

ABSTRACT

Title of Dissertation: **LOW TEMPERATURE
SCANNING TUNNELING MICROSCOPY
OF TOPOLOGICAL MATERIALS
AND MAGNETIC STRUCTURES**

Joseph Murray
Doctor of Philosophy, 2023

Dissertation Directed by: **Professor Christopher Lobb**
Department of Physics

Scanning tunneling microscopy (STM) provides an opportunity to study the physical and electromagnetic properties of surfaces at the atomic scale. When performed at low temperatures, in high magnetic fields, and with a variety of different probes, it offers a wide range of methods by which novel materials of great practical and theoretical interest can be evaluated, characterized, and even fabricated with atomic precision.

This thesis describes three independent STM studies performed at cryogenic temperatures. In the first, I present an in-situ modification to our 4K STM which permits us to current-bias our samples during STM operation. The modification can be used to study non-equilibrium effects such as spin accumulation induced by a current through a spin Hall material and the spin-momentum locking which is present at the surface of topological insulators.

Next, I examined oxygen-doped aluminum films with anomalously high kinetic inductance. A suggested explanation was the migration of oxygen to the grain boundaries, forming a percola-

tion network separated by Josephson links. To determine the coupling between grains, I studied the films using milliKelvin STM performed with a superconducting tip.

Finally, transport measurements performed by our collaborators indicated the possible presence of a topological Hall effect in thin films of Cr_2Te_3 , induced by the presence of topologically non-trivial magnetic textures called magnetic skyrmions. In order to provide more decisive evidence, I studied the films using spin-polarized STM at 4K.

In addition to the experimental studies, this work explores the theoretical underpinnings of the novel materials which constitute the frontier of our current understanding of condensed matter physics. An emphatically pedagogical viewpoint is adopted throughout as part of a continuing effort to bridge the gap between experiment and theory.

LOW TEMPERATURE SCANNING TUNNELING MICROSCOPY OF TOPOLOGICAL MATERIALS AND MAGNETIC STRUCTURES

by

Joseph Edward Murray

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:

Professor Christopher Lobb, Chair/Advisor
Dr. Robert E. Butera, Co-Advisor
Dr. Michael Dreyer
Professor Johnpierre Paglione
Professor John Cumings

© Copyright by
Joseph Edward Murray
2023

Dedication

To my mother, whose unconditional love and support were the bedrock of my youth, and whose memory will reverberate through the rest of my life. I love you.

Acknowledgments

This work represents the end of a very long and winding road which I would never have reached without the love and support of a vast community of mentors, friends and loved ones.

I owe an enormous debt of gratitude to my advisor Prof. Chris Lobb and my supervisor Dr. Robert Butera, who took me on as a student and guided me to where I am today. Special thanks are due to Dr. Michael Dreyer, who spent years teaching me the lion's share of what I know about experimental physics in general and scanning probe microscopy in particular. I am also indebted to Profs. Johnpierre Paglione and John Cumings for adding their unique insights and perspectives to my dissertation committee.

I would also like to thank my peers at the Laboratory for Physical Sciences. In particular, my heartfelt appreciation goes out to the LPS STM group which at various times included Wan-Ting Liao, Kevin Dwyer, Sungha Baek, Azadeh Farzaneh, Jon Marbey, and Giovanni Franco-Rivera; to Jimmy Kotsakidis for his experience, expertise, encouragement, and steadfast Australian-ness; to Jennifer DeMell for her relentless support and friendship; and to the LPS machinists Don Crouse, Ruben Brun, and Paul Davis, without whom much of my work would have been impossible.

I would not have made it through the turmoil of my first few years in graduate school without the support of many dear friends I made at the University of Maryland, including Anastasia Marchenkova, Gina Quan, Steve Ragole, Caroline Figgatt, Antony Speranza, David Somers, Matt

Anthony, Connor Roncaioli, Pete Volpe, Min-A Cho Zeno, and Dalia Ornelas. But my support network extended far beyond UMD - I have the greatest lifelong friends that could ever be asked for (///), and could not begin to express the depth of my gratitude to each and every one of them.

Finally, I could not have made it to the end of this road without my father John, my brother Steve, and my partner Elaine, who was a rock of support through good days and bad, and who generously kept me alive during the long days and nights of writing.

Table of Contents

Dedication	ii
Acknowledgements	iii
Table of Contents	v
List of Tables	viii
List of Figures	ix
Chapter 1: Introduction	1
Chapter 2: A Brief Introduction to Topology	4
2.1 The Main Ideas	4
2.1.1 Constructing Important Topologies	9
2.1.2 Topological Ideas	15
2.2 Topological Equivalence of Spaces	18
2.3 Homotopy	23
2.3.1 Homotopy Invariants	26
2.3.2 Homotopy Equivalence of Topological Spaces	30
Chapter 3: Physical Applications of Topology I	37
3.1 Foundational Ideas	37
3.1.1 A Mathematical Example	39
3.1.2 An Important Note Regarding Stability	41
3.2 Magnetic Skyrmions	42
3.2.1 Homotopic Preliminaries	42
3.2.2 2D Magnetic Skyrmions	44
3.2.3 Skyrmion Stability	45
3.3 Abrikosov Vortices in Type-II Superconductors	54
3.3.1 Homotopic Preliminaries	54
3.3.2 A Note on Magnetic Free Energy	56
3.3.3 Ginzburg-Landau Theory of Superconductivity	59
3.3.4 Abrikosov Vortices	60
Chapter 4: Elementary Band Theory	65
4.1 Preliminary Remarks	65

4.2	Bloch Theorem in the Continuum	66
4.2.1	Formulation of the Theorem	66
4.2.2	Example Models	69
4.3	Bloch Theorem on a Lattice	72
4.3.1	Formulation of the Theorem	74
4.3.2	Example Models	76
4.4	Finite Systems and the Thermodynamic Limit	79
4.5	The Emergence of Forbidden Bands	82
4.6	The Wannier Representation	87
Chapter 5: Physical Applications of Topology II: Topological Band Theory		91
5.1	The Bloch Bundle	92
5.1.1	The Nontrivial Topology of the Brillouin Zone	92
5.1.2	The Bundle Picture	95
5.1.3	States in the Bloch Bundle	96
5.1.4	The Valence Bundle	96
5.1.5	Transport of States	99
5.1.6	Wilson Lines	101
5.1.7	The Role of Gauge Invariance	103
5.1.8	Wilson Loops	104
5.1.9	The Berry Curvature	106
5.1.10	The Abelian Simplification	108
5.2	The Physicality of the Wilson Loop	109
5.2.1	The Zak Phase in 1D	109
5.2.2	The Bulk-Boundary Correspondence	111
5.2.3	The Surface Dispersion	113
5.3	Symmetry-Protected Topological Phases	116
5.3.1	The Chern Insulator in 2D	117
5.3.2	Inversion Symmetry	120
5.3.3	Time Reversal Symmetry	123
Chapter 6: Scanning Tunneling Microscopy in Theory and Practice		128
6.1	Principles of Operation	128
6.1.1	The Bardeen Tunneling Model	129
6.1.2	Beyond Bardeen: The Tersoff-Hamman Formula	138
6.2	Spin-Polarized STM	139
6.2.1	Polarized Tunneling	140
6.2.2	Differential Conductance	141
6.2.3	Establishing a Spin-Polarized Tip	142
6.2.4	Extracting a Spin-Polarized Signal	145
6.3	STM at Cryogenic Temperatures	147
6.3.1	Cooling Mechanisms	147
Chapter 7: Development of Cryogenic Current-Biased STM		154
7.1	Motivation for Current-biased STM	155

7.1.1	Scanning Tunneling Potentiometry	155
7.1.2	Local Sensing of Spin Transport Phenomena	159
7.2	Description of the LPS 4K STM	160
7.2.1	The UHV Subsystem	160
7.2.2	The Cryostat and Magnet Dewar	163
7.2.3	The Microscope	164
7.2.4	The Electronics and STM Control	167
7.3	Modification Design and Technical Specifications	167
7.3.1	The Sample Stud	167
7.3.2	The Compensation Circuit	170
7.4	Experimental Characterization	171
7.4.1	Thermal Characteristics	171
7.4.2	Validation via Potentiometry	173
7.5	Future Work	175
7.5.1	Current Mapping and Potentiometry	175
7.5.2	Manipulation of Solitons	177
Chapter 8:	Experiment: Superconducting Gap Variation in Granular Aluminum Films	180
8.1	Overview	180
8.2	Experimental Description	182
8.2.1	The mK-STM	183
8.2.2	Sample Preparation	187
8.2.3	Results	187
8.3	Discussion	190
8.3.1	Comparison with Previous Work	191
8.3.2	A Future Experiment: Vortex Matter in Granular Superconductors	193
8.3.3	Conclusion and Outlook	194
Chapter 9:	Experiment: Magnetic Skyrmions in Cr_2Te_3 Thin Films	196
9.1	The Topological Hall Effect	196
9.1.1	2DEG Model	198
9.1.2	An Alternative Model	205
9.2	Possible Topological Hall Effect in Cr_2Te_3	207
9.2.1	Structure and Properties of Cr_2Te_3	208
9.2.2	Experimental Description	210
	Bibliography	218
	Bibliography	218

List of Tables

- 3.1 Exact and approximate expressions for the various contributions to the total energy of an isolated, axially-symmetric magnetic skyrmion in the 2D limit $d \rightarrow 0$. The I 's are numerical functions of the reduced skyrmion radius $\rho \equiv R/\Delta$, where R is the radius of the $m_z = 0$ contour and Δ is the domain wall thickness. 51

List of Figures

2.1	(a) A set is <i>open</i> if it constitutes a neighborhood for each point it contains. (b) This set is not open because it contains points at the edge, and does not constitute a neighborhood for those points.	5
2.2	Given an arbitrary set U , its interior U° is the largest open subset of U , the closure is the smallest closed superset of U , and the boundary ∂U is the complement of U° in $\overline{U}$. U is open if and only if it does not contain any boundary points, i.e. $U \cap \partial U = \emptyset$	8
2.3	A set $U \subseteq S^1$ is open in the subset topology inherited from $\mathbb{R}^2$ if there exists a set $V \subseteq \mathbb{R}^2$ which is open in the topology on $\mathbb{R}^2$ such that $U = V \cap S^1$	11
2.4	A set $U \subseteq S \times T$ is open in the product topology, if for all $p \in U$, there exist open sets $A \subseteq S$ and $B \subseteq T$ such that $p \in A \times B \subseteq U$	12
2.5	The set $[0, 2\pi]$ is mapped to the quotient set $[0, 2\pi]/\sim$ by the quotient map q . A set $U \subseteq [0, 2\pi]/\sim$ is open in the quotient topology if its preimage under q is open in the topology on the original space $[0, 2\pi]$	14
2.6	A function $f : A \rightarrow B$ is said to be continuous at $a \in A$ if its output can be restricted to an arbitrarily small open neighborhood $V \ni f(a)$ by restricting its <i>input</i> to a sufficiently small open neighborhood $U \ni a$	16
2.7	A sequence $\{f_n\}$ is said to converge to a limit L if every open set containing L contains all but finitely-many points in the sequence.	17
2.8	Given a parent space $(M, \mathcal{O}_M)$ and two subsets A and B , the subspace $(S, \mathcal{O}_S)$ (where $S = A \cup B$ and $\mathcal{O}_S$ is the subset topology) is disconnected if and only if $A \cap B = \emptyset$	19
2.9	(a) In T_0 spaces, no two distinct points $s \neq t$ are in exactly the same open sets, but it's possible that every open set containing s also contains t (but not the reverse). (b) In T_1 spaces, we can always find open sets containing s but not t and t but not s , but it's possible that such sets necessarily intersect somewhere. (c) In T_2 spaces, we can always find open sets around any two distinct points which don't intersect at all.. . . .	21
2.10	If M is a bounded subset of $\mathbb{R}^n$, then its closure $\overline{M}$ equipped with the subset topology is a compact space.	22
2.11	f and g are homotopic if they can be continuously deformed into one another. The curves in (a) are homotopic (with varying stages of the homotopy shown with dashed lines), but the curves in (b) are non-homotopic because they may not cross over the puncture in the center of the disk.	24

2.12	(a) A contractible space is homotopy-equivalent to a point. (b) The punctured disk is not contractible, but it is homotopy-equivalent to S^1 (i.e. its boundary). (c) The surface of a hollow cube is homotopy equivalent to S^2 (the surface of a ball).	32
2.13	The punctured disk D^* has a non-trivial fundamental group given by $\pi_1(D^*; d_0) \simeq \mathbb{Z}$ for arbitrary d_0 . Curve f belongs to the equivalence class with winding number $+1$. g_1, g_2, g_3 are all homotopic and belong to the class with winding number 0 . h belongs to the class with winding number -2 , with the minus sign originating from the clockwise orientation.	35
3.1	The homotopic equivalence of n -dimensional loops is preserved under the action of a continuous state $\psi : R \rightarrow \mathcal{M}$, which motivates the definition of the maps $\psi_n^* : \pi_n(R) \rightarrow \pi_n(\mathcal{M})$ which map families of equivalent n -loops in R to corresponding families in $\mathcal{M}$.	38
3.2	An example of three states $\psi_1, \psi_2, \psi_3 : R \rightarrow \mathcal{M}$ and their corresponding topological indices (ν_+, ν_-) . Under our classification scheme, ψ_1 has indices $(0, 0)$, ψ_2 has indices $(1, 1)$, and ψ_3 has indices $(1, 0)$.	40
3.3	The punctured space $\mathbb{R}^3 - \vec{0}$ is homotopic to the 2-sphere S^2 . On this domain, point defects take the form of monopole configurations, and our classification scheme essentially counts the number of monopoles which are present at the location of the puncture.	43
3.4	The stereographic projection map Π provides a continuous, one-to-one correspondence between points on S^2 and points in $\mathbb{R}^2 \cup \{\infty\}$, where the so-called <i>point at infinity</i> is associated to the north pole of the sphere.	44
3.5	An example of a magnetic texture with skyrmion number $\Sigma = +1$.	46
3.6	The profile of a magnetic skyrmion in the 360° domain wall model with $\rho_0 = 0.25$. The color scale denotes the polar angle θ , with the asymptotic magnetization pointing down. The inner circle is the parameter ρ_0 , while the outer circle is the more intuitive $\rho \equiv \sinh^{-1}(\cosh(\rho_0))$.	49
3.7	A $U(1)$ -valued order parameter exhibits non-trivial states only if $\pi_1(R) \neq \{0\}$. This applies to point defects in 2D (a) and line defects in 3D (b), but not point defects in 3D (c).	55
3.8	Abrikosov vortices and their associated magnetic fields for (a) $\kappa = 0.2$ and (b) $\kappa = 10$. The density ρ is reflected in the surface shape, while the color reflects the magnitude of the magnetic field. Note that in the $\kappa \ll 1$ limit, the flux is restricted to the very center of the core, whereas for $\kappa \gg 1$ the flux extends deep into the bulk.	63
4.1	The Wigner-Seitz unit cell of a 2D triangular lattice, consisting of all of the points nearer to the origin than to any other lattice site.	67
4.2	The free electron dispersion relation for the $n = 0$ and $n = \pm 1$ bands. The degenerate points where the bands intersect are denoted by red dots.	70

