not due to the Born-Mayer forces (summarized in R_0) only: it also has a sizeable contribution from the Coulomb forces. This was to be expected, since the Coulomb forces are essential to crystal stability as well. We notice that ω_{TO}^2 is the eigenvalue of the *analytical* part of the dynamical matrix: Born-Mayer forces provide obviously an analytic term, but there is another analytic term of Coulomb origin, coming from the $\mathbf{G} \neq 0$ terms in Eq. (D.46).

In a first-principle approach there is no way of disentangling the "short-range" from the "Coulomb" forces: the interatomic forces are due to quantum mechanics and electrostatics intertwined. We are only allowed to partition into analytic and nonanalytic contributions. The key ingredient to the latter is $q_\alpha q_\beta / q^2$, i.e. an homogeneous function of degree zero, whose value for $\mathbf{q} \to 0$ depends on the $\mathbf{q}$ direction.

D.8   Zone-center modes: general case

Exploiting the results in Secs. D.1 and D.7, the *exact* zone-center dynamical matrix of a cubic binary crystal is

$$D_{ss',\alpha\beta}(\mathbf{q}) = \left[M\omega_{\mathrm{TO}}^2 \delta_{\alpha\beta} + \frac{4\pi e^2 (Z^*)^2}{\varepsilon_\infty V_{\mathrm{cell}}} \frac{q_\alpha q_\beta}{q^2} \right] \mathcal{M}_{ss'}, \qquad (\mathrm{D.60})$$

where the 2×2 matrix $\mathcal{M}$ is given in Eq. (D.57), and M is the reduced mass. The nonanalytic expression in Eq. (D.60) is correct to order zero in $\mathbf{q}$. In order to proceed further, it is expedient to define the Born charges of each sublattice Z_s^*, with $Z_1^* = -Z_2^*$, and rewrite Eq. (D.60) as

$$D_{ss',\alpha\beta}(\mathbf{q}) = \frac{1}{\sqrt{M_s M_{s'}}} \left[C_{ss',\alpha\beta}^{(\mathrm{analytic})} + \frac{4\pi e^2 Z_s^* Z_{s'}^*}{\varepsilon_\infty V_{\mathrm{cell}}} \frac{q_\alpha q_\beta}{q^2} \right], \qquad (\mathrm{D.61})$$

where—owing to cubic symmetry—$C_{ss',\alpha\beta}^{(\mathrm{analytic})}$ is diagonal in its Cartesian indices.
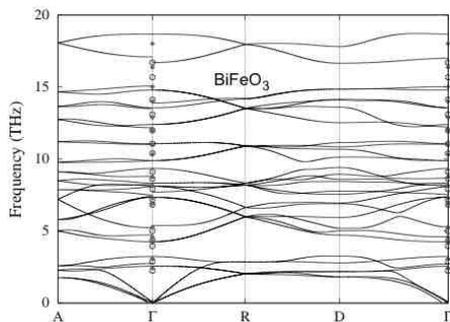
Figure D.2: A typical phonon spectrum of a low-symmetry dielectric (crystalline $BiFeO_3$). Owing to Coulomb interactions, the dynamical matrix is nonanalytic at the zone center. The $\mathbf{q} \to 0$ limit along different $\mathbf{q}$ directions may take different values.

Several first-principle electronic structure codes [104, 105, 106] routinely compute $C_{ss',\alpha\beta}^{(\text{analytic})}$, Z_s^*, and ε_∞, as well as their generalisations, discussed below. We notice that Z_s^* and ε_∞ require ad-hoc linear-response algorithms, while $C_{ss',\alpha\beta}^{(\text{analytic})}$ could even be computed in principle via "frozen phonons", i.e. evaluating the ground-state energy of the crystal for non equilibrium geometries, and expanding it to second order. We remind that standard ground-state calculations implicitly assume a lattice-periodical Hamiltonian, hence $\mathbf{E} = 0$. Any lattice-periodical calculation is unable to provide the longitudinal phonon frequency.

In a generic crystal with a basis of n atoms Eq. (D.61) generalises to

$$D_{ss',\alpha\beta} = \frac{1}{\sqrt{M_s M_{s'}}} \left(C_{ss',\alpha\beta}^{(\text{analytic})} + C_{ss',\alpha\beta}^{(\text{nonanalytic})}(\mathbf{q}) \right), \qquad (\text{D.62})$$

where $C_{ss',\alpha\beta}^{(\text{nonanalytic})}(\mathbf{q})$ is a nonanalytic function of degree zero in $\mathbf{q}$.

In a low-symmetry crystal the phonon spectrum is more complex: a typical example is shown in Fig. D.2, illustrating a novel qualitative feature. We have seen that for a cubic binary crystal the nonanalytic term in Eqs. (D.60) and (D.61) is responsible for the LT splitting, but the $\mathbf{q} \to 0$ limit is independent of the $\mathbf{q}$ directions, even for the longitudinal mode. In a low-symmetry crystal the zone-center modes are neither longitudinal nor transverse, and the $\mathbf{q} \to 0$ limit along different $\mathbf{q}$ directions may take different values.

D.8.1 Free energy and equations of motion

In order to obtain the explicit expression for $C_{ss',\alpha\beta}^{(\text{nonanalytic})}(\mathbf{q})$ in a generic insulating crystal we need to generalize our starting Eq. (D.1) to an arbitrary lattice. We remind that a zone-center mode is lattice-periodical, hence independent variables in the free energy per cell $\mathcal{F}$ are now the field $\mathbf{E}$ and the sublattice displacements $\mathbf{u}_s$, with $s = 1, n$. In a ferroelectric crystal the free energy includes a term *linear* in $\mathbf{E}$, and $\mathbf{u}_s$-independent.

We limit the present treatment to crystals which are neither pyroelectric nor ferroelectric. The most general second order expansion of $\mathcal{F}(E, \{\mathbf{u}_s\})$ reads then

$$
\begin{aligned}
\mathcal{F}(E, \{\mathbf{u}_s\}) &= \mathcal{F}_0 + \frac{1}{2} \sum_{ss'} C^{(\text{analytic})}_{ss',\alpha\beta} u_{s,\alpha} u_{s',\beta} - \frac{V_{\text{cell}}}{8\pi} \varepsilon_{\infty\alpha\beta} E_\alpha E_\beta \\
&\quad - e \sum_s Z^*_{s,\beta\alpha} u_{s,\alpha} E_\beta .,
\end{aligned} \tag{D.63}
$$

where the sum over Cartesian (Greek) indices is left implicit. Therein $C^{(\text{analytic})}_{ss',\alpha\beta}$ are the standard force constants, computed at zero $\mathbf{E}$ field, $\varepsilon_{\infty\alpha\beta}$ is the electronic dielectric tensor, and $Z^*_{s,\alpha\beta}$ is the dimensionless Born (a.k.a. infrared, a.k.a. dynamical) effective charge tensor of sublattice s. It is defined as the mixed second derivative

$$
e Z^*_{s,\beta\alpha} = -\frac{\partial^2 \mathcal{F}}{\partial u_{s,\alpha} \partial E_\beta}, \quad e Z^*_{s,\alpha\beta} = -\frac{\partial^2 \mathcal{F}}{\partial E_\alpha \partial u_{s,\beta}}. \tag{D.64}
$$

Notice that, while $C^{(\text{analytic})}_{ss',\alpha\beta}$ and $\varepsilon_{\infty\alpha\beta}$ are symmetric tensors in their Cartesian indices, the Born charge $Z^*_{s,\alpha\beta}$ in general is *not* a symmetric Cartesian tensor. The Born effective charge tensors must fulfill the acoustic sum rule [273]:

$$
\sum_s Z^*_{s,\alpha\beta} = 0. \tag{D.65}
$$

If the sum of the tensors does not vanish identically, some acoustic phonons do not have a vanishing frequency at $\mathbf{q} = 0$.

The equations of motion are:

$$
\begin{aligned}
f_{s,\alpha} &= -\frac{\partial \mathcal{F}}{\partial u_{s,\alpha}} = -\sum_{s'} C^{(\text{analytic})}_{ss',\alpha\beta} u_{s',\beta} + e Z^*_{s,\beta\alpha} E_\beta \\
D_\alpha &= -\frac{4\pi}{V_{\text{cell}}} \frac{\partial \mathcal{F}}{\partial E_\alpha} = \frac{4\pi e}{V_{\text{cell}}} \sum_s Z^*_{s,\alpha\beta} u_{s,\beta} + \varepsilon_{\infty\alpha\beta} E_\beta.
\end{aligned} \tag{D.66}
$$

We remind once more that all the coefficients occurring in Eqs. (D.63) and (D.66) are routinely computed in some electronic-structure codes [104, 105, 106].

D.8.2 Microscopic meaning of the Born charges

We have said above that the s-th Born effective charge tensor measures the macroscopic polarization linearly induced by a unit displacement of the s-th sublattice in zero $\mathbf{E}$ field, or equivalently the force on the s-th atom induced by a unit macroscopic $\mathbf{E}$ field at zero displacement. We adopt the latter viewpoint in

the following, and we notice that the force on the s-th atom is also equal to the *microscopic* field $\mathbf{E}^{(\mathrm{micro})}(\mathbf{r})$ at the nuclear site. Therefore

$$f_{s,\alpha} = e Z^*_{s,\beta\alpha} E_\beta = e Z_s E^{(\mathrm{micro})}_\alpha(\mathbf{R}_s),\qquad\text{(D.67)}$$

where eZ_s is the bare nuclear charge. This applies in an all-electron picture; the force has a different expression in a pseudopotential framework.

The dimensionless s-th Born charge tensor measures therefore the ratio of the microscopic field at site s to the macroscopic field. More precisely

$$Z^*_{s,\beta\alpha} = Z_s\,\frac{\partial E^{(\mathrm{micro})}_\alpha(\mathbf{R}_s)}{\partial E_\beta}.\qquad\text{(D.68)}$$

Since the acoustic sum rule [273] requires $\sum_s Z^*_{s,\beta\alpha} = 0$, the microscopic field must be—in simple cases where the tensors are diagonal—parallel and antiparallel to the macroscopic field on different sites.

All experiments (either anelastic neutron scattering or infrared spectroscopy) measure quantities which are *quadratic* in the Born tensors: therefore the sign of Z^*_s cannot be experimentally determined. Quantum-mechanical calculations of the response functions, instead, do determine the sign. The first pioneering calculations appeared in the early 1980s [274, 275]; nowadays computation of the Born charge tensors is a standard feature of many electronic structure codes [104, 105, 108, 106]. Data are available for many complex materials, even noncrystalline.

Based on a simple rigid-ion viewpoint, one would expect the diagonal elements of Z^*_s to be negative on anions and positive on cations, or equivalently that the force on the nuclei is parallel to the macroscopic field on cations, and antiparallel on anions. This is indeed what actual computations confirm, for normal materials at least; some materials with counterintuitive Z^*_s tensors are discussed in the following. The counterintuitive features, where present, are typically due to covalency and/or correlation.

The reasons why the Born charges are trivial or non trivial may be examined in terms of Wannier functions: in fact the modern theory of polarization, in one of its formulations, states that the polarization *difference* induced by a zone-center phonon can be simply expressed in terms of the displacements of the bare nuclear charges and of the Wannier-function centers (see Sec. 5.3.3). Let us start with the simple case of alkali halides, where the computed Born charges are close to their intuitive value of ± 1. In the paradigmatic example of NaCl the nuclear charges are 11 and 17, respectively, and in the unperturbed solid at equilibrium there are 5 and 9 doubly occupied maximally localized Wannier functions centered at the two sites, respectively. When the two sublattices are displaced, the Wannier centers are displaced as well. The modern theory of polarization tells that Z^*_s would be exactly ± 1 if the Wannier centers follow rigidly the nuclear motion. What actually
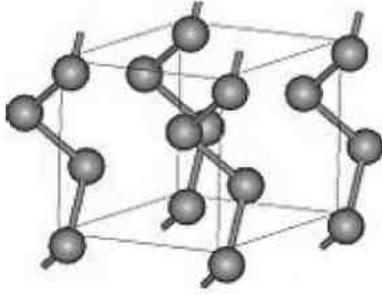
Figure D.3: The crystal structure of selenium. There are three equivalent atoms per cell, arranged on a spiral. The bonding is covalent within the spiral (coordination two), but there are weaker, "almost covalent", inter-spiral bonds.

happens is that, in alkali halides, the Wannier-center displacements are almost rigid. Notice that we are *not* assuming that the static charge, belonging to each atom, is ± 1: the static charge is an approximate concept, ill defined as a matter of principle [83, 84]. The Born (a.k.a. dynamical) charge, instead, is well defined both from first principles and experimentally. So much for alkali halides; at variance with them, in materials where covalency plays an important role the Wannier-center displacements are remarkably nonrigid.

In elemental crystals with a binary lattice, such as those having the diamond structure, the Born tensors vanish and there are no infrared active modes. This is due to symmetry and to the acoustic sum rule. However, elemental crystals of lower symmetry in general *do have* infrared active modes. One known case is solid hydrogen (a molecular crystal with several molecules in the elementary cell).

Another interesting case is Se: its crystal structure, with three equivalent atoms in the elementary cell, is shown in Fig. D.3. Owing to the acoustic sum rule $\sum_s Z^*_{s,\beta\alpha} = 0$, the $Z^*_{s,\beta\alpha}$ are purely off-diagonal, and they are equivalent by rotations of $2\pi/3$. This means that a macroscopic field exerts on an atom a force which is *orthogonal* to the field direction. The infrared activity can be qualitatively understood as an effect of the bond charges belonging to the weak interchain bonds

The next case that we wish to outline is the paradigmatic example of the perovskite $BaTiO_3$ in its undistorted cubic structure (nonferroelectric, a.k.a.
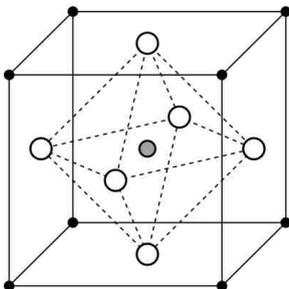

Figure D.4: The undistorted cubic perovskite structure (aristotype) of $BaTiO_3$. Solid, shaded, and empty circles represent Ba, Ti, and O atoms, respectively. The O atoms sit at orthorhombic, and *not* cubic, sites. The oxygen Born tensors are diagonal but anisotropic.

206
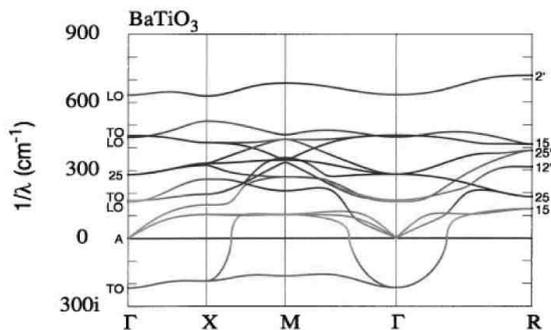
Figure D.5: The phonon spectrum of BaTiO$_3$ in its cubic structure (aristotype). Notice the giant LO-TO splitting, and also the occurrence of imaginary modes. After Ref. [276].

aristotype), where the oxygen ions sit at a noncubic sites. The Cartesian effective mass tensor Z_O^* is diagonal, but strongly anysotropic. Based on a rigid-ion picture, one would expect the elements of Z_O^* to be about -2; instead, one of them is of the order of -6. The tensor Z_{Ba}^* is isotropic, but is about $+6$, much larger than the nominal Ba ionicity $+4$. The effect is due to the mixed ionic-covalent character of this material, and of many perovskites as well. As a consequence, the LO-TO splitting at the zone center is giant [277, 276].

The phonon spectrum of BaTiO$_3$ in its cubic (aristotype) structure, computed from first principles by Ghosez et al. [276], is shown in Fig. D.5. Besides the giant LO-TO splitting, another prominent feature is worth a comment. Some phonon frequencies are *imaginary*, i.e. the dynamical matrix has negative eigenvalues. We have already observed that this material undergoes a spontaneous symmetry breaking, and therefore the cubic structure is not the relaxed equilibrium geometry. In other words the occurrence of imaginary modes is a manifestation of the fact that our harmonic expansion, Eqs. (D.33) and (D.63), is in this case performed at a saddle point and not at a minimum.

Finally, we wish to discuss the hypothetical case of an highly correlated alkali halide. It might happen that the electron-electron repulsion dominates over the standard behaviour discussed above. In this highly correlated regime the anion would have a positive Z_s^*, and the cation a negative one. This feature has been discovered via simulations on a two-band Hubbard model in one dimension [186].

D.8.3 Cochran-Cowley formula

We find expedient from now on to switch to compact formulæ adopting tensorial-vectorial notations, leaving the Cartesian indices implicit, and *without a change of symbols*; the dagger indicates the transpose. We thus equivalently cast Eq. (D.66) as

$$\mathbf{f}_s = -\sum_{s'} C_{ss'}^{(\text{analytic})} \mathbf{u}_{s'} + e\, Z_s^{*\dagger} \mathbf{E} \tag{D.69}$$

$$\mathbf{D} = \frac{4\pi e}{V_{\text{cell}}} \sum_s Z_s^* \, \mathbf{u}_s + \varepsilon_\infty \mathbf{E}. \tag{D.70}$$

We have learned from the paradigmatic example of the cubic binary crystal that the field $\mathbf{E}$ must depend on the relative orientation between the wavevector $\mathbf{q}$ and the sublattice displacements $\mathbf{u}_s(\mathbf{q})$. We make this explicit by rewriting Eqs. (D.69) and (D.70) as

$$\mathbf{f}_s(\mathbf{q}) = -\sum_{s'} C_{ss'}^{\text{(analytic)}} \mathbf{u}_{s'}(\mathbf{q}) + e\, Z_s^{*\dagger} \mathbf{E}(\mathbf{q}) \tag{D.71}$$

$$\mathbf{D}(\mathbf{q}) = \frac{4\pi e}{V_{\text{cell}}} \sum_s Z_s^* \mathbf{u}_s(\mathbf{q}) + \varepsilon_\infty \mathbf{E}(\mathbf{q}), \tag{D.72}$$

and we remind that we remain at the lowest order: the dependence is nonanalytic and of order zero in $\mathbf{q}$. Eq. (D.72) tells us that the polarization in zero field is

$$\mathbf{P}_0(\mathbf{q}) = \frac{e}{V_{\text{cell}}} \sum_s Z_s^* \mathbf{u}_s(\mathbf{q}). \tag{D.73}$$