4.3	Illustrations of the allowed energies in the Kronig-Penney model for several values of the barrier strength. The allowed energies $E \equiv q^2 \hbar^2 / 2m$ are those for which $\cos(ka) = \cos(qa) + \alpha \sin(qa)/qa \in [-1, 1]$. When $\alpha = 3$ (Figs. <i>a</i> and <i>b</i>), the δ -barriers are fairly weak and the forbidden gaps are small. When $\alpha = 10$ (Figs. <i>c</i> and <i>d</i>), the δ -barriers are strong and the gaps are correspondingly large compared to the width of the bands.	73
4.4	The dispersion relation for the bipartite chain with on-site energy $\pm\epsilon$ and intracell and intercell hopping amplitudes v and w . Allowed energy ranges are highlighted in green. In (a), $(\epsilon, v, w) = (0.5, 1, 1)$ and the system is gapped. In (b), $(\epsilon, v, w) = (0, 1, 1)$ and the gap closes, making the system a semimetal.	78
4.5	The origin of the band gap in nearly free electron models at the edge of the Brillouin zone. States at the edge of BZ can be written as (co)sinusoidal waves whose corresponding probability densities are maximized (red) or minimized (blue) on the lattice sites.	84
4.6	A cartoon illustrating the breaking of the isolated atom energy degeneracy by inclusion of nearest-neighbor interactions. The dashed lines correspond to the unperturbed wavefunctions centered at lattice sites n and $n + 1$. The symmetric linear combination (red) is larger than the antisymmetric combination (green) in between the lattice sites, and therefore feels the potential due to the neighboring sites more strongly.	86
4.7	A representation of the magnitude of the conduction band Wannier function $\mathcal{W}_{1,0}$ for the bipartite lattice (see Sec. 4.3.2.2), where the intercell hopping amplitude w is half the value of the intracell hopping amplitude v . Whereas the Bloch waves u_{nk} have uniform probability density on the A and B sublattices, the Wannier representation clearly shows a strong intra-cell bond.	88
5.1	A cartoon schematic of the Bloch bundle. The Brillouin zone BZ is called the <i>base space</i> , and the copies of $L^2(\mathcal{C})$ which are attached to each point of BZ are called the <i>fibers</i> . This type of structure is called a fiber bundle; because $L^2(\mathcal{C})$ is a vector space, the Bloch bundle is a special type of fiber bundle called a <i>vector bundle</i>	95
5.2	A genuine function $f : \mathbf{BZ} \rightarrow L^2(\mathcal{C})$ (left) vs. a section σ of the Bloch bundle (right). f sends all points in BZ to the same vector space, while σ does not. This has critical implications for differentiation, which fundamentally relies on expressions like $f(\mathbf{k} + d\mathbf{k}) - f(\mathbf{k})$ which don't make any sense if $f(\mathbf{k} + d\mathbf{k})$ and $f(\mathbf{k})$ live in different vector spaces.	97
5.3	A generic band structure before (a) and after (b) subtracting the center of the gap at each point in the Brillouin zone. This can always be done to a gapped system without changing its topology, so we will always assume that the gap is centered at 0.	99
5.4	The transport procedure defined by the projectors $\mathbb{P}_{\mathbf{k}}$. As the initial vector is moved in infinitesimal steps along the curve, it generically leaves the subspace of occupied states, but the relevant projection operator acts to project it back. In the limit as $ \delta\mathbf{k}_i \rightarrow 0$, this process preserves the normalization of the initial vector, and the operator $W[\gamma]$ will be seen to be unitary.	100

5.5	The periodic evolution of the Wannier charge center Δ_y as k_x is swept from $-\pi$ to π . The trivial phase (a) can be continuously deformed to $\Delta_y = 0$, but the topological phase (b) cannot, as it undergoes non-zero winding.	112
5.6	(a) In the trivial case, the spectrum of $\mathbb{Z}$ is periodic in $k_{\parallel}$ and exhibits no winding, so when it is squashed into a compact interval (b) there is no guarantee that it will cross the Fermi level. (c) In the trivial case, the spectrum of $\mathbb{Z}$ exhibits winding as $k_{\parallel}$ varies across the BZ, and so the squashed edge spectrum (d) <i>must</i> intersect the Fermi level.	115
5.7	As the flat-band Hamiltonian is deformed back into a realistic one, the edge state spectrum should change continuously, in particular retaining its topological properties (e.g. whether or not it exhibits winding).	116
5.8	The Wannier charge center evolution (a), the spectrum of the projected position operator along y (b), the top edge state dispersion (c), and the helical edge mode (d) of a 2D Chern insulator with Chern number $\text{Ch} = +1$. In (c), the bulk bands are shown in gray, and the Fermi level is denoted by a wavy line.	119
5.9	Several possible Wilson loop spectra which obey the constraints of time-reversal symmetry - namely, symmetry under $k_x \mapsto -k_x$ and double degeneracy at $k_x = 0, \pi$. The trivial phase (a) can be continuously deformed to $\Delta_y = 0$, but the non-trivial phases (b) and (c) cannot. Note that the non-trivial phases can nonetheless be deformed into one another, which corresponds to the Fu-Kane invariant being an element of $\mathbb{Z}_2$	125
5.10	The Wannier charge center evolution (a), the spectrum of the projected position operator along y (b), the top edge state dispersion (c), and the chiral edge modes (d) of a 2D time-reversal symmetric topological insulator. The bulk states in (c) are shown in gray, and the Fermi level is denoted by a wavy line. Solid and dashed curves reflect Kramers partners with opposite spin.	126
6.1	A simple model of tunneling due to Bardeen, in which we treat the two electrodes as quasi-independent systems weakly coupled by a barrier potential.	129
6.2	The decomposition of the potential $V(x)$ as described in Eq. 6.1. Note that the potentials do not sum to $V(x)$, but $H_A + H_B + \Delta$ <i>does</i> sum to H	131
6.3	Examples of spin-polarized STM. (a) is a cartoon depiction of SPSTM with an in-plane polarized tip on a body centered tetragonal (bct) Mn(001) film. The step-wise antiferromagnetism of bct-Mn(001) results in an enhanced tunneling current when the tip and sample magnetizations are parallel and a suppression when they are antiparallel. (b) shows the topography (left) and differential conductance (right) of a real scan performed with a Fe-coated tungsten tip, taken at +0.2 V. The spin-polarization is not immediately obvious in topography, but is manifested as a very clear pattern of alternating bright and dark steps in dI/dV . (c) and (d) are the topography and differential conductance of a scan I performed with a bulk Cr tip on a Cr(001) single crystal, in which the characteristic, step-wise pattern is visible in dI/dV . (a) and (b) are reused with permission from [1].	143

6.4	A photograph and schematic diagram of a liquid helium transfer dewar. Helium is transferred in and out of the dewar using an insulated hose which is inserted into a fill port at the top of the dewar. A vent port enables the user to control the pressure of the helium atmosphere which exists above the liquid. A safety valve (or more typically, several of them) release excessive pressure if needed. The crosshatched region is typically vacuum space and multilayered radiation shielding.	148
6.5	A phase diagram for ^3He - ^4He mixtures. Published by Mets501 as Helium phase diagram and licensed under CC BY-SA 3.0.	150
6.6	A 1K pot, capable of reaching temperatures of ~ 1.5 K. Liquid helium is drawn into the pot by a pump, where it evaporates and provides cooling power. The helium level is controlled by a needle valve, allowing the user to ensure that the helium level is neither above nor below the pot itself.	151
6.7	A schematic diagram of a dilution refrigerator. Reused with permission from [2].	152
7.1	Simulated IV curves and their corresponding standard deviations. (a) simulates curves taken at five different values of the tunneling resistance R_T , resulting in a sharper peak in (c). (b) simulates five curves taken at the same value of R_T , and differ due to thermal noise which is minimized at the true zero of the tip-sample voltage. As shown in (d), this method results in a less precise measurement of the current and voltage offsets.	156
7.2	A schematic drawing of a sample mounted for scanning tunneling potentiometry. Current is injected and extracted through the gold contacts.	158
7.3	A schematic drawing of the 4K STM system, reused with permission from [3]. Samples are loaded into the preparation chamber via a load lock, then into the transfer chamber via a horizontal magnetic transfer rod. A vertical transfer rod picks up the sample and transfers it into the cryostat which is housed within the LHe dewar.	161
7.4	A schematic drawing of the sample holder and transfer rod, reused with permission from [3]. The sample holder (1) is outfitted with a spring plate (2) which holds the sample stud (3) in place. Once inserted into the holder through the keyhole in the back, it is rotated to lock it in place. The sample holder is manipulated in the UHV subsystem via the tab at the top. The sample stud is extracted from the sample holder and moved into the cryostat via a transfer rod with a leaf spring friction coupling (right).	162
7.5	The 4K cryostat probe. The vacuum can which houses the cryostat is screwed in to the top flange (1). Copper heat sinks (2) keep the microscope wiring cool, while thin metallic shutters (3) block room-temperature electromagnetic radiation. A keyhole (4) ensures that the transfer rod remains aligned so it can reach the microscope stage (5) and microscope (6). Figure reused with permission from [3].	165
7.6	The locking mechanism at the top of the 4K microscope. The sample stud is inserted sample-side down into a keyhole and then twisted to lock it in place. The annular bias plate (shown in gold) provides electrical contact to the sample as well as mechanical stability. In this figure, the bias plate has been split in half to allow for current-biasing of the sample.	166

7.7	The original (a) and redesigned (b) sample studs. (1) The body of the stud is made from SS, to which metallic straps (2) are spot welded. A sample (3) is loaded into the trench before being pressed upward into the straps by set screws (4). In the redesigned stud with copper body, a ceramic insert (5) insulates the sample from the stud body. The straps are fixed to the insert by M1 screws (6). A second ceramic insert (7) isolates a contact on one of the wings (8), which is connected to the adjacent strap via an enameled wire which sits in a groove (9). The other strap is connected to the sample body.	168
7.8	The voltage compensation circuit which allows for a dynamic modulation of the bias current. A fraction ξ of the potential difference across the film is subtracted from the nominal bias voltage. When ξ is properly calibrated to the tip location, the local tip-sample voltage remains at V_0 while the bias current and ΔV are allowed to vary.	170
7.9	Equilibration times for the z position of the tip (a) and the thermometers on the microscope body and sample stage (b). The change in vertical tip position is well within the dynamic range of the z piezo, and can easily be accommodated by the feedback controls. The far longer equilibration time of the thermometers is due to the large thermal masses of the microscope body and sample stage.	173
7.10	Scanning tunneling potentiometry of a NbTiN thin film, with $I_{bias} = -3$ mA (left) and $+3$ mA (right). The average potential difference across the scan window agrees with the expected value (assuming a homogeneous film) to within 5%. . .	174
8.1	The mK-STM system at the Laboratory for Physical Sciences, consisting of a room temperature UHV subsystem for sample preparation and transport and a cryostat outfitted with a 1K pot and dilution refrigerator. Figure published in [4]. .	182
8.2	A close-up photograph of the outer and inner tip configuration (a) and an overall schematic (b) of the dual-tip mK STM. The outer (1) and inner (2) tips are mounted on separate outer (3) and inner (4) piezo tubes for lateral scanning control. Coarse motion for the outer assembly is provided by external piezo stacks (5) which move an outer sapphire prism (6). The inner tip can be independently controlled via an inner prism (7) and piezo stacks (8), while also riding along with the outer assembly (9). Figure published in [4].	185
8.3	AFM topography and histograms of grain sizes for films grown with varying pressures of O_2 in the deposition chamber. The scan windows are all 500×500 nm ² , with the exception of the 9 μ Torr scan which was truncated (but remains to scale).	186
8.4	(a) The empirical fit function $I(V)$ given in 8.2 and its derivative. (b) An example of the fit being applied to one of the measured spectra. Despite its simplicity, it fits the data well and serves a straightforward way to extract the gap with substantial computational efficiency.	188
8.5	The topography (a) and measured gap widths (b) of a granular Al film, taken over a 500×500 nm ² scan area. The scale on the topography map is in nm, while the scale on the gap map is in mV. Both images have a lateral resolution of 128×128 pixels.	189