Next, it is expedient to define the unit vectors in the $\mathbf{q}$-direction as

$$\hat{\mathbf{q}} = \begin{pmatrix} \hat{q}_x \\ \hat{q}_y \\ \hat{q}_z \end{pmatrix}, \qquad \hat{\mathbf{q}}^\dagger = \begin{pmatrix} \hat{q}_x & \hat{q}_y & \hat{q}_z \end{pmatrix}. \tag{D.74}$$

Therefore the norm is $\hat{\mathbf{q}}^\dagger \hat{\mathbf{q}} = 1$, while the dyadic product $\hat{\mathbf{q}}\,\hat{\mathbf{q}}^\dagger$ is the projector in the direction of $\mathbf{q}$, whose components are $q_\alpha q_\beta / q^2$

We have already observed that in presence of a phonon of wavevector $\mathbf{q}$, the solid is macroscopically homogeneous in the plane normal to $\mathbf{q}$, while all macroscopic properties display a modulation in the direction of $\mathbf{q}$. It is immediate to realize that the component of $\mathbf{D}(\mathbf{q})$ parallel to $\mathbf{q}$ and the component of $\mathbf{E}(\mathbf{q})$ normal to $\mathbf{q}$ both vanish [95]:

$$\hat{\mathbf{q}}^\dagger \mathbf{D}(\mathbf{q}) = 0, \qquad (1 - \hat{\mathbf{q}}\,\hat{\mathbf{q}}^\dagger)\mathbf{E}(\mathbf{q}) = 0. \tag{D.75}$$

Whenever nonvanishing, both $\mathbf{D}(\mathbf{q})$ and $\mathbf{E}(\mathbf{q})$ are nonanalytic functions of order zero in $\mathbf{q}$. In the following equations, it is tacitly assumed that only the leading term in $\mathbf{q}$ is considered. Whenever convenient, we may therefore replace $\mathbf{D}(\mathbf{q})$ and $\mathbf{E}(\mathbf{q})$ with $\mathbf{D}(\hat{\mathbf{q}})$ and $\mathbf{E}(\hat{\mathbf{q}})$, respectively. Eqs. (D.72) and (D.73) are clearly equivalent to

$$\mathbf{D}(\mathbf{q}) = \epsilon_\infty \mathbf{E}(\mathbf{q}) + 4\pi \mathbf{P}_0(\mathbf{q}). \tag{D.76}$$

We now exploit Eqs. (D.75) and (D.76) as follows:

$$0 = \mathbf{q}^\dagger \mathbf{D}(\mathbf{q}) = \mathbf{q}^\dagger \varepsilon_\infty \mathbf{E}(\mathbf{q}) + 4\pi \mathbf{q}^\dagger \mathbf{P}_0(\mathbf{q}) \tag{D.77}$$

$$\mathbf{E}(\mathbf{q}) = \hat{\mathbf{q}}\,\hat{\mathbf{q}}^\dagger \mathbf{E}(\mathbf{q}). \tag{D.78}$$

From these it easily follows that

$$0 = \hat{\mathbf{q}}^\dagger \varepsilon_\infty \hat{\mathbf{q}}\,\hat{\mathbf{q}}^\dagger \mathbf{E}(\mathbf{q}) + 4\pi \hat{\mathbf{q}}^\dagger \mathbf{P}_0(\mathbf{q}) \tag{D.79}$$

$$\mathbf{E}(\mathbf{q}) = -\frac{4\pi}{\hat{\mathbf{q}}^\dagger \varepsilon_\infty \hat{\mathbf{q}}}\hat{\mathbf{q}}\,\hat{\mathbf{q}}^\dagger \mathbf{P}_0(\mathbf{q}),$$

$$= -\frac{4\pi e}{V_{\text{cell}}\hat{\mathbf{q}}^\dagger \varepsilon_\infty \hat{\mathbf{q}}}\hat{\mathbf{q}}\sum_s \hat{\mathbf{q}}^\dagger Z_s^* \mathbf{u}_s(\mathbf{q}) \tag{D.80}$$

which can be interpreted as the depolarization field for an arbitrary $\hat{\mathbf{q}}$-direction. We can finally eliminate the field in the equations of motion, Eq. (D.71), which reads

$$\mathbf{f}_s(\mathbf{q}) = -\sum_{s'} \left[C_{ss'}^{(\text{analytic})} + \frac{4\pi e^2}{V_{\text{cell}}} \frac{(Z_s^{*\dagger}\hat{\mathbf{q}})(\hat{\mathbf{q}}^\dagger Z_{s'}^*)}{\hat{\mathbf{q}}^\dagger \varepsilon_\infty \hat{\mathbf{q}}} \right] \mathbf{u}_{s'}(\mathbf{q}). \tag{D.81}$$

The quantity in parenthesis is indeed the usual expression for the force-constant matrix at the zone center, including the nonanalytic term, first obtained in 1962 by Cochran and Cowley [278] and implemented much later in some first-principle codes [279, 134, 106, 104]. This confirms that the matrix elements at the zone center are indeed nonanalytic functions, homogeneous of degree zero in $\mathbf{q}$; we also remind that Eq. (D.81) applies to crystals of any symmetry and is *exact* at the harmonic level. The simple expression previously found in Eq. (D.61) is clearly a special case of Eq. (D.81).

When we restore the Cartesian subscripts the nonanalytic contribution to the dynamical matrix in Eq. (D.62) takes the more familiar (and prolix) form

$$C_{ss',\alpha\beta}^{(\text{nonanalytic})}(\mathbf{q}) = \frac{4\pi e^2}{V_{\text{cell}}} \frac{\sum_\gamma Z_{s,\gamma\alpha}^* q_\gamma \sum_\nu Z_{s',\nu\beta}^* q_\nu}{\sum_{\gamma\nu} q_\nu \varepsilon_{\infty\nu\gamma} q_\gamma}, \tag{D.82}$$

where even the sums on dummy indices are explicitated.

To inspect where nonanalyticity enters the dynamical matrix, we exploit associativity in the matrix products, and we write

$$C_{ss'}^{(\text{nonanalytic})}(\mathbf{q}) = \frac{4\pi e^2}{V_{\text{cell}}(\hat{\mathbf{q}}^\dagger \varepsilon_\infty \hat{\mathbf{q}})} Z_s^{*\dagger}(\hat{\mathbf{q}}\hat{\mathbf{q}}^\dagger) Z_{s'}^*. \tag{D.83}$$

The scalar $\hat{\mathbf{q}}^\dagger \varepsilon_\infty \hat{\mathbf{q}}$ is clearly analytic, hence the only source of nonanalyticity is the projector in the $\mathbf{q}$-direction $\hat{\mathbf{q}}\hat{\mathbf{q}}^\dagger$, discussed above, and whose elements are $q_\alpha q_\beta/q^2$.

Finally we observe that in a crystal of arbitrarily low symmetry all zone-center phonons are coupled to the field. Therefore a zero-field calculation of the force constants, as routinely provided by the popular computer codes [106, 104] does not provide by itself the frequency (and the eigenvectors) of any zone-center mode. It only provides $C_{ss'}^{(\text{analytic})}$, which must be used into the full Cochran-Cowley formula in order to get all zone-center modes.

D.9 High-symmetry cases

We consider here only crystals whose symmetry is orthorombic or higher; then all crystalline tensors can be simultaneously diagonalized. We obviously orient the Cartesian axes along the orthorhombic crystal axes.

Since all Z_s^* are diagonal Cartesian tensors, we rewrite Eq. (D.83) as

$$C_{ss',\alpha\beta}^{(\mathrm{nonanalytic})}(\mathbf{q}) = \frac{4\pi e^2}{V_{\mathrm{cell}}(\hat{\mathbf{q}}^\dagger \varepsilon_\infty \hat{\mathbf{q}})} (Z_s^{*\dagger} Z_{s'}^*)(\hat{\mathbf{q}}\hat{\mathbf{q}}^\dagger). \qquad (\mathrm{D.84})$$

Even the products $Z_s^{*\dagger} Z_{s'}^*$ are diagonal, and the modes are purely longitudinal or purely transverse.

For phonons polarised along the principal axes, the dynamical matrix can be dealt with as a scalar (in general different in different directions), and one may proceed analogously to our treatment of a cubic binary crystal. In that case the 6×6 dynamical matrix factorized in three 2×2 blocks, each with one acoustic mode and only one optical mode; of the three blocks one was longitudinal and two transverse. In the generalised case the matrix is $3n \times 3n$, with $n-1$ optical modes per direction.

Adopting from now on a scalar notation, we write the free energy per cell using as independent variables the field E and the transverse (zero field) normal mode coordinates u_n. The second order expansion of the free energy per cell is

$$\mathcal{F}(E, \{u_j\}) = \mathcal{F}_0 + \frac{1}{2}\sum_{j=1}^{n-1}\omega_j^2 u_j^2 - \frac{V_{\mathrm{cell}}}{8\pi}\varepsilon_\infty E^2 - eE\sum_{j=1}^{n-1} Z_j^* u_j. \qquad (\mathrm{D.85})$$

We have used for the expansion only $1/3$ of the optic modes: those polarised along the given Cartesian axis. We are thus generalizing the Huang single-mode theory to the $(n-1)$-mode case. Notice however that here the normal-mode coordinates u_j include a factor with the dimensions of $(\mathrm{mass})^{1/2}$, while the mode effective charges Z_j^* include a factor with the dimensions of $(\mathrm{mass})^{-1/2}$.

In Eq. (D.85) the mode frequencies ω_j are obviously the transverse ones ($E = 0$). The equation of motion for the normal modes are

$$f_j = -\omega_j^2 u_j + eZ_j^* E \qquad (\mathrm{D.86})$$

$$D = \frac{4\pi e}{V_{\mathrm{cell}}}\sum_{j=1}^{n-1} Z_j^* u_j + \varepsilon_\infty E. \qquad (\mathrm{D.87})$$

Proceeding as above the dielectric constant in the given direction is

$$\varepsilon'(\omega) = \varepsilon_\infty + \frac{4\pi e^2}{V_{\mathrm{cell}}}\sum_{j=1}^{n-1}\frac{(Z_j^*)^2}{\omega_j^2 - \omega^2} \qquad (\mathrm{D.88})$$

210

$$\varepsilon''(\omega) \;=\; \frac{4\pi^2 e^2}{V_{\text{cell}}} \sum_{j=1}^{n-1} (Z_j^*)^2 \delta(\omega_j^2 - \omega^2), \tag{D.89}$$

and obviously obeys Kramers-Kronig relationships. While the poles of $\varepsilon'(\omega)$ are the frequencies of the transverse zone-center modes, its zeros are the frequencies of the longitudinal ones: it is not straightforward to find such zeros from Eq. (D.88).

We write therefore the equations of motion for the longitudinal modes by setting $D = 0$ in Eqs. (D.86) and (D.87):

$$-\omega u_j \;=\; -\omega_j u_j + e Z_j^* E$$
$$0 \;=\; e \sum_{j=1}^{n-1} Z_j^* u_j + \varepsilon_\infty E, \tag{D.90}$$

which yields

$$\omega^2 u_j = \omega_j^2 u_j + \frac{4\pi e^2}{V_{\text{cell}} \varepsilon_\infty} \sum_{j'=1}^{n-1} Z_j^* Z_{j'}^* u_{j'}. \tag{D.91}$$

This equation shows that, in general, the longitudinal normal mode coordinates are *different* from the transverse ones. This may happen even in cubic crystals, like e.g. a cubic perovskite [277], whose structure is shown in Fig. D.4.

Suppose then we have diagonalized the longitudinal dynamical matrix, Eq. (D.91), and found its eigenvalues $\tilde{\omega}_j^2$: since these are the zeros of $\varepsilon'(\omega)$, we may rewrite identically Eq. (D.88) as

$$\frac{\varepsilon'(\omega)}{\varepsilon_\infty} = \frac{\prod_j (\tilde{\omega}_j^2 - \omega)^2}{\prod_j (\omega_j^2 - \omega)^2}, \tag{D.92}$$

thus generalising the analogous single-mode formula, Eq. (D.20). This result was found in 1961 by Kurosawa [280]. We thus immediately get the generalised Lyddane-Sachs-Teller relationship in the form

$$\frac{\varepsilon_0}{\varepsilon_\infty} = \frac{\prod_j \tilde{\omega}_j^2}{\prod_j \omega_j^2}, \tag{D.93}$$

and we remind that this only holds if all the Born charge tensors are diagonal on the Cartesian axes.

For a cubic crystal all the zone-center modes are threefold degenerate (at $\mathbf{E} = 0$): we may thus rewrite Eqs. (D.88) and (D.89) as

$$\varepsilon'(\omega) \;=\; \varepsilon_\infty + \frac{4\pi e^2}{3 V_{\text{cell}}} \sum_{j=1}^{3n-3} \frac{(Z_j^*)^2}{\omega_j^2 - \omega^2} \tag{D.94}$$

$$\varepsilon''(\omega) \;=\; \frac{4\pi^2 e^2}{3 V_{\text{cell}}} \sum_{j=1}^{3n-3} (Z_j^*)^2 \delta(\omega_j^2 - \omega^2), \tag{D.95}$$

211

Where now the sum is over *all* the $3n - 3$ zone-center optic modes.

D.10 Infrared spectra at finite temperature

In the present Appendix, we have implicitly assumed zero temperature so far. In this framework, the imaginary part of the isotropic dielectric response $\varepsilon''(\omega)$ is given by Eq. (D.95) for a cubic crystalline system. We wish to compare with the fluctuation formula, Eq. (C.41) valid for a general classical system at finite T. We reproduce here Eq. (C.41) for the sake of clarity:

$$\varepsilon''(\omega) = 4\pi\chi''(\omega) = \frac{2\pi\beta\omega}{3V} \int_{-\infty}^{\infty} dt \, e^{i\omega t} \langle \mathbf{d}(t) \cdot \mathbf{d}(0) \rangle. \tag{D.96}$$

As discussed in Sec. C.6 this formula applies when the equilibrium fluctuations are evaluated at $\mathbf{E} = 0$, and $\mathbf{d} = V\mathbf{P}$, where V is the volume of the periodic simulation cell. Ideally, the thermodynamic limit obtains for $V \to \infty$.

We pause to comment on a very important feature: our expression in Eq. (D.89) was evaluated at zero T, while Eq. (D.96) provides by definition the finite-T response of the system. The key point is that—for a purely harmonic system—the correlation function in Eq. (D.96) is exactly proportional to $1/\beta$, and $\varepsilon''(\omega)$ is T-independent. In fact in a harmonic system at thermal equilibrium the energy is in average half kinetic and half potential. From the equipartition theorem and from Eq. (D.85) at $\mathbf{E} = 0$ the average potential energy per degree of freedom is

$$\frac{1}{2}\omega_j^2 \langle u_j^2 \rangle = \frac{1}{2\beta}, \tag{D.97}$$

where we remind that u_j^2 includes a factor with the dimension of a mass. The $1/\beta$ factor cancel the β factor in Eq. (D.96): a virtue of harmonic systems only.

We are now going to do the bookkeeping, in order to show in detail that for our harmonic system Eq. (D.95) and Eq. (D.96) are indeed identical. The free evolution of each normal mode—with zero field—is $u_j(t) = u_j(0)e^{-i\omega_j t}$, hence the dipole of each normal mode is

$$d_j(t) = eZ_j^* u_j(0)e^{-i\omega_j t}. \tag{D.98}$$

We remind that here the normal-mode coordinates u_j include a factor with the dimensions of $(\text{mass})^{1/2}$, while the mode effective charges Z_j^* include a factor with the dimensions of $(\text{mass})^{-1/2}$. Due to cubic symmetry, the j-th normal-mode coordinate is parallel to one of the Cartesian axes.

Since correlation functions are quadratic, it is safer to separate the real from the imaginary part:

$$u_j(t) = u_j'(t) + iu_j''(t) = [u_j'(0) + iu_j''(0)](\cos\omega_j t - i\sin\omega_j t)$$

$$
\begin{aligned}
u'_j(t) &= u'_j(0)\cos\omega_j t + u''_j(0)\sin\omega_j t \\
u'_j(0)u'_{j'}(t) &= u'_j(0)u'_{j'}(0)\cos\omega_j t + u'_j(0)u''_{j'}(0)\sin\omega_j t.
\end{aligned}
\tag{D.99}
$$

When the system is at equilibrium with a thermostat at the inverse temperature β the $j \neq j'$ fluctuations are uncorrelated. Eq. (D.97) yields

$$
\langle u'_j(0)u'_{j'}(t)\rangle = \delta_{jj'}\langle[u'_j(0)]\rangle^2 \cos\omega_j t = \frac{\delta_{jj'}}{\beta\omega_j^2}\cos\omega_j t
\tag{D.100}
$$

for any given Cartesian direction; orthogonal fluctuations are uncorrelated. Summing over the modes polarised in the three directions, the dipole-dipole time correlation function is

$$
\langle \mathbf{d}(t)\cdot\mathbf{d}(0)\rangle = \frac{e^2}{\beta}\sum_j \frac{(Z_j^*)^2}{\omega_j^2}\cos\omega_j t.
\tag{D.101}
$$

Taking then the Fourier transform (a.k.a. "power spectrum") we get

$$
\begin{aligned}
\int_{-\infty}^{\infty} dt\, e^{i\omega t}\,\langle \mathbf{d}(t)\cdot\mathbf{d}(0)\rangle &= \frac{\pi e^2}{\beta}\sum_j \frac{(Z_j^*)^2}{\omega_j^2}\left[\delta(\omega_j - \omega) + \delta(\omega_j + \omega)\right] \\
&= \frac{2\pi e^2}{\beta\omega}\sum_j (Z_j^*)^2\delta(\omega_j^2 - \omega^2), \quad \omega > 0.
\end{aligned}
\tag{D.102}
$$

Inserting this into the fluctuation-dissipation expression, Eq. (D.96), we cancel—as anticipated—the apparent β-dependence, and we arrive at

$$
\varepsilon''(\omega) = \frac{4\pi^2 e^2}{3V}\sum_j (Z_j^*)^2\delta(\omega_j^2 - \omega^2),
\tag{D.103}
$$

which coincides with Eq. (D.95), the only difference being V vs. V_{cell} in the denominator. Clearly, the equipartition theorem, Eq. (D.97), holds even for a single cell with a microscopic number of degrees of freedom, *provided* these $3n - 3$ degrees of freedom fluctuate at equilibrium with a thermostat at inverse temperature β. In anharmonic systems with a non separable Hamiltonian, instead, a large supercell of volume V is mandatory for the use of the fluctuation formula, Eq. (D.96). In any case, the dipole of a cell is extensive and the response is intensive.