8.6	Histogram of the measured values for the superconducting map shown in Fig. 8.5. The mean gap is $\mu_{\Delta} = 1.75$ mV and the standard deviation is $\sigma_{\Delta} = 257$ μ V. .	189
8.7	The deviation of Δ from its mean value, shown as a spatial plot and as a histogram. The color map has been superimposed on top of a grayscale topography image for visual clarity.	190
8.8	The topography (a) and Δ map (b) of a 50 nm thick granular TiN film, taken with permission from [5]. The anticorrelation between the grain topography and Δ is qualitatively similar to that shown in Fig. 8.5.	192
9.1	A magnetic phase diagram for MnSi, featuring a Skyrmion lattice phase identified via the Hall resistivity ρ_{xy} (color scale). Figure published in [6] and reused under a Creative Commons license.	202
9.2	A typical presentation of the topological Hall effect, observed in a $\text{Bi}_2\text{Te}_3/\text{CrTe}_2$ heterostructure. The the sign of the anomalous Hall resistivity changes at $T = 40$ K, but the topological hall effect is not dramatically changed. Figure published with permission from [7], Copyright 2021 American Chemical Society.	204
9.3	A cartoon illustration of a two-component AHE, in which the dark curve is the sum of the two lighter curves. The superposition of two ordinary AHE signals with opposite signs and different coercive fields gives rise to localized features which mimic the characteristic sign of a true (skyrmionic) topological Hall effect.	206
9.4	Imaging of magnetic skyrmions by Lorentz transmission electron microscopy (a) and out-of-plane and in-plane spin-polarized STM (b,c). Lorentz TEM has the advantage of not relying on an atomically sharp tunnel junction, but SPSTM offers the ability to resolve both in-plane and out-of-plane magnetization. The LTEM image is reproduced with permission from [8], while the SPSTM image is reproduced with permission from [9].	206
9.5	Hall measurements performed on a quasi-2D Cr_2Te_3 film with a thickness of 6 unit cells. The emergence of pronounced, localized humps in the vicinity of the coercive field suggests either the presence of a topological Hall effect or of a multi-domain anomalous Hall effect. Figure published with permission from [10].	207
9.6	The crystal structures of CrTe_2 (left) and Cr_2Te_3 (right). CrTe_2 has a trigonal unit cell ($\alpha = \beta = 90^\circ, \gamma = 120^\circ$) with lattice constants $a = b = 3.86$ Å, $c = 6.2$ Å. Cr_2Te_3 also has a trigonal unit cell, but the lattice constants are enlarged to $a = b = 6.9$ Å, $c = 12.5$ Å, as the larger unit cell is required by the intercalation layers.	208
9.7	Surface morphology of a Cr_2Te_3 film at (a) 2 μm and (b) 250 nm scales with an electrochemically etched bulk Cr tip. (c) shows the height profile of the line cut shown in (b). (d) is a filtered atomic resolution image of one of the island features and was taken with an electrochemically etched tungsten tip. All images taken at $I_t = 100$ pA and $V = 100$ mV.	211
9.8	(a) Topography and (b) differential conductance maps of a Cr_2Te_3 film for $\mu H_z = +2$ T, 0 T, -0.5 T, and -0.9 T, taken with an in-plane polarized STM tip. dI/dV images were processed via Fourier filtering and decorrelation from topography. Possible stripe domains emerge in the -0.5 T and -0.9 T scans, but no signs of a skyrmionic phase are visible.	213

9.9	(a) Topography and (b) differential conductance maps of a Cr_2Te_3 film for $\mu H_z = -2$ T, 0 T, 0.5 T, and 0.9 T, taken with an in-plane polarized STM tip. A mask was applied to emphasize the terrace under examination. dI/dV images were processed via Fourier filtering and decorrelation from topography. Possible magnetic skyrmions emerge at $\mu_0 H_z = 0.9$ T before disappearing at higher fields.	215
9.10	A comparison of transport measurements (a) vs. the emergence of possibly skyrmionic structures at $\mu_0 H_z = 0.9$ T. Based on the hysteresis curves, this is in the immediate vicinity of the coercive field of the film. (c) shows a line cut through the differential conductance map which exhibits symmetric positive and negative variation across the structure, as one would expect from a magnetic skyrmion. Figure (a) published with permission from [10].	216

Chapter 1: Introduction

In 1981, Binnig and Rohrer [11] developed the scanning tunneling microscope (STM) - an instrument capable of measuring the topography and electrical characteristics of surfaces with sub-nanometer precision without direct physical contact with the sample. If a small voltage is applied between a sharp metallic probe and a sample which are separated by a few angstroms, the quantum tunneling effect allows electrons to traverse the classically-forbidden vacuum gap and establish a small but measurable current. In the tunneling regime, this current varies exponentially with the tip-sample distance; if the latter decreases by 1 Å, then the former typically increases by roughly a factor of 10. By slowly scanning the probe across the surface and adjusting its position to maintain a constant tunneling current, the surface topography (as well as its electromagnetic properties) can be measured with sub-nanometer precision.

The STM has revolutionized the way that we study and manipulate physical systems on the smallest of scales. Its sub-nanometer spatial resolution has enabled the characterization of the topography and physical structure of surfaces, atomic features, and composite heterostructures with unprecedented detail, while its wealth of variations has provided a window into the electromagnetic and chemical properties of materials and how they vary over scales from microns to angstroms. The rise of STM-based atomic precision fabrication [12, 13] has even allowed us the ability to build our own nanostructures atom-by-atom. It has become one of the most essential

and direct tools in the arsenal of experimental solid-state physics.

Over roughly the same period of time, topology has emerged as an indispensable tool for our *theoretical* understanding of physical systems. In 1980, Klaus von Klitzing discovered the quantum Hall effect [14], in which a material exhibits a quantized Hall conductance of such high quality that it initially led to a suspicion of the laboratory equipment. The incredible robustness of this quantization in the presence of even quite strong disorder suggested an underlying mechanism which in some sense operated globally, completely glossing over the localized messiness of an imperfectly clean system. Since then, the mathematical tools and concepts from topology have permeated essentially every subfield of condensed matter physics and beyond, informing a complementary picture of global and local influences.

This work reports several novel studies performed using STM as the principal experimental tool, while simultaneously developing a unified theoretical understanding of the phenomena under investigation through the lens of topology. Chapter 2 serves as a self-contained introduction to the basic ideas of topology from a purely mathematical viewpoint before Chapter 3 applies those ideas to the study and characterization of defects in ordered media such as magnetic materials and superconductors. Chapter 4 reviews the fundamentals of the electronic structure of solids before Chapter 5 develops an understanding of the novel topological phases which can emerge in quantum matter and of the surprising correspondence between topological properties of the bulk of a material and the unique and robust electronic states which are localized to its boundaries.

Having built this broad theoretical base, Chapter 6 reviews the subject of STM, the principles on which it operates, and the practical side of operating an STM under cryogenic conditions before the remaining chapters report some of the experimental studies performed over the last several years in the STM group at the Laboratory for Physical Sciences. Chapter 7 describes an

in-situ modification which I implemented on our low-temperature STM to permit the study of non-equilibrium effects which arise only in the presence of a transverse electrical current. Chapter 8 is a report on a study I performed on superconducting granular aluminum at mK temperatures to study the coupling between grains and the somewhat unconventional nature of granular superconductivity in general. Finally, Chapter 9 describes my search for topological magnetic structures as part of a larger, collaborative effort to explain the unexpected electromagnetic behavior of the quasi-2D layered ferromagnet Cr_2Te_3 .

Chapter 2: A Brief Introduction to Topology

This chapter constitutes a purely mathematical but hopefully fairly gentle introduction to topology for someone who had never seen it before. It is my belief that it is helpful to be willing to set physics aside for a moment and wrestle with some mathematical ideas without being burdened by the necessity of practical application. On the other hand, being a physicist, I want to get to the good stuff eventually. The best resources I've found which manage to bridge this divide are the books by Nakahara [15] and Frankel [16]. Additionally, I must credit the lecture course by Dr. Frederic Schuller entitled *The Geometrical Anatomy of Theoretical Physics* [17] as perhaps the most self-contained introduction to mathematics for physicists that I've ever seen, which spurred me in this direction some years ago. At the time of writing, it is freely available on YouTube and from FAU's video archives.

2.1 The Main Ideas

The most fundamental concept in topology is that of a *neighborhood*. Given a set S and a point $p \in S$, a *topology* on S amounts to a definition of what constitutes a neighborhood of p . The following constraints are imposed on all such definitions:

1. If N is a neighborhood of p , then $p \in N$
2. If N is a neighborhood of p and $N \subseteq M$, then M is also a neighborhood of p

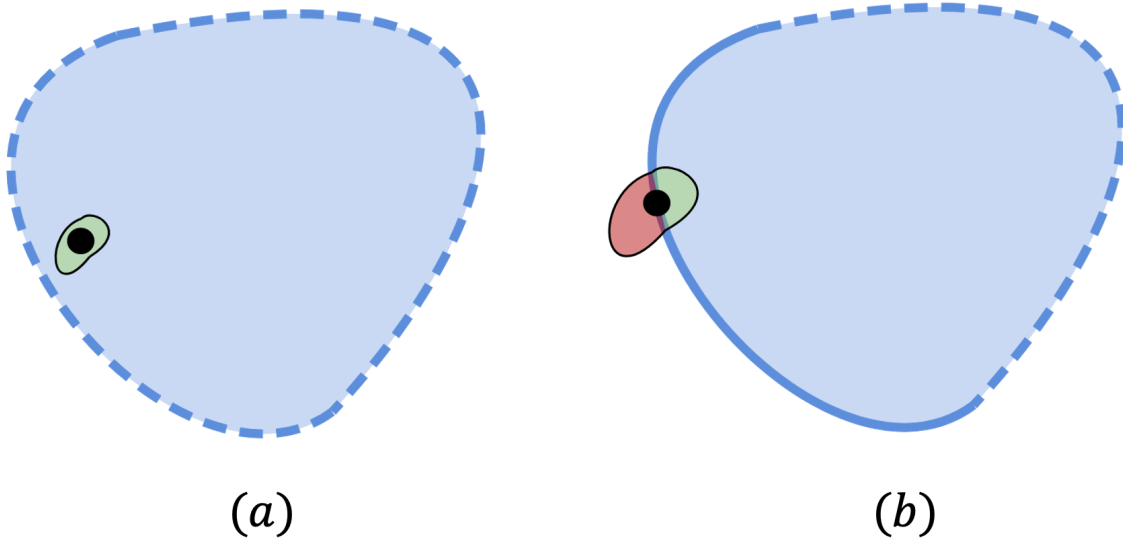


Figure 2.1: (a) A set is *open* if it constitutes a neighborhood for each point it contains. (b) This set is not open because it contains points at the edge, and does not constitute a neighborhood for those points.

3. If N and M are neighborhoods of p , then their intersection $N \cap M$ is a neighborhood of p
4. If N is a neighborhood of p , then there exists some neighborhood M of p such that N is a neighborhood of every point in M .

Given a notion of neighborhood as defined above, we define a set U to be *open* if it is a neighborhood of every point $p \in U$ which it contains. All topological properties of a space can be framed in terms of these open sets. Given a set S , the set of all open sets $U \subseteq S$ - which we denote by $\mathcal{O}$ - is called the *topology* on S . The pair $(S, \mathcal{O})$ is then called a *topological space*.

These requirements do not uniquely associate a topology to any given set. In fact, it's quite the opposite - for a generic set, there are a vast number of topologies which satisfy these conditions.

2.1.0.1 A Few Simple Topologies

Let S be any set. The following definitions of neighborhoods are consistent with the requirements of a topology.

- **The Discrete Topology** Every set containing a point $p \in S$ is a neighborhood of p .
- **The Indiscrete Topology** The *only* neighborhood of a point $p \in S$ is S itself.
- **The Particular Point Topology** Let $q \in S$ be an arbitrary point. A set N containing $p \in S$ is a neighborhood of p if and only if N also contains q .
- **The Cofinite Topology** We call a set $N \subseteq S$ *cofinite* if its complement of N in S contains an infinite number of points. A set $N \subseteq S$ is a neighborhood of p in the cofinite topology if $p \in N$ and N is cofinite.

The discrete and indiscrete topologies are not particularly interesting because of how inclusive and exclusive they are, respectively. The particular point and cofinite topologies is less trivial, but are also of limited importance for our purposes. However, it is easy to show that all four satisfy the requirements of a topology, so it is a good exercise to do so.

It is worth emphasizing that given a set S , whether a subset $U \subseteq S$ is open clearly depends on the topology we choose for S . For example, if we equip S with the discrete topology defined above, then every possible $U \subseteq S$ is open. On the other hand, if we equip S with the indiscrete topology, then *no* subsets $U \subseteq S$ are open, with the exception of $\emptyset$ and S itself. Therefore, when there is a possibility for ambiguity it is important to specify that a set is open *in a particular topology*. Since a topology $\mathcal{O}$ formally consists of the collection of all sets which are to be considered open, the statement that a set U is open in that topology can also be written as $U \in \mathcal{O}$.

2.1.0.2 Terminology

It's worth taking a moment to explicitly mention some important terminology. Let $(S, \mathcal{O}_S)$ be a topological space. As mentioned above, a subset $U \subseteq S$ is called *open* if it is an element of $\mathcal{O}_S$. If the complement of U is open, then U is called *closed*. Confusingly, observe that openness and closedness are neither exhaustive nor mutually exclusive; a set could be either open or closed, both¹ open and closed, or neither open nor closed.

Given a set $U \subseteq S$, the *interior* of U , denoted U° , is the collection of points $p \in U$ such that there exists a neighborhood N of p which lies entirely within U (note that if U is open, then $U^\circ = U$). Equivalently, U° is the union of all open sets $X \subseteq U$. The *closure* of U , denoted $\overline{U}$, is the intersection of all closed sets $X \supseteq U$; as one might expect, $\overline{U}$ is a closed set (in some sense, it is the smallest closed set containing U). The *boundary* of U , denoted ∂U , is the set of points in $\overline{U}$ which are not elements of U° .

In more intuitive terms, a set U is open if it does not contain any of its boundary points. For the topologies of general interest to physics, this means that open sets have the property that their elements remain elements when they are shifted or changed by small amounts. In contrast, the boundary points are teetering on the edge, such that any arbitrarily small shift may be enough to move it outside. Put differently, if U is open then membership in U is a property which is robust to tiny perturbations.

¹Note that for any topological space $(S, \mathcal{O}_S)$, both S and the empty set $\emptyset$ are always both open and closed.

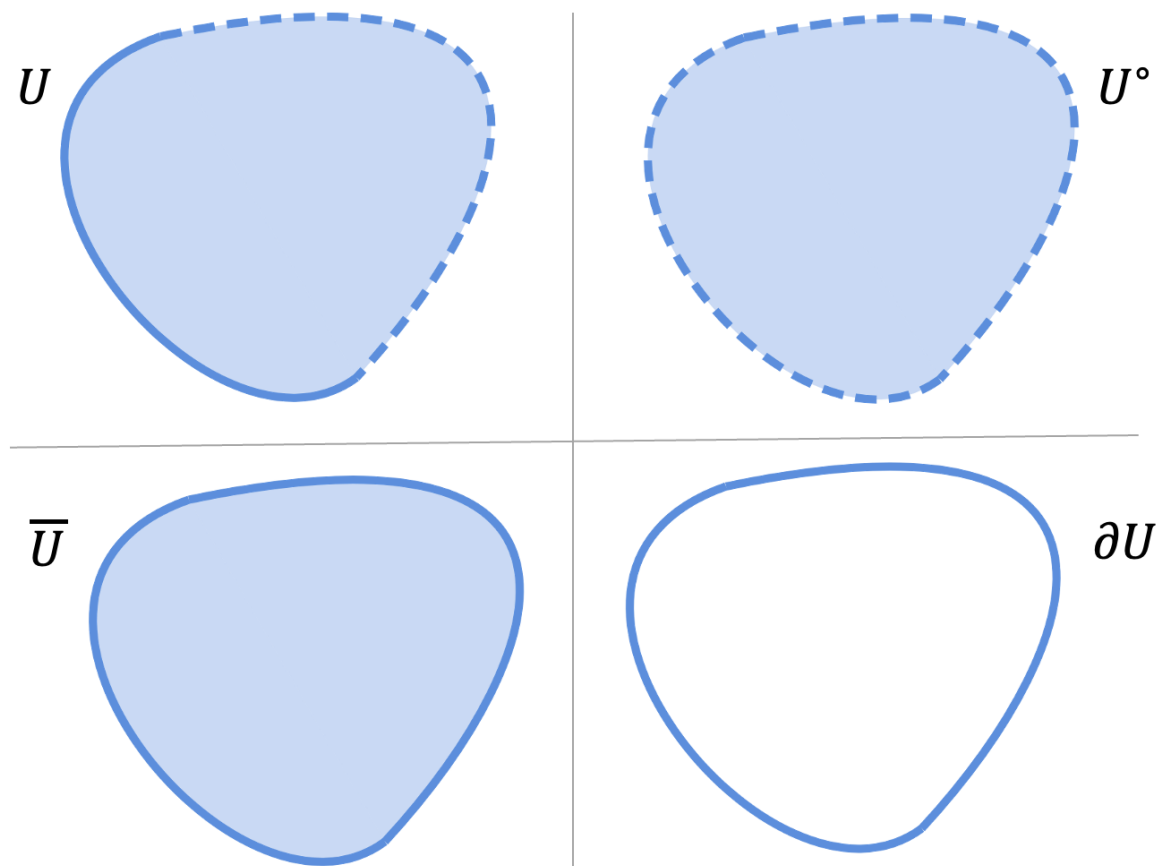


Figure 2.2: Given an arbitrary set U , its interior U° is the largest open subset of U , the closure is the smallest closed superset of U , and the boundary ∂U is the complement of U° in $\overline{U}$. U is open if and only if it does not contain any boundary points, i.e. $U \cap \partial U = \emptyset$.

2.1.1 Constructing Important Topologies

The topologies given above are straightforward to define, and serve as a good warm-up to test our understanding of important ideas. However, in physics we are interested in more complex topologies. The most important topology for our purposes is the so-called *standard topology* $\mathcal{O}_{\mathbb{R}^d}$ on $\mathbb{R}^d$. Once we've defined this fundamental topology, we can inherit a large number of additional topologies from it. For example, any subset $M \subseteq \mathbb{R}^d$ can be endowed with the subset topology inherited from $\mathcal{O}_{\mathbb{R}^d}$.

2.1.1.1 The Standard Topology

To define this topology, we introduce the concept of the open ball of radius $r > 0$ centered at a point $x \in \mathbb{R}^d$, denoted $B_r(x)$; it is defined as

$$B_r(x) := \left\{ y \in \mathbb{R}^d \mid \sum_{i=1}^d (y_i - x_i)^2 < r^2 \right\}$$

That is, $B_r(x)$ is the set of points within a distance r of x , where the notion of distance is defined in the preceding expression. Having made this definition, we say that a set N is a neighborhood of a point $p \in \mathbb{R}^d$ if there exists some $r > 0$ such that $B_r(p) \subseteq N$, and that a set $U \subseteq \mathbb{R}^d$ is open (in the standard topology) if it is a neighborhood of every point $p \in U$. See Fig. 2.1 for an illustration of this topology.

This construction provides a topology on $\mathbb{R}^d$, and whenever we speak of $\mathbb{R}^d$ in the future we will assume it to be equipped with this topology unless otherwise specified. However, any set which can be expressed as a subset of $\mathbb{R}^d$ can inherit a so-called subset topology from its parent,

as we see from the following construction.

2.1.1.2 The Subset Topology

Let $(S, \mathcal{O}_S)$ be a topological space, and let $M \subseteq S$. We can define a topology $\mathcal{O}_M$ on M called the *subset topology* inherited from $\mathcal{O}_S$ as follows. A set $U \subseteq M$ is an element of $\mathcal{O}_M$ if there exists some $V \in \mathcal{O}_S$ such that $U = V \cap M$. In words, U is open in the subset topology if it can be expressed as the intersection of the subset M with a set V which is open in the parent topology.

As an example, let $S = \mathbb{R}^2$ and $\mathcal{O}_S$ be the standard topology on $\mathbb{R}^2$, and let $M = S^1$ where

$$S^1 := \left\{ (a, b) \in \mathbb{R}^2 \mid a^2 + b^2 = 1 \right\}$$

is the unit circle (also called the 1-sphere). Let $U = \{(a, b) \in S^1 \mid a > 0\}$ be the right semicircle of S^1 . It's easy to see that U is *not* an element of $\mathcal{O}_S$ (see the definition above), but it *is* an element of the subset topology $\mathcal{O}_M$. This can be seen by letting $V = \{(a, b) \in \mathbb{R}^2 \mid a > 0\}$ and noting that $U = V \cap S^1$.

The subset topology allows us to start from the standard topology on $\mathbb{R}^d$ and construct a topology for any of its subsets. A very large number of topological spaces² can be expressed in this way, making this a very powerful construction. The product topology construction adds another crucial tool to our toolbox, which provides a topology to the Cartesian product of spaces.

²The n -spheres, the n -tori, and spaces of $n \times n$ matrices (interpreted as elements of $\mathbb{R}^{n^2}$) are classic examples.

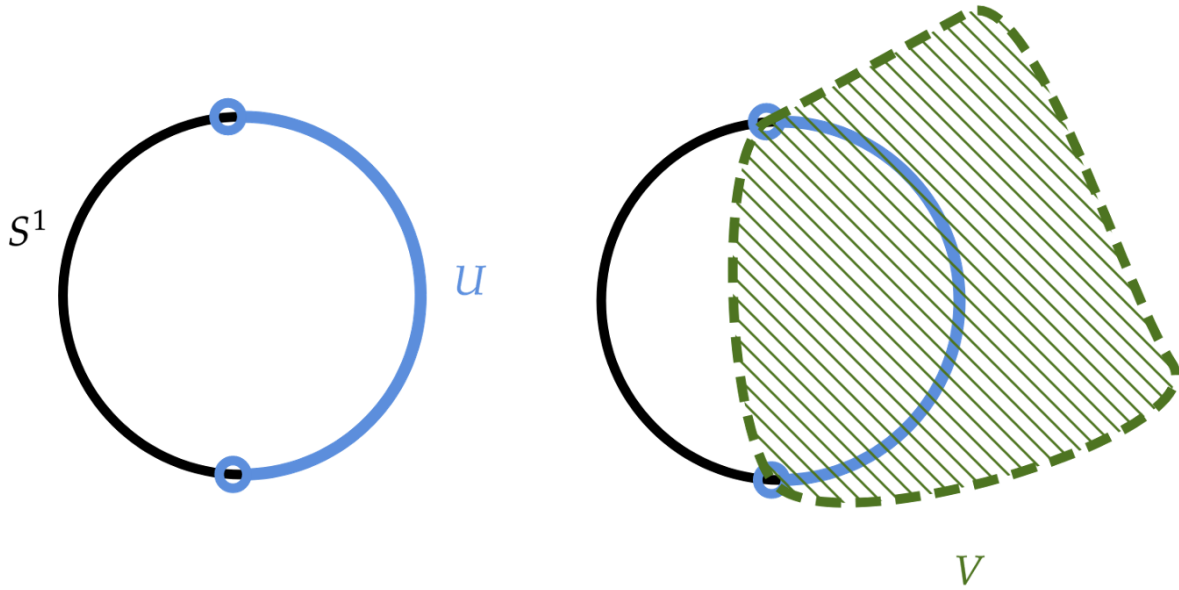


Figure 2.3: A set $U \subseteq S^1$ is open in the subset topology inherited from $\mathbb{R}^2$ if there exists a set $V \subseteq \mathbb{R}^2$ which is open in the topology on $\mathbb{R}^2$ such that $U = V \cap S^1$.

2.1.1.3 The Product Topology

Let $(S, \mathcal{O}_S)$ and $(T, \mathcal{O}_T)$ be topological spaces, and let $S \times T$ be the Cartesian product set whose elements are the ordered pairs (s, t) where $s \in S$ and $t \in T$. A subset $U \subseteq S \times T$ is open in the product topology if for all $p \in U$ there exists some $A \in \mathcal{O}_S$ and $B \in \mathcal{O}_T$ such that $p \in A \times B$ and $A \times B \subseteq U$.

The final method for constructing a topology which we will consider here is the *quotient topology* construction, which is particularly useful for constructing topologies on spaces which are obtained by identifying elements of other spaces as equivalent. For example, the circle S^1 can be obtained by starting from an interval $[0, 2\pi]$ and then identifying the points 0 and 2π - hence "gluing" the ends of the interval together.

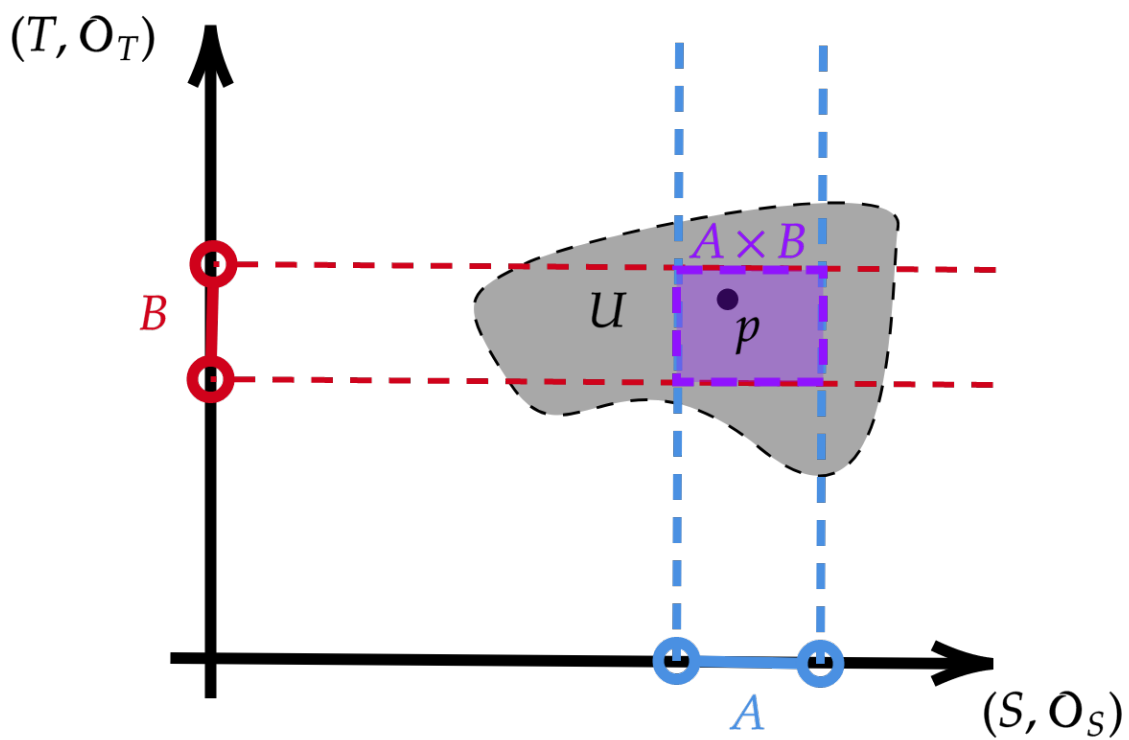


Figure 2.4: A set $U \subseteq S \times T$ is open in the product topology, if for all $p \in U$, there exist open sets $A \subseteq S$ and $B \subseteq T$ such that $p \in A \times B \subseteq U$.

2.1.1.4 The Quotient Topology

An *equivalence relation* on a set S , denoted by $\sim$, is a relation which satisfies the following conditions:

- $s \sim s$ for all $s \in S$ (reflexivity)
- $s \sim t \iff t \sim s$ for all $s, t \in S$ (symmetry)
- $s \sim t$ and $t \sim u$ implies that $s \sim u$ for all $s, t, u \in S$ (transitivity)

If $s \sim t$, then we say that s and t are *equivalent* under $\sim$.

Every equivalence relation on S partitions S into so-called *equivalence classes*. Given an element $s \in S$, the equivalence class $[s]$ is the subset of all elements of S which are equivalent to s . It's easy to show that each element of S corresponds to one and only one equivalence class. The set of all equivalence classes is denoted $S/\sim$ and is called the *quotient space* of S under $\sim$. Finally, the *quotient map* $q : S \rightarrow S/\sim$ is the map $s \mapsto [s]$ which takes each element of S to the corresponding equivalence class.

Having made these definitions, let $(S, \mathcal{O}_S)$ be a topological space and let $\sim$ be an equivalence relation on S . A subset $U \subseteq S/\sim$ is open in the *quotient topology*³ if its preimage⁴ under the quotient map q is open in the parent topology $\mathcal{O}_S$.

To flesh out the example mentioned previously, let $S = [0, 2\pi]$ and $\mathcal{O}_S$ be the topology inherited from the standard topology on $\mathbb{R}$. Define the equivalence relation $\sim$ on S such that for

³This construction actually works for an arbitrary function $f : A \rightarrow B$, and is more generally called the *final topology* on B inherited from f (where "final" refers to the fact that f takes points in A to some destination in B). The quotient topology is then the final topology applied to a quotient map which arises due to an equivalence relation.

⁴Given a map $f : A \rightarrow B$, the preimage of a set $U \subseteq B$, denoted $\text{preim}_f(U)$ or $f^{-1}(U)$, is the set of all $a \in A$ such that $f(a) \in U$.