The equipartition theorem no longer holds in quantum statistical mechanics; nonetheless in the harmonic case it is easy to correct Eq. (D.96) for quantum effects. It is enough to replace

$$
\frac{1}{2}\omega_j^2\langle u_j^2\rangle = \frac{1}{2\beta} \quad \rightarrow \quad \frac{1}{2}\frac{\frac{\hbar\omega}{2}}{\text{tgh}\frac{\beta\hbar\omega}{2}}.
\tag{D.104}
$$

One obtains in this way the T-dependent quantum response of the harmonic system as

$$\varepsilon''(\omega) = 4\pi\chi''(\omega) = \frac{4\pi\beta\omega}{3\hbar V}\mathrm{tgh}\frac{\beta\hbar\omega}{2}\int_{-\infty}^{\infty} dt\, e^{i\omega t}\langle \mathbf{d}(t)\cdot\mathbf{d}(0)\rangle. \qquad (D.105)$$

D.11   Anharmonic systems

As anticipated, the dipole-dipole autocorrelation functions are a standard tool to compute the infrared spectra of disordered, anharmonic systems, from molecular dynamics simulations, both classical and Car-Parrinello. The previous discussion of the harmonic case was provided for pedagogical purposes only.

There is some disagreement between different authors about the relative merits of Eq. (D.105) vs. Eq. (D.96). In fact Eq. (D.96) is the exact formula at the purely classical level, while the quantum-corrected formula applies in principle to the harmonic case only. The issue mostly concerns liquid water and in general hydrogen-bonded systems, since H atoms are not classical particles at room temperature. The relative merits of the different quantum corrections for hydrogen-bonded anharmonic systems are investigated in Ref. [281]. We point out, nonetheless, that only Eq. (D.96) obeys Kramers-Kronig relationships, as it is clear from the previous derivation.

The absorption coefficient per unit path length is related to $\varepsilon''(\omega)$ as [263]:

$$\alpha(\omega) = \frac{\omega}{c\,n(\omega)}\varepsilon''(\omega) \qquad (D.106)$$

where $n(\omega)$ is the index of refraction. We therefore obtain the two main formulæ on the market, from either Eq. (D.96) or Eq. (D.105), as:

$$\alpha(\omega) = \frac{2\pi\omega^2\beta}{3cVn(\omega)}\int dt\,\langle \mathbf{d}(0)\cdot\mathbf{d}(t)\rangle\,e^{i\omega t}; \qquad (D.107)$$

$$\alpha(\omega) = \frac{4\pi\omega}{3\hbar cVn(\omega)}\mathrm{tgh}\frac{\beta\hbar\omega}{2}\int dt\,\langle \mathbf{d}(0)\cdot\mathbf{d}(t)\rangle\,e^{i\omega t}. \qquad (D.108)$$

Within Car-Parrinello simulations the dipole of the simulation cell is evaluated as $\mathbf{d} = V\mathbf{P}$, where the electronic term in $\mathbf{P}$ is given by the modern theory of polarization, by computing the single-point Berry phase (see Sec. 5.6.2) at each time step or—equivalently—the Wannier centers. Not surprising, the material whose infrared spectrum has been most studied is liquid water. The very first Car-Parrinello infrared spectrum for liquid water appeared in 1997 [93]. Many other followed over the years; the literature is also flooded by simulations based on a vast zoology of classical force models: all of them inadequate (in my view) to capture the key features of hydrogen bonding.

Notice that the simulation only needs polarization *differences* in small time steps. In fact at any discretized time $n\Delta t$ the polarization is

$$
\begin{aligned}
\mathbf{P}(n\Delta t) &= \mathbf{P}(0) + [\mathbf{P}(\Delta t) - \mathbf{P}(0)] + [\mathbf{P}(2\Delta t) - \mathbf{P}(\Delta t)] + \dots \\
&+ [\mathbf{P}(n\Delta t) - \mathbf{P}((n-1)\Delta t).
\end{aligned}
\tag{D.109}
$$

For small enough Δt the polarization quantum is harmless, as discussed in Sec. 5.6.2.

Bibliography

[1] R. Resta, Rev. Mod. Phys. **66**, 899 (1994).

[2] R. Resta, in: *Quantum-Mechanical Ab-initio Calculation of the Properties of Crystalline Materials*, Lecture Notes in Chemistry, Vol. **67**, edited by C. Pisani (Springer, Berlin, 1996), p. 273.

[3] R. Resta, J. Phys.: Condens. Matter **12**, R107 (2000).

[4] R. Resta, J. Phys.: Condens. Matter **14**, R625 (2002).

[5] D. Vanderbilt and R. Resta, in: *Conceptual foundations of materials: A standard model for ground- and excited-state properties*, S.G. Louie and M.L. Cohen, eds. (Elsevier, 2006), p. 139.

[6] R. Resta and D. Vanderbilt, in: *Physics of Ferroelectrics: a Modern Perspective*, Topics in Applied Physics Vol. **105**, Ch. H. Ahn, K. M. Rabe, and J.-M. Triscone, eds. (Springer-Verlag, 2007), p. 31.

[7] R. Resta, J Phys.: Conference Series **117**, 012024 (2008).

[8] R. Resta, J. Phys.: Condens. Matter **22** 123201 (2010).

[9] R. Resta, Eur. Phys. J. B **79**, 121 (2011).

[10] A. Z. Hasan and C. L. Kane, Rev. Mod. Phys. **82**, 3045 (2010).

[11] J. E. Moore, Nature **464**, 194 (2010).

[12] N. Nagaosa, J. Sinova, S. Onoda, A. H. MacDonald, and N. P. Ong, Rev. Mod. Phys. **82**, 1539 (2010).

[13] D. Xiao, M.-C. Chang, and Q. Niu, Rev. Mod. Phys. **82**, 1959 (2010).

[14] E. Prodan, J. Phys. A **44**, 113001 (2011).

[15] X.-L. Qi and S.-C. Zhang, Rev. Mod Phys. **83**, 1057 (2011).

[16] M. Z. Hasan and J. E. Moore, Annu. Rev. Condens. Matter Phys. **2**, 55 (2011).

[17] J. Maciejko, T. L. Hughes, and S.-C. Zhang, Annu. Rev. Condens. Matter Phys. **2**, 31 (2011).

[18] T. Thonhauser, Int. J. Mod. Phys. B **25**, 1429 (2011).

[19] N. Marzari, A. A. Mostofi, J. R. Yates, I. Souza, and D. Vanderbilt, Rev. Mod. Phys. **84**, 1419 (2012).

[20] N. A. Spaldin, J. Solid State Chem. **195**, 2 (2012).

[21] R. Resta, Riv. Nuovo Cimento **41**, 463 (2018).

[22] R. Resta, *Electrical polarization and orbital magnetization: The position operator tamed*, in: *Handbook of Materials Modeling*, W. Andreoni and S. Yip, eds.
`https://doi.org/10.1007/978-3-319-42913- 7_12- 1`.

[23] R. Resta, in: *Topology, Entanglement, and Strong Correlations Modeling and Simulation* Vol. **10**, E. Pavarini and E. Koch, eds. (Forschungszentrum Juelich, 2020), Ch. 10. `https://https://www.cond-mat.de/events/correl20/manuscripts/resta.pdf`.

[24] D. Vanderbilt, *Berry Phases in Electronic Structure Theory* (Cambridge University Press, Cambridge, 2018).

[25] M. V. Berry, Nature Phys. **6**, 148 (2010).

[26] Y. Aharonov and D. Bohm, Phys. Rev. **115**, 485 (1959); reprinted in *Geometric Phases in Physics*, edited by A. Shapere and F. Wilczek (World Scientific, Singapore, 1989), p.104.

[27] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures in Physics, Vol. 2* (Addison Wesley, Reading, 1964), Sect. 15-4.

[28] R. G. Chambers, Phys. Rev. Lett. **5**, 3 (1960).

[29] M. Peshkin and A. Tonomura, *The Aharonov–Bohm Effect* (Springer, Berlin, 1989).

[30] `http://en.wikipedia.org/wiki/SQUID`.

[31] P. Bocchieri and A. Loinger, Nuovo Cimento **47**, 475 (1978).

[32] `http://www.phy.bris.ac.uk/people/berry_mv/index.html`.

[33] M. V. Berry, Proc. Roy. Soc. Lond. A **392**, 45 (1984).

[34] H. C. Longuet-Higgins, U. Öpik, M. H. L. Pryce, and R. A. Sack, Proc. Roy. Soc. A **244**, 1 (1958).

[35] G. Herzberg and H. C. Longuet-Higgins, Discuss. Faraday Soc. **35**, 77 (1963).

[36] C. A. Mead and D. G. Truhlar, J. Chem. Phys. **70**, 2284 (1979).

[37] C. A. Mead, Chemical Physics **49**, 23 (1980).