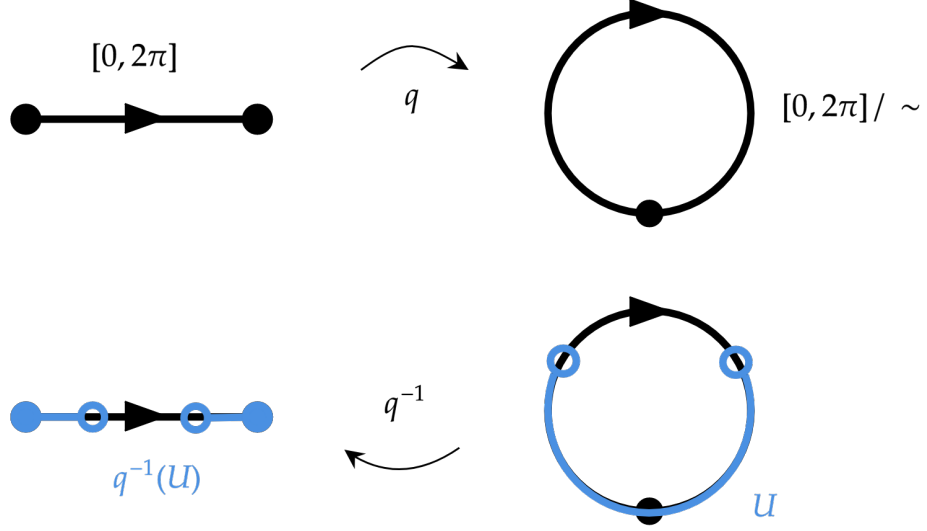


Figure 2.5: The set $[0, 2\pi]$ is mapped to the quotient set $[0, 2\pi]/\sim$ by the quotient map q . A set $U \subseteq [0, 2\pi]/\sim$ is open in the quotient topology if its preimage under q is open in the topology on the original space $[0, 2\pi]$.

all $s \in S, s \sim s$ and $0 \sim 2\pi$. Consider the set of points

$$U = \left\{ [s] \in S/\sim \mid s \in [0, \pi/2) \cup (3\pi/2, 2\pi] \right\}$$

The set $\text{preim}_q(U) = [0, \pi/2) \cup (3\pi/2, 2\pi]$ is open in the parent topology $\mathcal{O}_S$, and so U is open in the quotient topology.

We can find compatibility with our earlier constructions by noting that if we embed $S/\sim$ into $\mathbb{R}^2$ with the map $[s] \mapsto (\cos(s), \sin(s))$, then we obtain the familiar circle S^1 , where the quotient topology on $S/\sim$ coincides with the subset topology on S^1 inherited from the standard topology on $\mathbb{R}^2$.

2.1.2 Topological Ideas

Armed with the ability to endow a large variety of different spaces with topologies, we can now explore some of the ideas which a topology allows us to define.

2.1.2.1 Continuity

One of the most important ideas which arises in topology is that of *continuity*. A function f between topological spaces is continuous if small changes in the input yield correspondingly small changes in the output. Formally, let $(A, \mathcal{O}_A)$ and $(B, \mathcal{O}_B)$ be topological spaces, and let f be a map from $A \rightarrow B$. We say that f is *continuous at a point* $a \in A$ if, for any open set $V \ni f(a)$, there exists an open set $U \ni a$ such that $f(U) \subseteq V$. If f is continuous at each point in its domain, we simply say that f is a continuous function.

Intuitively, $f : A \rightarrow B$ is continuous at a if we may restrict the output to an arbitrarily small neighborhood of $f(a) \in B$ by restricting the input to an sufficiently small neighborhood of $a \in A$. If we *cannot* do this, then that is an indication that the function makes a "jump" at a .

In practice, an extremely useful alternative definition of continuity goes as follows. Given topological spaces $(A, \mathcal{O}_A)$ and $(B, \mathcal{O}_B)$, a function $f : A \rightarrow B$ is continuous if the preimage $f^{-1}(U)$ of all open sets $U \subseteq B$ are open - that is, $U \in \mathcal{O}_B \implies f^{-1}(U) \in \mathcal{O}_A$. It is an excellent exercise to demonstrate that this definition is equivalent to the more intuitive one given above.

2.1.2.2 Convergence

Another crucial idea which emerges from topology is that of *convergence*. A sequence $f : \mathbb{N} \rightarrow S$ is formally a map from the natural numbers $(0, 1, 2, \dots)$ to some other set S . By

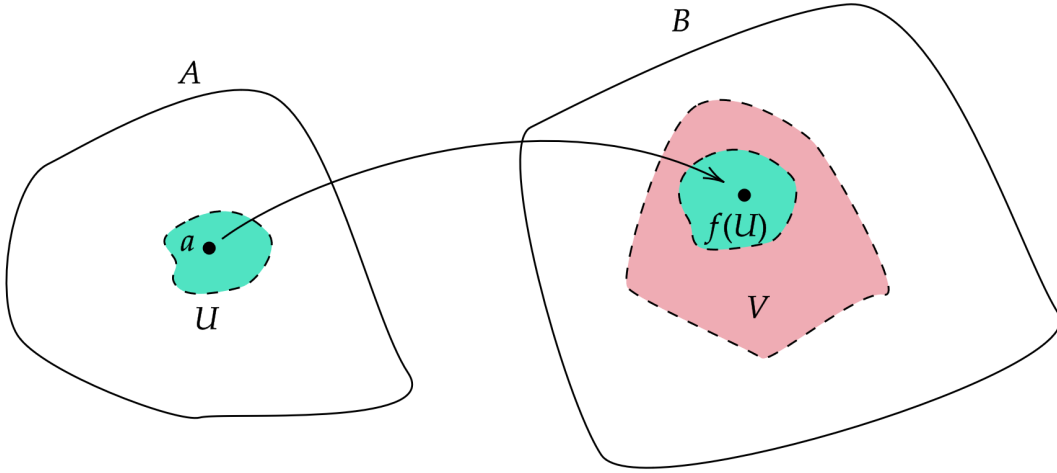


Figure 2.6: A function $f : A \rightarrow B$ is said to be continuous at $a \in A$ if its output can be restricted to an arbitrarily small open neighborhood $V \ni f(a)$ by restricting its *input* to a sufficiently small open neighborhood $U \ni a$.

convention, we often write $f(n) \equiv f_n$.

If $(S, \mathcal{O}_S)$ is a topological space, then we can talk about a sequence converging to a limit L . The formal definition goes as follows. A sequence $f : \mathbb{N} \rightarrow S$ converges to a limit $L \in S$ in the topology $\mathcal{O}_S$ if every open set containing L contains all but finitely-many points in the sequence. Alternatively, we say f converges to L if, given any open set $U \ni L$, there exist some $N \in \mathbb{N}$ such that $f_n \in U$ for all $n > N$. Proving that these definitions are equivalent is a useful exercise.

Intuitively, for $\{f_n\}$ to converge to L we want its elements to get arbitrarily close to L (where we define arbitrary close as belonging to an arbitrary open neighborhood of L) and stay there (which we ensure by demanding that past a certain N , all points f_n with $n > N$ remain in that neighborhood).

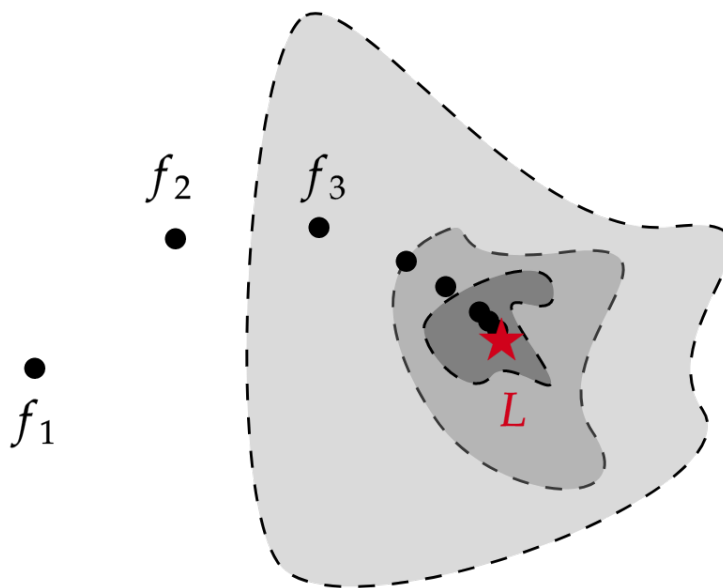


Figure 2.7: A sequence $\{f_n\}$ is said to converge to a limit L if every open set containing L contains all but finitely-many points in the sequence.

2.2 Topological Equivalence of Spaces

If $(A, \mathcal{O}_A)$ and $(B, \mathcal{O}_B)$ are topological spaces, it is worth asking whether there is a sense in which they are the same space.

For example, we first defined the circle S^1 as the topological space whose underlying set is the unit circle in $\mathbb{R}^2$ and whose topology is inherited by the subset topology construction. However, when discussing the quotient topology we started with the interval $[0, 2\pi]$, identified the endpoints as equivalent via an equivalence relation $\sim$, and then equipped $[0, 2\pi]/\sim$ with the quotient topology. We then claimed that this was an equivalent construction of the circle.

Obviously the underlying sets of the two constructions are totally different - the former consists of ordered pairs of $\mathbb{R}$ -numbers, while the latter consists of equivalent classes of elements of the interval $[0, 2\pi]$. However, from a *topological* perspective, all of their properties are exactly the same. The reason for this is the existence of the function $\phi : [s] \mapsto (\cos(s), \sin(s))$ which was mentioned during the discussion of the quotient topology. This function is invertible, and both ϕ and ϕ^{-1} are continuous; as a result, the open sets in the domain are in one-to-one correspondence with the open sets in the codomain. Any property which can be phrased entirely in terms of the open sets of a topological space must therefore be shared between the two.

Formally, let $(A, \mathcal{O}_A)$ and $(B, \mathcal{O}_B)$ be topological spaces and let $\phi : A \rightarrow B$ be a map between them. If ϕ is invertible and ϕ and ϕ^{-1} are both continuous, then ϕ is called a *homeomorphism* (or *topological isomorphism*). If there exists a homeomorphism between two topological spaces, then the spaces are called *homeomorphic* and they are equivalent from a topological standpoint.

The following are several examples of topological properties, their definitions in terms of

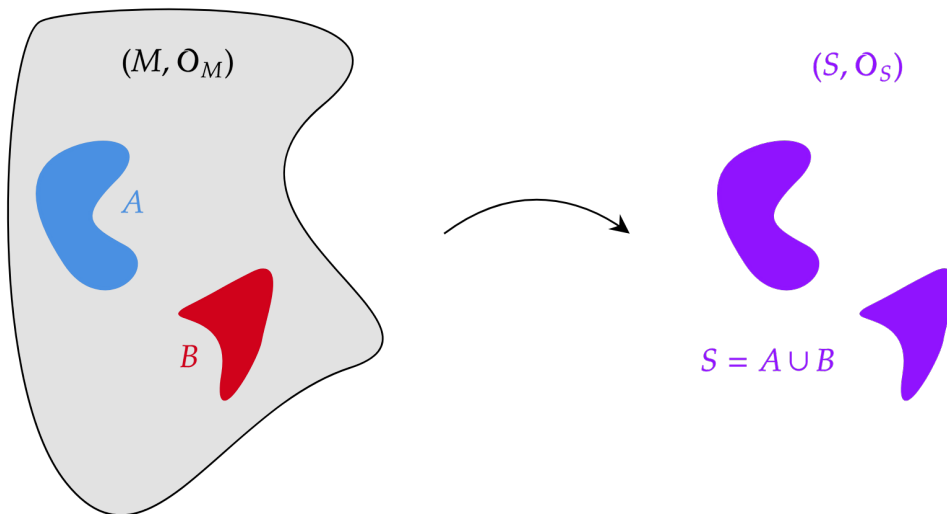


Figure 2.8: Given a parent space $(M, \mathcal{O}_M)$ and two subsets A and B , the subspace $(S, \mathcal{O}_S)$ (where $S = A \cup B$ and $\mathcal{O}_S$ is the subset topology) is disconnected if and only if $A \cap B = \emptyset$.

open sets, and an explicit demonstration that they are preserved under homeomorphisms:

2.2.0.1 Connectedness

A topological space $(S, \mathcal{O}_S)$ is called *disconnected* if it can be written as the union of two disjoint, non-empty open subsets. Otherwise, it is called *connected*. This definition can be motivated by the observation that if we start with an arbitrary space $(M, \mathcal{O}_M)$, take two non-empty subsets $U, V \subset M$, and let $S = U \cup V$ and $\mathcal{O}_S$ be the subset topology, then $(S, \mathcal{O}_S)$ is disconnected if and only if U and V are disjoint. That is, if $(S, \mathcal{O}_S)$ arises as a subspace of a larger space $(M, \mathcal{O}_M)$, then S is disconnected precisely if it can be written as disjoint pieces of M . The non-intuitive definition is given because it does not require any reference to a "parent" space.

Let $(A, \mathcal{O}_A)$ and $(B, \mathcal{O}_B)$ be topological spaces and let $f : A \rightarrow B$ be a homeomorphism.

If $B = U \cup V$ for non-empty $U, V \in \mathcal{O}_B$, then

$$A = f^{-1}(B) = f^{-1}(U \cup V) = f^{-1}(U) \cup f^{-1}(V) \quad (2.1)$$

where $f^{-1}(U)$ and $f^{-1}(V)$ are non-empty (because f is surjective) and open (because f is continuous). As a result, if B is (dis)connected then A is (dis)connected, and vice-versa.

2.2.0.2 Separation (T_0, T_1, T_2)

A topological space $(S, \mathcal{O}_S)$ is called:

- T_0 if for any two distinct points $s \neq t \in S$, there exists an open set U which contains s but not t **or** t but not s . This property means that any two points in S are topologically distinguishable; otherwise, the topology is too coarse to tell them apart. For an example of a topological space for which this fails, consider any set equipped with the indiscrete topology.
- T_1 if for any two distinct points $s \neq t \in S$, there exists an open set U containing s which does not contain t . Note the subtle distinction between this and T_0 - the latter admits the possibility that there are some points t such that the only open set containing t is the entire set S , while the latter does not. For an example of a space which is T_0 but not T_1 , consider any set equipped with the particular point topology.
- T_2 if for any two distinct points $s \neq t \in S$, there exist open sets $U_s \ni s$ and $U_t \ni t$ such that $U_s \cap U_t = \emptyset$. In such spaces, any two distinct points can be respectively enclosed within their own open sets which do not intersect. This is crucially important for convergence

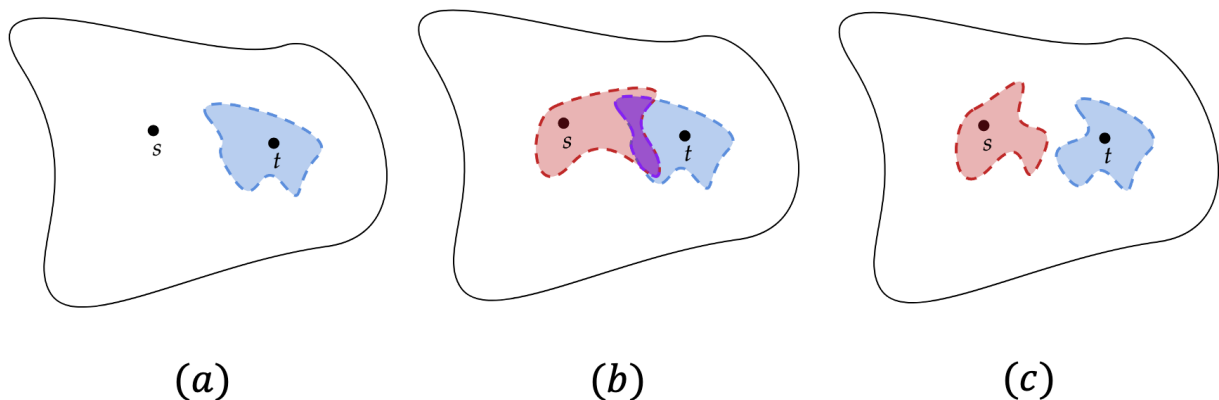


Figure 2.9: (a) In T_0 spaces, no two distinct points $s \neq t$ are in exactly the same open sets, but it's possible that every open set containing s also contains t (but not the reverse). (b) In T_1 spaces, we can always find open sets containing s but not t and t but not s , but it's possible that such sets necessarily intersect somewhere. (c) In T_2 spaces, we can always find open sets around any two distinct points which don't intersect at all..

because it implies and is required by the uniqueness of limits - a convergent sequence has a unique limit if and only if the space in question is T_2 . All topologies we will work with are T_2 , but for an example of a topology which is T_1 but not T_2 , consider any set equipped with the cofinite topology.

To prove that these properties are preserved under homeomorphism, let $(A, \mathcal{O}_A)$ and $(B, \mathcal{O}_B)$ be topological spaces and let $f : A \rightarrow B$ be a homeomorphism between them. Assuming that B has any of the properties listed above, simply replace sets and points with their preimages under f to prove that A has the same properties.

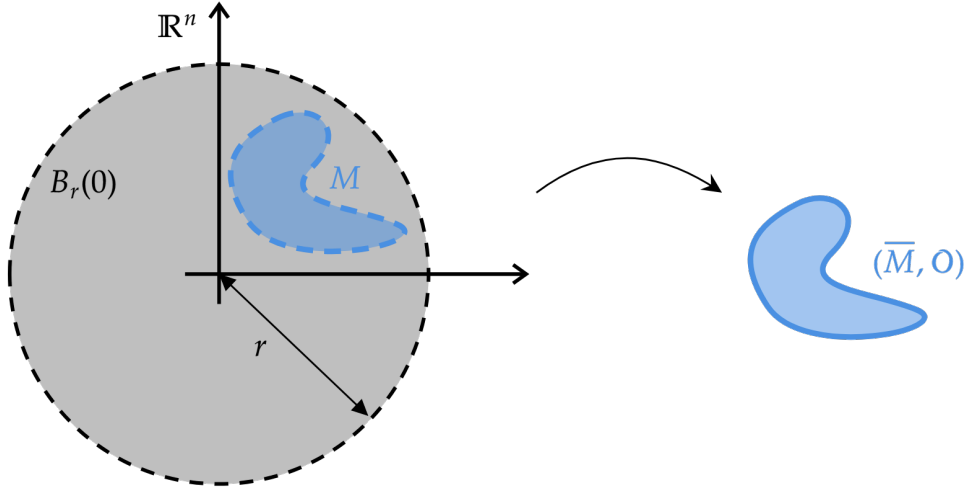


Figure 2.10: If M is a bounded subset of $\mathbb{R}^n$, then its closure $\overline{M}$ equipped with the subset topology is a compact space.

2.2.0.3 Compactness

An *open cover* C of a topological space $(S, \mathcal{O}_S)$ is a collection of open sets $\{U_1, U_2, \dots\}$ whose union is equal to S . A subcover C' of a cover C if $C' \subseteq C$ and C' is also a cover - intuitively we obtain C' from C by throwing away some of the U 's.

We say that $(S, \mathcal{O}_S)$ is *non-compact* if there exists an open cover $C = \{U_i\}$ which has no finite subcover - that is, there is no finite subset of these U 's which covers the space. Otherwise, we say that the space is *compact*. Some intuition for compactness is obtained by considering subsets of $\mathbb{R}^n$. A subset $M \subset \mathbb{R}^n$ is called *bounded* if there exists some $r > 0$ such that $M \subseteq B_r(0)$, i.e. M can be contained inside a ball of finite radius centered at the origin. If M is bounded then $(\overline{M}, \mathcal{O})$ is a compact space, where $\overline{M}$ is the closure of M and $\mathcal{O}$ is the induced subset topology.

Compact spaces have a number of very useful properties. For example, any $\mathbb{R}$ -valued

continuous functions on a metric space are bounded, and achieve both minimum and maximum values. More generally, the image of a compact space under a continuous function is also compact. More sophisticated operations such as integration are often better-behaved on compact spaces, and a number of important theorems such as the Chern-Gauss-Bonnet theorem hold only for compact spaces.

To see that this property is preserved under homeomorphism, let $(A, \mathcal{O}_A)$ and $(B, \mathcal{O}_B)$ be topological spaces and $f : A \rightarrow B$ be a homeomorphism. Assume that $(A, \mathcal{O}_A)$ is compact. Then any open cover $\{U_i\}$ of B can be mapped to an open cover $\{f^{-1}(U_i)\}$ of A . Because A is compact, this implies that this cover admits a finite subcover, which can be mapped via $f = (f^{-1})^{-1}$ to a finite open cover of B .

2.3 Homotopy

Sometimes we are interested not in the equivalence of different spaces, but rather in the equivalence of maps between spaces. For example, a global map of the magnetic field orientation could be understood as a function from S^2 (whose points represent positions on the surface of the Earth) to S^2 (whose points represent unit vectors which specify the direction of the magnetic field at the Earth's surface). Two different configurations can be viewed as topologically equivalent if it is possible to continuously deform one into the other (in a sense to be defined momentarily). If this is not possible, it suggests that the two configurations are distinct in some topological sense, and that a transition between the two is possible only by introducing a discontinuity in the map.

The concept at play here is called *homotopy*. Let $(A, \mathcal{O}_A)$ and $(B, \mathcal{O}_B)$ be topological spaces and let $f, g : A \rightarrow B$ be two continuous functions between them. If there exists a

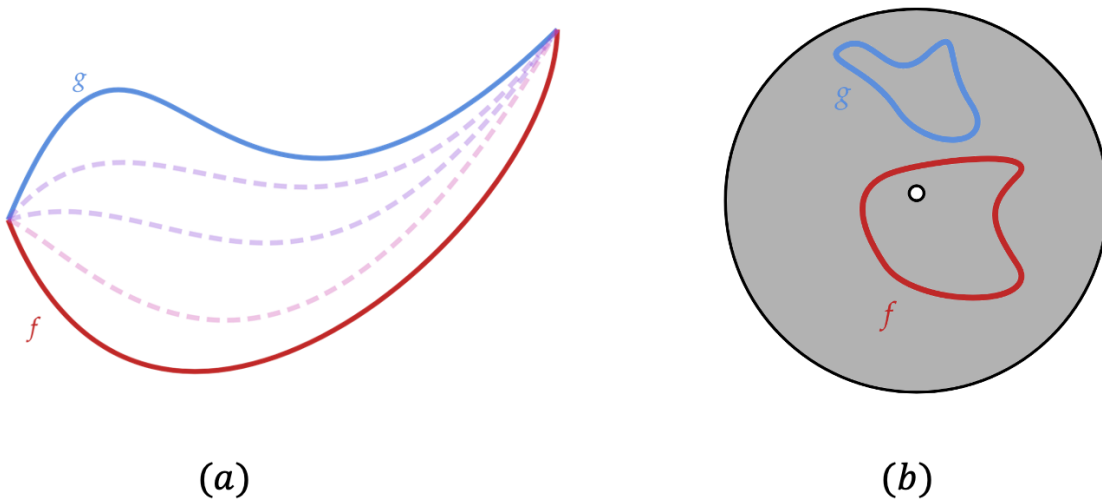


Figure 2.11: f and g are homotopic if they can be continuously deformed into one another. The curves in (a) are homotopic (with varying stages of the homotopy shown with dashed lines), but the curves in (b) are non-homotopic because they may not cross over the puncture in the center of the disk.

continuous map $h : [0, 1] \times A \rightarrow B$ such that $h(0, a) = f(a)$ and $h(1, a) = g(a)$, then it is called a *homotopy* between f and g , and f and g are called *homotopic*. The intuition for this is that the parameter $\lambda \in [0, 1]$ in the expression $h(\lambda, a)$ parameterizes the deformation of f into g ; as λ varies from 0 to 1, $h(\lambda, a)$ varies in a continuous way between $f(a)$ and $g(a)$.

As a trivial example, consider two functions $f, g : \mathbb{R} \rightarrow \mathbb{R}^2$ such that

$$f(t) = (t, 0) \quad g(t) = (t, t^2)$$

We can easily find a homotopy h between these maps:

$$h(\lambda, t) = (t, \lambda t^2) \tag{2.2}$$

Note that h can be written as $h(\lambda, t) = (1 - \lambda)f(t) + \lambda g(t)$. This raises the question as to why

we couldn't *always* do this - why can't we always find a homotopy between any two functions in a similar way?

The answer is that when we multiplied f and g by $(1 - \lambda)$ and λ and again when we added the results together, we made implicit use of the canonical vector space structure on $\mathbb{R}^2$, where $\lambda(a, b) = (\lambda a, \lambda b)$ and $(a, b) + (a', b') \equiv (a + a', b + b')$. If a space *does* have such a vector space structure, then indeed all continuous functions are homotopic to one another. However, this assumption fails for almost all spaces. Consider points on the unit circle - what does it mean to add them together, or to multiply them by scalars?

For a less trivial example, consider functions $f, g : [0, 2\pi] \rightarrow S^1$ such that

$$f(t) = (\cos(t), \sin(t)) \quad g(t) = (\cos(2t), \sin(2t))$$

These functions are also homotopic via the homotopy

$$h(\lambda, t) = (\cos([1 + \lambda]t), \sin([1 + \lambda]t))$$

Though this homotopy is slightly less trivial, it may not yet be clear why it is impossible to construct a similar homotopy for arbitrary functions. Functions from $[0, 2\pi] \rightarrow S^1$ can be imagined as continuous arcs on the unit circle (which may double back on themselves), and intuitively any such arc should be continuously deformable into any other. And indeed, this is also true, a fact which can be traced back to the triviality of the domain $[0, 2\pi]$.

To see an example of when it is not true, consider functions $f, g : S^1 \rightarrow S^1$. For ease of notation, we will identify S^1 with points z on the unit circle in the complex plane, so $z \in \mathbb{C}$ and

$|z| = 1$. Let

$$f(z) = 1 \quad g(z) = z^2 \quad (2.3)$$

I claim that it is *not* possible to produce a homotopy $h : [0, 1] \times S^1 \rightarrow S^1$ between these two functions. Intuitively, the image of f is a single point while the image of g wraps around S^1 , and it is not possible to continuously expand f to g (or shrink g to f) without leaving the unit circle or having an intermediate stage in which the loop is "cut" - but a continuous function whose domain is S^1 must form a closed loop at all times.

2.3.1 Homotopy Invariants

The intuitive argument is convincing, but it would be nice to have a rigorous proof which extends to more general functions and spaces. The problem lies in the difficulty of proving that it's impossible to construct something. The general way to do this is to define some quantity which is *homotopy-invariant*, meaning that if two functions are homotopic then the quantity must be the same for both of them. Once we have this, it follows that if two functions *differ* in the value of this quantity, then they must *not* be homotopic to one another (and so a homotopy cannot possibly be constructed).

2.3.1.1 The Winding Number

An extremely useful homotopy invariant for maps from S^1 to S^1 is given by the *winding number* ν , which counts the number of times the output of the function winds around the circle when the input winds around once. A complex number u with modulus 1 can be assigned an angle in the complex plane given by $\theta(u) = \log(u)/i$; along a contour γ , the total change in that

angle is given by

$$\Delta\theta = \int_{\gamma} d\theta = \frac{1}{i} \int_{\gamma} \frac{du}{u} \quad (2.4)$$

Since the angle changes by 2π for each winding, we divide by 2π to yield the winding number of the curve γ . If $u = f(z)$ is the image of a point $z \in S^1$ under the differentiable mapping f , then $du = f'(z)dz$ and it follows that the winding number associated to the function f is

$$\nu[f] = \frac{1}{2\pi i} \oint_{|z|=1} \frac{f'(z)}{f(z)} dz \quad (2.5)$$

To show that this is homotopy-invariant, consider two functions f, g and a homotopy h between them. We may write

$$\nu[h, \lambda] = \frac{1}{2\pi i} \oint_{|z|=1} \frac{\frac{d}{d\lambda} h(z, \lambda)}{h(z, \lambda)} dz$$

Because h is a continuous function of λ , ν must also be a continuous function of λ . However, we know that ν takes only integer values, while λ varies continuously from 0 to 1. The crucial insight⁵ is that the only way to have a continuous function of λ take only discrete values is for it to be constant - otherwise it will have a discontinuity somewhere. As a result, $\nu[f] = \nu[h, 0] = \nu[h, 1] = \nu[g]$.