[38] V. Heine, Phys. Rev.**145**,593 (1966).

[39] J. A. Appelbaum and D. R. Hamann, Phys. Rev. B **10**, 4973 (1974).

[40] L. Kleinman, Phys. Rev. B **11**, 858 (1975).

[41] F. Claro, Phys. Rev. B **17**, 699 (1977).

[42] J. A. Appelbaum, G. A. Baraff, and D. R. Hamann, Phys. Rev. B **14**, 1623 (1976).

[43] P. Streda, J. Phys. C **15**, L717 (1982).

[44] K. von Klitzing, G. Dorda, and M. Pepper, Phys. Rev. Lett. **45**, 494 (1980).

[45] T. Ando, J. Phys. Soc. Jpn. **37**, 622 (1974).

[46] R. B. Laughlin, Phys. Rev. B **23**, 5632 (1981).

[47] R. E. Prange, S. M. Girvin, M. E. Cage, and K. von Klitzing, *The Quantum Hall Effect*, Second Edition (Springer, New York, 1990).

[48] D. Yoshioka, *The Quantum Hall Effect* (Springer, Berlin, 2002).

[49] D. Bures, Trans. Am. Math. Soc. **135**, 199 (1969).

[50] *Geometric Phases in Physics*, edited by A. Shapere and F. Wilczek (World Scientific, Singapore, 1989).

[51] D. J. Thouless, *Topological Quantum Numbers in Nonrelativistic Physics* (World Scientific, Singapore, 1998).

[52] A. Bohm, A. Mostafazadeh, H. Koizumi, Q. Niu, and J. Zwanzinger, *The Geometric Phase in Quantum Systems* (Springer, Berlin, 2003).

[53] J. J. Sakurai, *Modern Quantum Mechanics* (Addison-Wesley, Reading, 1994), p.140.

[54] J. P. Provost and G. Vallee, Commun. Math Phys. **76**, 289 (1980).

[55] J. Zak, Phys. Rev. Lett. **62**, 2747 (1989).

[56] N. Marzari and D. Vanderbilt, Phys. Rev. B **56**, 12847 (1997).

[57] D. J. Thouless, M. Kohmoto, M. P. Nightingale, and M. den Nijs, Phys. Rev. Lett. **49**, 405 (1982).

[58] D. J. Thouless, Phys. Rev. B **27**, 6083 (1983).

[59] F. D. M. Haldane, Rev. Mod. Phys. **89**, 040502 (2017).

[60] S. Coh, D. Vanderbilt, A. Malashevich, I. Souza and D. Vanderbilt, Phys. Rev. B **83**, 085108 (2011).

[61] J. Liu and D. Vanderbilt, Phys. Rev. B **92**, 245138 (2015).

[62] X.-L. Qi, T. L. Hughes, and S.-C. Zhang, Phys. Rev. B **78**, 195424 (2008).

[63] A. M. Essin, A. M. Turner, J. E. Moore, and D. Vanderbilt, Phys. Rev. B **81**, 205104 (2010).

[64] R. Resta, Eur. Phys J. B **91**, 100 (2018).

[65] C. A. Mead, Rev. Mod. Phys. **64**, 51 (1992).

[66] M. Kohmoto, Ann. Phys. **160**, 343 (1985).

[67] D. N. Zubarev, *Non-Equilibrium Statistical Thermodynamic* (Consultants Bureau, New York, 1974).

[68] P. Schmelcher, L. S. Cederbaum, and H.–D. Meyer, Phys. Rev. A **38**, 6066 (1988).

[69] L. Yin and C. A. Mead, J. Chem. Phys. **100**, 8125 (1994).

[70] D. Ceresoli, R. Marchetti and E. Tosatti, Phys. Rev. B **75**, 161101(R) (2007).

[71] D. Ceresoli and E. Tosatti, Phys. Rev. Lett. **89**, 116402 (2002).

[72] P. Lazzeretti and R. Zanasi, Phys. Rev. A **32**,2607 (1985) .

[73] N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Saunders, Philadelphia, 1976).

[74] J. C. Slater, Phys. Rev. **76**, 1592 (1959).

[75] J. M. Luttinger, Phys. Rev. **84**, 814 (1951).

[76] J. Zak, Phys. Rev. **168**, 686 (1968).

[77] M.-C. Chang and Q. Niu, Phys. Rev. B **53**, 7010 (1996).

[78] G. Sundaram and Q. Niu, Phys. Rev. B **59**, 14915 (1999).

[79] G. Panati, H. Spohn, and S. Teufel, Commun. Math Phys. **242**, 547 (2003).

[80] D. Xiao, J. Shi, and Q. Niu, Phys. Rev. Lett. **95**, 137204 (2005).

[81] C. Duval *et al.*, Mod. Phys. Lett. B **20**, 373 (2006).

[82] J. B. Pendry and C. H. Hodges, J. Phys. C **17**, 1269 (1984).

[83] J. Meister and W. H. E. Schwarz, J. Phys. Chem. **98**, 8245 (1994).

[84] Ph. Ghosez, J.-P. Michenaud, and X. Gonze, Phys. Rev. B **58**, 6224 (1998).

[85] A. Walsh, A. A. Sokol, J. Buckeridge, D. O. Scanlon, and C. R. A. Catlow, Nat. Materials **17**, 958 (2018).

[86] H. Raebiger, S. Lany, R. Resta, and A. Zunger, Nat. Prec. (2009): https://doi.org/10.1038/npre.2009.4012.1.

[87] R. Resta, Nature **453**, 735 (2008).

[88] H.-L. Sit, R. Car, M. H. Cohen, and A. Selloni, Inorg. Chem. **50**, 10259 (2011).

[89] L. Jiang, S. V. Levchenko, and A. W. Rappe, Phys. Rev. Lett. **108**, 166403 (2012).

[90] F. Grasselli and S. Baroni, Nature Mat. **15**, 967 (2019).

[91] J. P. Hansen and I. R. McDonald, *Theory of Simple Liquids* (Academic, New York, 1986).

[92] D. Frenkel and B. Smit, *Understanding Molecular Simulation: from Algorithms to Applications. 2nd ed.* (Academic, San Diego, 2002).

[93] P. L. Silvestrelli, M. Bernasconi, and M. Parrinello, Chem. Phys. Lett. **277**, 478 (1997).

[94] A. Pasquarello and R. Resta, Phys. Rev. B **68**, 174302 (2003).

[95] L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media* (Pergamon Press, Oxford, 1984).

[96] J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975).

[97] C. Kittel, *Introduction to Solid State Physics*, 8th. edition (Wiley, Hoboken, NJ, 2005).

[98] R. M. Martin, Phys. Rev. B **9**, 1998 (1974).

[99] M. Posternak, A. Baldereschi, A. Catellani and R. Resta, Phys. Rev. Lett. **64**, 1777 (1990).

[100] R. Resta, Ferroelectrics **136**, 51 (1992).

[101] R. D. King-Smith and D. Vanderbilt, Phys. Rev. B **47**, 1651 (1993).

[102] G. Ortíz and R. M. Martin, Phys. Rev. B **49**, 14202 (1994).

[103] R. Resta, Phys. Rev. Lett. **80**, 1800 (1998).

[104] http://www.abinit.org/.

[105] http://www.crystal.unito.it/.

[106] http://www.quantum-espresso.org.

[107] http://www.uam.es/departamentos/ciencias/fismateriac/siesta/.

[108] http://cms.mpi.univie.ac.at/vasp/.

[109] http://www.cpmd.org/.

[110] R. Resta, Modelling Simul. Mater. Sci. Eng. **11**, R69 (2003).

[111] W. H. Duan and Z. R. Liu, Curr. Opin. Solid State Mater. Sci. **10**, 40 (2006).

[112] M. Rabe, and J.-M. Triscone, eds., *Physics of Ferroelectrics: a Modern Perspective*, Topics in Applied Physics Vol. **105**, Ch. H. Ahn, K. (Springer-Verlag, 2007).

[113] M.P. Marder, *Condensed Matter Physics* (Wiley, New York, 2000).

[114] G. Grosso and G. Pastori-Parravicini, *Solid State Physics*, second edition (Elsevier, Amsterdam, 2014).

[115] I. Souza, J. Ìñìguez, and D. Vanderbilt, Phys. Rev. Lett. **89**, 117602 (2002).

[116] P. Umari and A. Pasquarello, Phys. Rev. Lett. **89**, 157602 (2002).

[117] D. Vanderbilt and R. D. King-Smith, Phys. Rev. B **48**, 4442 (1993).

[118] R. Resta, M. Posternak, and A. Baldereschi, Phys. Rev. Lett. **70**, 1010 (1993).

[119] G. H. Wannier, Phys. Rev. **52**, 191 (1937).

[120] A. A. Mostofi, Y.-S. Lee, I. Souza, D. Vanderbilt, and N. Marzari, Comput. Phys. Commun. **178**, 685 (2008).

[121] J. R. Yates, C. J. Pickard, and F. Mauri, Phys. Rev. B **76**, 024401 (2007).

[122] E. I. Blount, in *Solid State Physics*, edited by H. Ehrenreich, F. Seitz and D. Turnbull, vol **13** (Academic, New York, 1962), p. 305.