Applying this to our previous example 2.3, we see immediately that $\nu[f] = 0$ and $\nu[g] = 1$, and so the two functions are not homotopic.

⁵Proving this is a nice exercise!

2.3.1.2 The Wrapping Number

The concept of a winding number can be extended to higher dimensions. Consider a continuous map $f : S^2 \rightarrow S^2$, where S^2 is the 2-sphere defined as

$$S^2 := \left\{ (a, b, c) \in \mathbb{R}^3 \mid a^2 + b^2 + c^2 = 1 \right\} \quad (2.6)$$

By way of analogy with the winding number, we begin by writing a normalized integral over the *area* of S^2 :

$$\frac{1}{4\pi} \oint \sin(\alpha) d\alpha d\beta \quad (2.7)$$

where $\alpha \in [0, \pi]$ and $\beta \in [0, 2\pi]$ are the polar and azimuthal angles, respectively. Letting $\alpha = \alpha(\theta, \phi)$ and $\beta = \beta(\theta, \phi)$, we find the wrapping number w of the map f to be

$$\begin{aligned} w[n] &= \frac{1}{4\pi} \oint \sin(\alpha) d\alpha d\beta \\ &= \frac{1}{4\pi} \oint \sin(\alpha(\theta, \phi)) \left(\frac{\partial \alpha}{\partial \theta} \frac{\partial \beta}{\partial \phi} - \frac{\partial \beta}{\partial \theta} \frac{\partial \alpha}{\partial \phi} \right) d\theta d\phi \end{aligned} \quad (2.8)$$

where $\frac{\partial \alpha}{\partial \theta} \frac{\partial \beta}{\partial \phi} - \frac{\partial \beta}{\partial \theta} \frac{\partial \alpha}{\partial \phi}$ is the Jacobian of the coordinate transformation $(\alpha, \beta) \mapsto (\theta, \phi)$.

From an intuitive standpoint, one can imagine a region $R \subset S^2$ with boundary ∂R and image $f(R)$. The image of the boundary $f(\partial R)$ traces out a loop in the target space of the function. As R grows to cover the entire domain, this loop shrinks down to a point and disappears. It could do this by wrapping around the target space once, or by wrapping partway around the target

space and doubling back on itself (think of a deflated balloon being draped over a basketball), or by wrapping itself around the unit sphere any integer number of times (with negative wrapping number corresponding to a reversal of orientation, where the image is turned ”inside out”).

2.3.1.3 The Degree of A Continuous Map

The construction we’ve used so far seems readily generalizable, which turns out to be true. Let M and N be smooth manifolds, which for the moment we will loosely describe as topological spaces which can be smoothly and consistently coordinatized⁶. The dimension of a smooth manifold is equal to the number of coordinates necessary to specify a point.

If M and N are compact, boundary-less smooth manifolds of the same dimension, then every continuous function $f : M \rightarrow N$ can be characterized by an integer homotopy invariant called its *degree*, which can be obtained by first writing down the natural volume⁷ integral over N (suitably normalized) and then taking its coordinates to be the images of points in M under the map f . If the volume of the space N is given by

$$\text{vol}(N) = \oint \omega(\mathbf{y}) dy^1 \dots dy^d \quad (2.9)$$

where ω is the so-called *volume form*⁸, then the degree of a function $f : \mathbf{x} \mapsto \mathbf{y}$ is given by

$$\text{deg}[f] = \frac{1}{\text{vol}(N)} \oint_M \omega(\mathbf{y}(\mathbf{x})) \det \left(\frac{\partial y^i}{\partial x^j} \right) dx^1 \dots dx^d \quad (2.10)$$

⁶A given coordinate system may only be valid on a particular patch of the manifold, but by coordinatizing each patch and smoothly transitioning between coordinate systems in the overlap region we can cover the entire manifold.

⁷Here we mean volume in a generalized sense - for a one- or two-dimensional manifold, we mean length and area, respectively.

⁸For our parameterization of S^1 , the volume form was 1; for S^2 , it was $\sin(\alpha)$.

For maps from $S^1 \rightarrow S^1$ and $S^2 \rightarrow S^2$, the degree of the map is given by the winding and wrapping numbers, respectively. More generally, one might consider the degree of a map between n -spheres, n -tori, and other similar spaces.

2.3.2 Homotopy Equivalence of Topological Spaces

The concept underlying homotopy is the continuous deformation of continuous maps into one another. Using this concept, one can speak of a kind of equivalence between the topological spaces themselves.

Let $(A, \mathcal{O}_A)$ and $(B, \mathcal{O}_B)$ be topological spaces, and let $f : A \rightarrow B$ and $g : B \rightarrow A$ be continuous maps. If $g \circ f$ is homotopic to id_A ⁹ and $f \circ g$ is homotopic to id_B , then the spaces are said to be *homotopy equivalent*.

Homotopy equivalence is a strictly weaker notion than that of topological equivalence (via a homeomorphism). For homotopy equivalence, we must be able to find some f and g such that that $g \circ f$ and $f \circ g$ are homotopic to id_A and id_B , respectively; if A and B are homeomorphic, it is possible to find f and g such that $g \circ f$ and $f \circ g$ are actually *equal* to id_A and id_B (in which case $g = f^{-1}$). Therefore, if two topological spaces are homeomorphic then they are necessarily homotopy equivalent, but the reverse is not true. With that being said, homotopy equivalence is often easier to work with in practice, and many topological invariants are also homotopy invariant.

2.3.2.1 Contractibility

An important notion for topological spaces is that of *contractibility*. A contractible space is a topological space which is homotopy-equivalent to a set with a single element, which means

⁹Given a set S , the function $\text{id}_S : S \rightarrow S$ is the identity map which maps each $s \in S$ to itself.

that it can be continuously shrunk down into a point. The interval $[-1, 1] \subset \mathbb{R}$ is contractible, but $[-1, 0) \cup (0, 1]$ is not. Similarly, the unit disk $D = \{(a, b) \in \mathbb{R}^2 \mid a^2 + b^2 < 1\}$ is contractible, but the punctured disk $D^* = \{(a, b) \in \mathbb{R}^2 \mid 0 < a^2 + b^2 < 1\}$ is not.

More generally, let $(S, \mathcal{O}_S)$ be a topological space and $P = \{0\}$ be a topological space with one element¹⁰. The only map from $S \rightarrow P$ is $f : s \mapsto 0$, which is easily seen to be continuous. Let $g : P \rightarrow S$ be a map defined by $g(0) = s_0$ for some arbitrarily chosen point $s_0 \in S$, which is also easily shown to be continuous. The map $g \circ f$ is the constant map which takes $s \mapsto s_0$, while the map $f \circ g$ is the identity map which takes $0 \mapsto 0$. Therefore, by definition $(S, \mathcal{O}_S)$ is homotopy equivalent to P (and is therefore contractible) if and only if the identity map $\text{id}_S : s \mapsto s$ is homotopic to any constant map $s \mapsto s_0$. A map which is homotopic to a constant map is called *null-homotopic*.

As a simple example, let $S = [-1, 1] \subset \mathbb{R}$ and $s_0 \in S$. The constant map id_S is homotopic to $f : s \mapsto s_0$ via the following homotopy:

$$h(\lambda, s) = \lambda s_0 + (1 - \lambda)s \quad (2.11)$$

As a result, $[-1, 1]$ is contractible.

To see the importance of contractibility, consider two spaces $(A, \mathcal{O}_A)$ and $(B, \mathcal{O}_B)$ and a continuous map $f : A \rightarrow B$ between them. If $(A, \mathcal{O}_A)$ is contractible, then there exists some $h_A : [0, 1] \times A \rightarrow A$ such that $h_A(0, a) = a$ and $h_A(1, a) = a_0$ for some constant $a_0 \in A$. Therefore, the map $f \circ h_A : [0, 1] \times A \rightarrow B$ has the property that $(f \circ h_A)(0, a) = f(a)$ and $(f \circ h_A)(1, a) = f(a_0)$, meaning that f is homotopic to a constant map (i.e. it is null-homotopic).

¹⁰Up to homeomorphism, there is only one topological space with one element, with the obvious (discrete) topology.

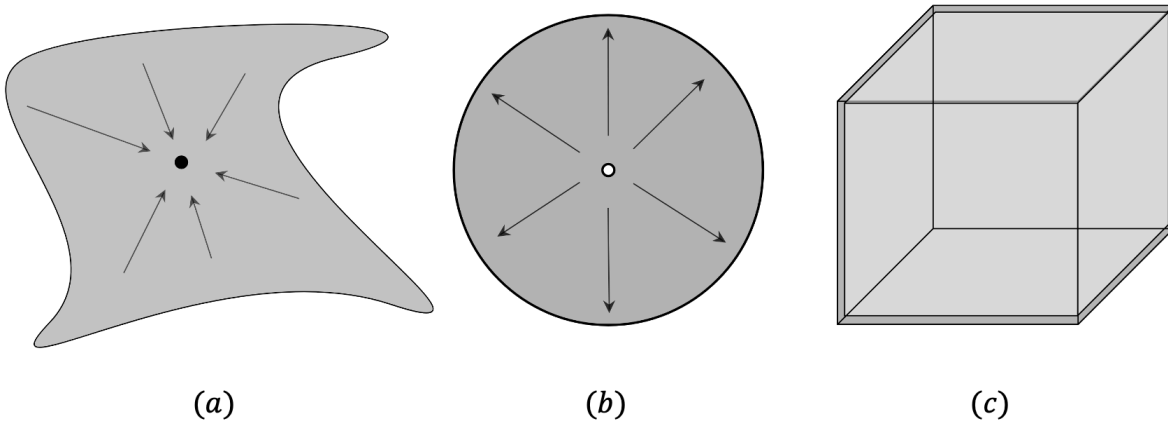


Figure 2.12: (a) A contractible space is homotopy-equivalent to a point. (b) The punctured disk is not contractible, but it is homotopy-equivalent to S^1 (i.e. its boundary). (c) The surface of a hollow cube is homotopy equivalent to S^2 (the surface of a ball).

If this is the case, then all of its homotopy-invariant properties (e.g. its wrapping number or, more generally, topological degree) must be trivial. Similarly, if $(B, \mathcal{O}_B)$ is contractible then there exists some $h_B : [0, 1] \times B \rightarrow B$ such that $h_B(0, b) = b$ and $h_B(1, b) = b_0$ for some $b_0 \in B$. From there, $h_B \circ f$ constitutes a homotopy between f and a constant map on b , and once again we find that f must be null-homotopic.

The conclusion, then, is that for a continuous map $f : A \rightarrow B$ to have non-trivial homotopy-invariant properties, both $(A, \mathcal{O}_A)$ and $(B, \mathcal{O}_B)$ must be non-contractible. In essence, contractible spaces are too trivial for maps to or from them to have interesting properties (at least through the lens of homotopy).

2.3.2.2 Homotopy Groups

If a space is *not* contractible, then it is because there is some kind of obstruction which prevents us from continuously shrinking it down to a point. The examples $[-1, 0) \cup (0, 1] \subset \mathbb{R}$

and the punctured disk cannot be shrunk down because they have holes in them. The same is true of the circle S^1 , the 2-sphere S^2 , and more generally the n -spheres¹¹. The homotopic non-triviality of a space can, in essence, be associated to the presence and nature of these holes.

Consider a topological space $(A, \mathcal{O}_A)$. We can define a *loop* in A to be a continuous map $f : S^1 \rightarrow A$. It is convenient to take S^1 to be constructed from interval $[0, 2\pi]$ with its endpoints glued together, in which case we can think of a loop as a continuous map $f : [0, 2\pi] \rightarrow A$ with $f(0) = f(2\pi)$. We call the point $f(0) \equiv a \in A$ the *base point* of the loop.

Two loops f and g which share the same base point can be concatenated to form a third loop $f * g$, which follows f and then subsequently g . Concretely, we define

$$(f * g)(t) = \begin{cases} f(2t) & 0 \leq t < \pi \\ g(2(t - \pi)) & \pi \leq t \leq 2\pi \end{cases} \quad (2.12)$$

Let $\mathcal{L}_a$ be the set of all loops with base point a . We define the equivalence relation $\sim$ on $\mathcal{L}_a$ such that $f \sim g \iff f$ and g are homotopic. Therefore, the equivalence classes $[f]$ consist of families of loops which can be continuously deformed into one another. We can define a kind of composition operation $\circ$ on $\mathcal{L}_a / \sim$ by letting $[f] \circ [g] := [f * g]$; it can be shown straightforwardly that this respects the equivalence relation. Equipped with this operation, the quotient space $\mathcal{L}_a / \sim$ has the structure of a *group* which we call $\pi^1(A; a)$ - the first homotopy group of A based at a .

The unit disk $D \subset \mathbb{R}^2$ is contractible, and so every loop based at any $d_0 \in D$ is homotopic to the constant loop $f : t \mapsto d_0$. This implies that every loop based at d_0 is homotopic to every

¹¹The n -sphere S^n is given by the set $\{(x_1, x_2, \dots, x_{n+1}) \in \mathbb{R}^{n+1} \mid \sum_{i=1}^{n+1} x_i^2 = 1\}$ equipped with the topology inherited from $\mathbb{R}^{n+1}$.

other, and therefore equivalent to every other under the equivalence relation $\sim$. As a result, $\pi^1(D; d_0)$ is the trivial group with a single element, which we may write as $\pi^1(D; d_0) \simeq \{0\}$. Clearly this is true for any contractible space. In general, if π_1 is trivial then we call the space *simply connected*¹².

On the other hand, the punctured disk $D^* \subset \mathbb{R}^2$ is *not* contractible. The constant loop $t \mapsto d_0 \in D^*$ is not homotopic to a loop which encircles the origin, which in turn is not homotopic to a loop which encircles the origin twice. It turns out that the equivalence classes $[f] \in \pi^1(D^*; d_0) \simeq \mathbb{Z}$, where the association $[f] \leftrightarrow n \in \mathbb{Z}$ means that every loop in $[f]$ winds¹³ around the origin n times.