[123] O. F. Mossotti, Memorie di Matematica e di Fisica della Società Italiana delle Scienze Residente in Modena, **24**, 49 (1850); R. Clausius, *Die Mechanische Behandlung der Electrica* (Vieweg, Berlin, 1879).

[124] I. Dabo, B. Kozinsky, N. E. Singh-Miller, and N. Marzari, Phys. Rev. B **77** , 115139 (2008); *ibid.* **84**, 159910(E) (2011).

[125] P. Umari, A. Dal Corso, and R. Resta, in: *Fundamental Physics of Ferroelectrics: 2001 Williamsburg Workshop*, H. Krakauer, ed. (AIP, Woodbury, New York, 2001), p. 107.

[126] K. N. Kudin, R. Car, and R. Resta, J. Chem. Phys. **127**, 194902 (2007).

[127] R. Resta, J. Chem. Phys. **154**, 050901 (2021).

[128] W. Kohn, Phys. Rev. **133**, A171 (1964).

[129] Q. Niu, Phys. Rev. **33**, 5368 (1986).

[130] M. Stengel, Phys. Rev. B **84**, 205432 (2011).

[131] N. C. Bristowe, P. B. Littlewood, and E. Artacho, J. Phys.: Condens. Matter **23**, 081001 (2011).

[132] A. Pasquarello and R. Car, Phys. Rev. Lett. **79**, 1766 (1997).

[133] X. Gonze, Ph. Ghosez, and R. W. Godby, Phys. Rev. Lett. **74**, 4035 (1995).

[134] S. Baroni, S. de Gironcoli, A. Dal Corso, and P. Giannozzi, Rev. Mod. Phys. **73**, 515 (2001).

[135] W. P. Su, J. R. Schrieffer, and A. J. Heeger, Phys. Rev. Lett. **42**, 1698 (1979).

[136] A. Malashevich, I. Souza, S. Coh, and D. Vanderbilt, New J. Phys. **12**, 053032 (2010).

[137] A. M. Essin, J. E. Moore, and D. Vanderbilt, Phys. Rev. Lett. **102**, 146805 (2009).

[138] N. Varnava and D. Vanderbilt, Phys. Rev. B **98**, 245117 (2018).

[139] F. Bloch, Z. Phys. **52**, 555 (1928).

[140] A. H. Wilson, Proc. Roy. Soc. A **133**, 458, and **134**, 277 (1931).

[141] N. F. Mott, Proc. Phys. Soc. (London) **62**, 416 (1949).

[142] P. W. Anderson, Phys. Rev. **109**, 1492 (1958).

[143] N. Mott, *Metal-Insulator Transitions*, 2nd ed. (Taylor & Francis, London, 1990).

[144] E. Abrahams (Ed.), *50 Years of Anderson Localization*, (World Scientific, Singapore, 2010).

[145] W. Kohn, in *Many–Body Physics*, edited by C. DeWitt and R. Balian (Gordon and Breach, New York, 1968), p. 351.

[146] R. Resta and S. Sorella, Phys. Rev. Lett. **82**, 370 (1999).

[147] I. Souza, T. Wilkens, and R. M. Martin, Phys. Rev. B **62**, 1666 (2000).

[148] C. Sgiarovello, M. Peressi, and R. Resta, Phys. Rev. **64**, 115202 (2001).

[149] R. Resta, J. Chem. Phys. **124**, 104104 (2006).

[150] A. Marrazzo and R. Resta, Phys. Rev. Lett. **122**, 166602 (2019).

[151] G. Bellomia and R. Resta, Phys. Rev. B **102**, 205123 (2020).

[152] D. J. Scalapino, S. R. White, and S. C. Zhang, Phys. Rev. **47**, 7995 (1993).

[153] P. B. Allen, in: *Conceptual foundations of materials: A standard model for ground- and excited-state properties*, S.G. Louie and M.L. Cohen, eds. (Elsevier, 2006), p. 139.

[154] R. Resta, J. Phys. Condens. Matter **30**, 414001 (2018).

[155] N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Saunders, Philadelphia, 1976), Ch. 1 and Ch. 13.

[156] D. N. Zubarev, Soviet Phys. Ushpekhi **3**, 320 (1960).

[157] R. McWeeny, *Methods of Molecular Quantum Mechanics*, Second Edition (Academic, London, 1992).

[158] D. Chandler, *Introduction to Modern Statistical Mechanics* (Oxford University Press, Oxford, 1987).

[159] E. Akkermans, J. Math. Phys. **38**, 1781 (1997).

[160] Y. Gao, S. A. Yang, and Q. Niu, Phys. Rev. B **91**, 214405 (2015).

[161] X. Wang, D. Vanderbilt, J. R. Yates, and I. Souza, Phys. Rev. B **76**, 195109 (2007).

[162] N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Saunders, Philadelphia, 1976), Appendix E.

[163] G. Grosso and G. Pastori-Parravicini, *Solid State Physics*, second edition (Elsevier, Amsterdam, 2014), Sec. 2.6.2.

[164] R. Resta, Phys. Rev. Lett. **96**, 137601 (2006).

[165] L. He and D. Vanderbilt, Phys. Rev. Lett. **86**, 5341 (2001).

[166] G. F. Giuliani and G. Vignale, *Quantum Theory of the Electron Liquid* (Cambridge University Press, Cambridge, 2005).

[167] G. L. Bendazzoli, S. Evangelisti, A. Monari, and R. Resta, J. Chem. Phys. **133**, 064703 (2010).

[168] G. L. Bendazzoli, S. Evangelisti, and A. Monari, Int. J. Quantum Chem. **112**, 653 (2012).

[169] E. Drigo and R. Resta, Phys. Rev. B **101**, 165120 (2020).

[170] M. Rigol and B. S. Shastry, Phys. Rev. B **77**, 161101(R) (2008).

[171] F. L. Pilar. *Elementary Quantum Chemistry* (McGraw-Hill, 1990).

[172] E. Abrahams, P. W. Anderson, D. C. Licciardello, and T. V. Ramakrishnan, Phys. Rev. Lett. **42**, 673 (1979).

[173] B. Kramer and A. MacKinnon, Rep. Prog. Phys. **56**, 1469 (1993).

[174] D. J. Thouless, Phys. Rep. **13**, 93 (1974).

[175] A. Lagendijk, B. van Tiggelen, and D. S. Wiersma, Phys. Today **62**(8), 24 (2009).

[176] A. MacKinnon and B. Kramer, Phys. Rev. Lett. **47**, 1546 (1981).

[177] E. Hofstetter and M. Schreiber, Phys. Rev. B **49**, 14726 (1994).

[178] K. Slevin and T. Ohtsuki, Phys. Rev. Lett. **82**, 382 (1999).

[179] A. Rodriguez, L. J. Vasquez, K. Slevin, and R. A. Römer, Phys. Rev. B **84**, 134209 (2011).

[180] F. Evers and A. D. Mirlin. Rev. Mod. Phys. **80**, 1355 (2008).

[181] T. Olsen, R. Resta, and I. Souza, Phys. Rev. B **95**, 045109 (2017).

[182] J. M. Ziman, *Models of Disorder* (Cambridge University Press, Cambridge, 1979).

[183] T. Wilkens and R. M. Martin, Phys. Rev. B **63**, 235108 (2001).

[184] S. Tamura and H. Yokoyama, JPS Conf. Proc. **3**, 013003 (2014).

[185] V. K. Varma and S. Pilati, Phys. Rev. B **92**, 134207 (2015).

[186] R. Resta and S. Sorella, Phys. Rev. Lett. **87**, 4738 (1995).

[187] E. H. Lieb and F.-Yu. Wu, Phys. Rev. Lett. **20**, 1445 (1968).

[188] M. Motta *et al.*, Phys. Rev. X **10**, 031058 (2020).

[189] L. Stella, C. Attaccalite, S. Sorella, and A. Rubio, Phys. Rev. B **84**, 245117 (2011).

[190] M. El Khatib *et al.*, J. Chem. Phys. **142**, 094113 (2015).

[191] D. R. Hofstadter, Phys. Rev. B **14**, 2239 (1976).

[192] F. D. M. Haldane, Phys. Rev. Lett. **61**, 2015 (1988).

[193] C.-Z. Chang et al. Science **340**, 167 (2013).

[194] C.-Z. Chang et al. Nature Materials **14**, 473 (2015).

[195] C. L. Kane and E. J. Mele, Phys. Rev. Lett. **95**, 226801 (2005).

[196] Q. Niu and D. J. Thouless, J. Phys A **17**, 2453 (1984).

[197] Q. Niu, D. J. Thouless, and Y. S. Wu, Phys. Rev. B **31**, 3372 (1985).

[198] R. Bianco, R. Resta, and I. Souza, Phys. Rev. B **90**, 125153 (2014).

[199] A. Marrazzo and R. Resta, Phys. Rev. B **95**, 121114(R) (2017).

[200] R. Resta, https://arxiv.org/abs/2101.10949.

[201] T. Jungwirth, Q. Niu, and A. H MacDonald, Phys. Rev. Lett. **88**, 207208 (2002).

[202] M. Onoda and N. Nagaosa, J. Phys. Soc. Jpn. **71**, 19 (2002).

[203] Y. Yao, L. Kleinman, A. H. MacDonald, J. Sinova, T. Jungwirth, D.-S. Wang, E. Wang, and Q. Niu, Phys. Rev. Lett. **92**, 037204 (2004).

[204] X. Wang, J. R. Yates, I. Souza, and D. Vanderbilt, Phys. Rev. B **74**, 195118 (2006).

[205] F. D. M. Haldane, Phys. Rev. Lett. **93**, 206602 (2004).

[206] T. Thonhauser and D. Vanderbilt, Phys. Rev. B **74**, 235111 (2006).

[207] D. J. Thouless, J. Phys. C **17**, L325 (1984).

[208] C. Brouder, G. Panati, M. Calandra, Ch. Mourougane, and N. Marzari, Phys. Rev. Lett. **98**, 046402 (2007).

[209] L. Fu and C. L. Kane, Phys. Rev. B **74**, 195312 (2006).

[210] T. Fukui, Y. Hatsugai, and H. Suzuki, J. Phys. Soc. Japan **74**, 1674 (2005).

[211] D. Ceresoli and R. Resta, Phys. Rev. B **76**, 012405 (2007).

[212] A. A. Soluyanov and D. Vanderbilt, Phys. Rev, B **83**, 035108 (2011).

[213] N. Hao *et al.*, Phys. Rev. B **78**, 075438 (2008).

226

[214] S. Coh and D. Vanderbilt, Phys. Rev. Lett. **102**, 107603 (2009).

[215] R. Bianco and R. Resta, Phys. Rev. B **84**, 241106(R) (2011).

[216] T. Thonhauser, D. Ceresoli, D. Vanderbilt, and R. Resta, Phys. Rev. Lett. **95**, 137205 (2005).

[217] D. Ceresoli, T. Thonhauser, D. Vanderbilt, and R. Resta, Phys. Rev. B **74**, 024408 (2006).

[218] R. Bianco and R. Resta, Phys. Rev. Lett. **110**, 087202 (2013).

[219] I. Sodemann and L. Fu, Phys. Rev. Lett. **115**, 216806 (2015).

[220] O. Matsyshyn and I. Sodemann, Phys. Rev. Lett. **123**, 246602 (2019).

[221] S. Nandy and I. Sodemann, Phys. Rev. B **100**, 195117 (2019).

[222] Q. Ma *et al.* Nature (London) **565**, 337 (2019).

[223] A. J. P. Meyer and G.Asch, J. Appl. Phys. **32**, S330 (1961).

[224] L. L. Hirst, Rev. Mod. Phys. **69**, 607 (1997).

[225] D. Xiao, Y. Yao, Z. Fang, and Q. Niu, Phys. Rev. Lett. **97**, 026603 (2006).

[226] J. Shi, G. Vignale, D. Xiao, and Q. Niu, Phys. Rev. Lett. **99**, 197202 (2008).

[227] I. Souza and D. Vanderbilt, Phys. Rev. B **77**, 054438 (2008).

[228] K.-T. Chen and P. A. Lee, Phys. Rev. B **84**, 205137 (2011).

[229] H. Schulz-Baldes and S. Teufel, Commun. Math. Phys. **319**, 649 (2013).

[230] T. Thonhauser, D. Ceresoli, A.A. Mostofi, N. Marzari, R. Resta, and D. Vanderbilt, J. Chem. Phys. **131**, 101101 (2009).

[231] D. Ceresoli, U. Gerstmann, A. P. Seitsonen, and F. Mauri, Phys. Rev. B **81**, 060409(R) (2010).

[232] M. G. Lopez, D. Vanderbilt, T. Thonhauser, and I. Souza, Phys. Rev. B **85**, 014435 (2012).

[233] A. Marrazzo and R. Resta, Phys. Rev. Lett. **116**, 137201 (2016).

[234] R. Bianco and R. Resta, Phys. Rev. B **93**, 174417 (2016).

227

[235] W. Kohn, Phys. Rev. Lett. **76**, 3168 (1996).

[236] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures in Physics, Vol. 2* (Addison Wesley, Reading, 1964), Sect. 36-6.

[237] D. J. Griffiths, *Introduction to Electrodynamics, 3rd Ed.* (Prentice-Hall, 1999).

[238] N. A. Hill, J. Phys. Chem. B **104**, 6694 (2000).

[239] I. E. Dzyaloshinskii, Sov. Phys. JETP **10**, 628 (1960).

[240] D. N. Astrov, Sov. Phys. JETP **11**, 708 (1960).

[241] J. Ìñiguez, Phys. Rev. Lett. **101**, 117201 (2008).

[242] J. C. Wojdel and J. Ìñiguez, Phys. Rev. Lett. **103**, 267205 (2009).

[243] R. Resta, Phys. Rev. Lett. **106**, 047202 (2011).

[244] R. Resta, Phys. Rev. B **84**, 214428 (2011).

[245] A. Malashevich, S. Coh, I. Souza, and D. Vanderbilt, Phys. Rev. B **86**, 094430 (2012).

[246] F. Bernardini, V. Fiorentini, and D. Vanderbilt, Phys. Rev. B **56**, 10024 (1997).

[247] S. E. Park and T. R. Shrout, J. Appl. Phys. **82**, 1804 (1997).

[248] R. M. Martin, Phys. Rev. B **5**, 1607 (1972).

[249] R. M. Martin, Phys. Rev. B **6**, 4874 (1972); W. F. Woo and W. Landauer, Phys. Rev. B **6**, 4876 (1972); R. Landauer, Solid St. Commun. **40**, 971 (1981); C. Kallin and B. J. Halperin, Phys. Rev. B **29**, 2175 (1984); R. Landauer, Ferroelectrics **73**, 41 (1987); A. K. Tagantsev, Phase Transitions **35**, 119 (1991).

[250] S. de Gironcoli, S. Baroni, and R. Resta, Phys. Rev. Lett. **62**, 2853 (1989).

[251] B. Alder and T. E. Wainwright, Phys. Rev. A **1**, 18 (1970).

[252] A. Widom, Phys. Rev. A **3**, 1394 (1971).

[253] H. Kornfeld, Z. Phys. 22, 27 (1924).

[254] J. G. Kirkwood, J. Chem. Phys. **7**, 911 (1939).

[255] H. Fröhlich, *Theory of Dielectrics*, 2nd edition (Clarendon, Oxford, 1958).

[256] M. Neumann, Mol. Phys. **50**, 841 (1983).

[257] S. W. de Leeuw, J. W. Perram, and E. R. Smith, Proc. Roy. Soc. London Ser. A **373**, 27 (1980); *ibid.* **373**, 57 (1980).

[258] J. Ford, Physics Reports **213**, 271 (1992).

[259] See Fermi-Pasta-Ulam on Wikipedia.

[260] K. Huang, Proc. Roy. Soc. **A203**, 178 (1950).

[261] M. Born and K. Huang, *Dynamical Theory of Crystal Lattices* (Oxford University Press, Oxford, 1954).

[262] R. H Lyddane, R. G. Sachs, and E. Teller, Phys. Rev. **59**, 673 (1941).

[263] D. A. McQuarrie, *Statistical Mechanics* (University Science Books, Sausalito, California, 2000).

[264] A. A. Maradudin, E. W. Montroll, G. H. Weiss, and I. P. Ipatova, *Theory of Lattice Dynamics in the Harmonic Approximation*, Solid State Physics, Suppl. **3** (Academic, New York, 1971).

[265] G. V. Chester, Adv. Phys. **10**, 357 (1961).

[266] E. G. Brovman and Yu. M. Kagan, in: *Dynamical Properties of Solids*, edited by G. K. Horton and A. A. Maradudin, vol. 1 (North-Holland, Amsterdam, 1974), p. 191.

[267] T. Toya, J. Res. Inst. Catalysis, Hokkaido Univ. **6**, 161 and 183 (1958).

[268] W. A. Harrison, *Pseudopotentials in the Theory of Metals* (Benjamin, New York, 1966).

[269] W. Kohn, Phys. Rev. Lett. **2**, 393 (1959).

[270] Brockhouse et al, Phys. Rev. Lett. **7**, 93 (1961).

[271] E. W. Kellermann, Phil. Trans. Roy. Soc. , **A238**, 513 (1940).

[272] H. Bilz and W. Kress, *Phonon Dispersion Relations in Insulators* (Springer, Berlin, 1979).

[273] R. Pick, M. H. Cohen, and R. M. Martin, Phys. Rev. B **1**, 910 (1970).

[274] R. Resta and A. Baldereschi, Phys. Rev. B **24**, 4839 (1981).

[275] R. M. Martin and K. Kunc, Phys. Rev. B **24**, 2081 (1981).

[276] Ph. Ghosez, E. Cockayne, U. V. Waghmare, and K. M. Rabe Phys. Rev. B **60**, 836 (1999).

[277] W. Zhong, R. D. King-Smith and D. Vanderbilt, Phys. Rev. Lett **72**, 3618 (1994).

[278] W. Cochran and R. A. Cowley, J. Phys. Chem. Solids **23**, 4471 (1962).

[279] P. Giannozzi, S. de Gironcoli, P. Pavone, and S. Baroni, Phys. Rev. B **43**, 7231 (1991).

[280] T. Kurosawa, J. Phys. Soc. Jpn. **16**, 1298 (1961).

[281] R. Ramirez, T. Lòpez-Ciudad, P. Kumar P, and D. Marx, J.Chem. Phys. **121**, 3973 (2004).