The general homotopy groups $\pi_n(A; a)$ are constructed analogously to $\pi_1(A; a)$, with the only difference being that the continuous maps in question are from $S^n \rightarrow A$. The 0-sphere S^0 is the set of points $x \in \mathbb{R}$ such that $x^2 = 1$ - that is, the set $\{-1, 1\}$. A map $f : \{-1, 1\} \rightarrow A$ with base point a must have $f(1) = a$, and so is completely characterized by $f(-1)$. Two such maps f, g are homotopic if there exists a homotopy h such that $h(0, -1) = f(-1)$ and $h(1, -1) = g(-1)$ - that is, if there exists a continuous path $p : [0, 1] \rightarrow A$ such that $p(0) = f(-1)$ and $p(1) = g(-1)$. This property is called *path-connectedness*; a generic space can be decomposed into pieces which are each path-connected, and $\pi_0(A; a)$ counts the number of such pieces.

One can show that if A is path-connected, then the homotopy groups $\pi_n(A; a)$ are independent of a , and in such cases the groups are often written as $\pi_n(A)$. However, in general this is not possible, and the homotopy groups of a generic space *do* depend on the chosen base point.

¹²Observe that contractibility implies simple-connectedness, but simple-connectedness does not imply contractibility

¹³We can define a generalized winding number about the origin (i.e. the puncture). Because the winding number is a homotopy invariant, $f \sim g \implies \nu[f] = \nu[g]$, and in this particular example the reverse is also true. As a result, the equivalence classes in $\pi_1(D^*; d_0)$ can be precisely identified by their winding numbers.

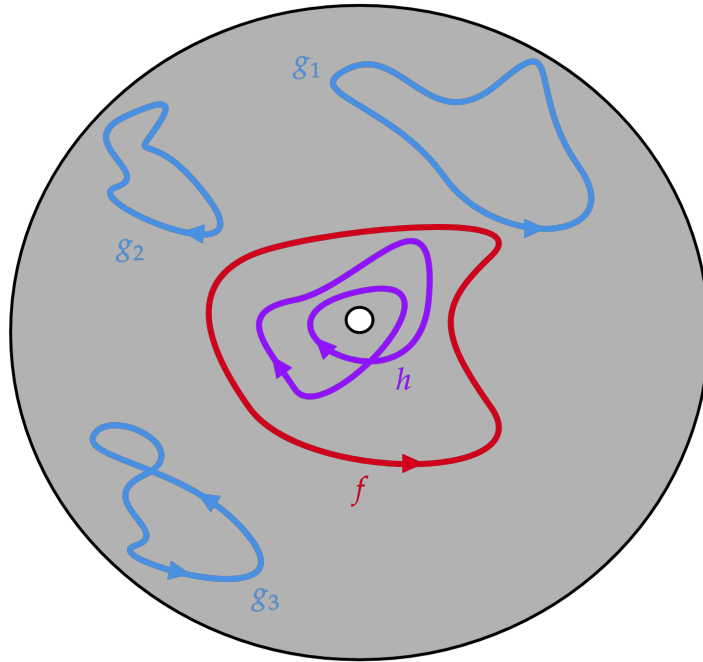


Figure 2.13: The punctured disk D^* has a non-trivial fundamental group given by $\pi_1(D^*; d_0) \simeq \mathbb{Z}$ for arbitrary d_0 . Curve f belongs to the equivalence class with winding number $+1$. g_1, g_2, g_3 are all homotopic and belong to the class with winding number 0 . h belongs to the class with winding number -2 , with the minus sign originating from the clockwise orientation.

2.3.2.3 Interplay between Homotopy Groups and Invariants

As one might be able to suspect by now, the homotopy groups π_n of a given space (and base point) are intimately related to the types of homotopy-invariant quantities which that space may support. As an example, π_1 (also called the *fundamental group*) describes the homotopy-equivalent families of loops which can exist in the space, and therefore constrains the types of winding numbers which can be defined. $\pi_1(D^*) \simeq \pi_1(S^1) \simeq \mathbb{Z}$, which means that these spaces support integer-valued winding numbers. For the 2-torus¹⁴, $\pi_2(\mathbb{T}^2) \simeq \mathbb{Z} \times \mathbb{Z}$ (loops may wind an integer number of times around the major (exterior) radius and an integer number of times around the minor (interior) radius, and so it supports *two* integer-valued winding numbers. The somewhat exotic space $\mathbb{RP}^3$ - which can mysteriously be associated with 3-dimensional rotations - has fundamental group $\pi_1(\mathbb{RP}^3) \simeq \mathbb{Z}/2\mathbb{Z}$ (the group of integers modulo 2), meaning that there are only two distinct classes of loops on this space. This is intimately related to the fact that there are only two types of fundamental particle (fermionic and bosonic) in three spatial dimensions.

¹⁴The 2-torus $\mathbb{T}^2$ is the surface of a doughnut.

Chapter 3: Physical Applications of Topology I

Much of the content of this chapter can be learned from the excellent overview of the topologically ordered media by Mermin [18]. His approach is somewhat more sophisticated and powerful, but it comes at the cost of working with quotient groups from the start. It would be an excellent idea to study his review if the material in this chapter is deemed interesting.

3.1 Foundational Ideas

An immediate application of topology arises when we seek to classify ordered media - physical systems occupying some region $R \subseteq \mathbb{R}^d$ which can be described by a continuous, spatially-varying quantity called the *order parameter*. The order parameter associates to each point $\mathbf{r} \in R$ a value which lies in some parameter space $\mathcal{M}$. The archetypal example of an order parameter arises in a magnetic material, in which each spatial point $\mathbf{x}$ is associated to the local magnetization $\mathbf{m}(\mathbf{x}) \in \mathbb{R}^3$; other examples include the translations, rotations, and distortions of infinitesimal elements of a crystalline solid, and the orientations of molecules in a liquid crystal.

The topological classification of various configurations of the order parameter (which we will refer to as *states*) is built on the idea that not all continuous states can be continuously deformed into one another; in the language of topology, not all continuous states (which are ultimately functions from $R \rightarrow \mathcal{M}$) are homotopic. Given a choice of domain R and parameter

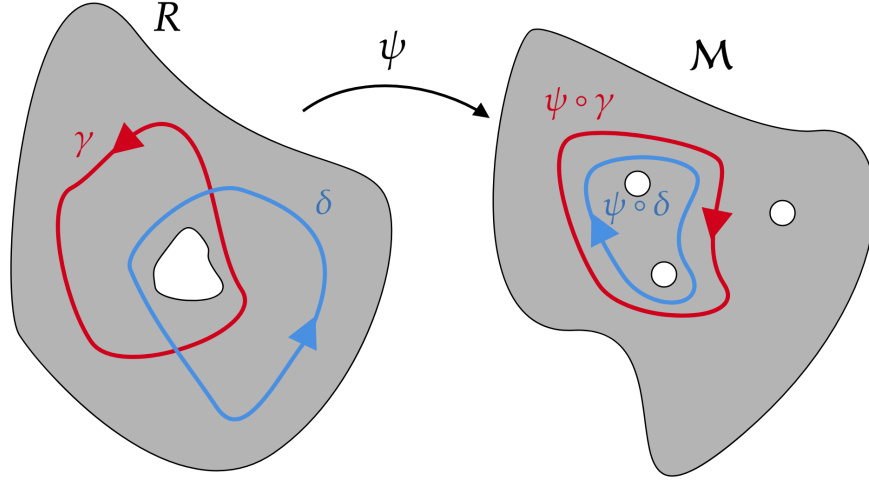


Figure 3.1: The homotopic equivalence of n -dimensional loops is preserved under the action of a continuous state $\psi : R \rightarrow \mathcal{M}$, which motivates the definition of the maps $\psi_n^* : \pi_n(R) \rightarrow \pi_n(\mathcal{M})$ which map families of equivalent n -loops in R to corresponding families in $\mathcal{M}$.

space M , we would like to decompose the space of all continuous states¹ into homotopically equivalent families.

Given a state $\psi : R \rightarrow \mathcal{M}$, any n -dimensional loop $\gamma : S^n \rightarrow R$ in the domain can be mapped to an n -dimensional loop $\psi \circ \gamma : S^n \rightarrow \mathcal{M}$ in the parameter space, as in Fig. 3.1. It's not difficult to show that if two loops γ, δ in R are homotopic, then the loops $\psi \circ \gamma$ and $\psi \circ \delta$ in $\mathcal{M}$ are homotopic. As a result, for each non-negative integer n , a state ψ induces a map $\psi_n^* : \pi_n(R) \rightarrow \pi_n(\mathcal{M})$ which acts as follows. Given any equivalence class $C \in \pi_n(R)$, choose a representative loop $\gamma \in C$ and map it to $\mathcal{M}$ via ψ . The output of function is the corresponding equivalence class in $\pi_n(\mathcal{M})$. Symbolically, let $[\gamma]$ be the equivalence class to which the loop γ

¹Note that we may consider only those states which are compatible with some prescribed boundary conditions.

belongs; then we may write

$$\begin{aligned}\psi_n^\star : \pi_n(R) &\rightarrow \pi_n(\mathcal{M}) \\ [\gamma] &\mapsto [\psi \circ \gamma]\end{aligned}\tag{3.1}$$

If two states ψ and ϕ are homotopic, then it is straightforward to show that $\psi_n^\star$ and $\phi_n^\star$ are the same maps for all n . This implies that if $\psi_n^\star$ and $\phi_n^\star$ are *different* (for any n) then ψ and ϕ cannot be homotopic. This provides the foundation for our classification scheme. Once we have specified R and $\mathcal{M}$, we may systematically ask what kinds of maps we can have from $\pi_n(R) \rightarrow \pi_n(\mathcal{M})$ for each n , and distinguish different states based on the results.

In particular, note that if $\pi_n(R) = \{0\}$ or $\pi_n(\mathcal{M}) = \{0\}$, then the only possible map from $\pi_n(R) \rightarrow \pi_n(\mathcal{M})$ is the zero map. In order to have a meaningful classification scheme for a given n , it is necessary (and, as it turns out, sufficient) that $\pi_n(R)$ and $\pi_n(\mathcal{M})$ be non-trivial. In the following, we restrict our attention to $n \leq 2$, while remarking that there is room for some very interesting research surrounding the application of higher homotopy groups to the classification of physical systems.

3.1.1 A Mathematical Example

To make this rather abstract discussion concrete, we will consider a visual (but still purely mathematical) example before moving on to a more in-depth discussion of physical phenomena. Let $R = D^\star$ be the punctured unit disk and $\mathcal{M} = \mathbb{C}^{\star\star}$ be the doubly-punctured complex plane, defined as

$$\mathbb{C}^{\star\star} := \{z \in \mathbb{C} \mid z \neq \pm 1\}$$

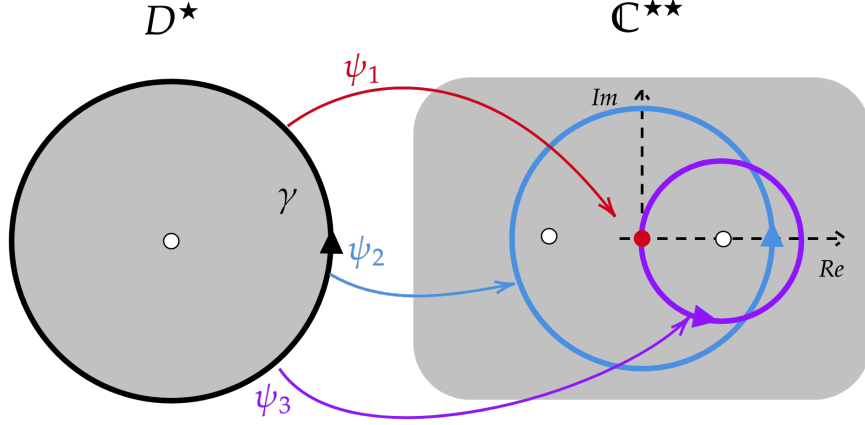


Figure 3.2: An example of three states $\psi_1, \psi_2, \psi_3 : R \rightarrow \mathcal{M}$ and their corresponding topological indices (ν_+, ν_-) . Under our classification scheme, ψ_1 has indices $(0, 0)$, ψ_2 has indices $(1, 1)$, and ψ_3 has indices $(1, 0)$.

as shown in Fig. 3.2. Note that $\pi_1(D^*) \simeq \mathbb{Z}$ and $\pi_1(\mathbb{C}^{**}) \simeq \mathbb{Z} \times \mathbb{Z}$, so for $n = 1$ we have a nontrivial classification scheme.

Loops in D^* are classified by an integer winding number which describes how many times they encircle the central puncture (with positive n denoting a counterclockwise loop). Loops in $\mathbb{C}^{**}$ are classified by a pair of integer winding numbers which describe how many times they encircle each of the two punctures. Therefore, we may classify a state ψ by choosing a loop γ in D^* with winding number 1 (e.g. the positively-oriented unit circle) and assigning it a pair of integers (ν_+, ν_-) which denote the respective winding numbers around the punctures at $+1$ and -1 in $\mathcal{M}$. It is an instructive exercise to work out that those winding numbers are given by

$$\nu_{\pm} = \frac{1}{2\pi} \int_0^{2\pi} d\theta \frac{d}{d\theta} \text{Im} \left[\log \left(\frac{\psi(r, \theta) \mp 1}{|\psi(r, \theta) \mp 1|} \right) \right] \quad (3.2)$$

where (r, θ) are the coordinates on D^* and $\log$ is the natural logarithm. To see this classification

scheme at work, consider the following three states:

$$\begin{aligned}\psi_1(r, \theta) &= 0 \\ \psi_2(r, \theta) &= (1 + r)e^{i\theta} \\ \psi_3(r, \theta) &= 1 + e^{-r}e^{i\theta}\end{aligned}\tag{3.3}$$

It is straightforward to verify that our classification scheme labels ψ_1 with $(0, 0)$, ψ_2 with $(1, 1)$, and ψ_3 with $(1, 0)$, which matches the picture shown in Fig. 3.2.

3.1.2 An Important Note Regarding Stability

Before studying physical examples, it's important to note that topological classification can have implications about the dynamical stability of various structures in ordered media, but these implications are often subtle. Topological stability and dynamical stability are often improperly conflated, and so disentangling these notions is crucially important to physical applications.

It is always worth remembering that modeling states of an ordered system as continuous maps is always, on some level, an approximation. Sometimes the approximation arises from some form of coarse-graining procedure, in which e.g. spins on a lattice may be approximated as a continuous function on a line in the suitable thermodynamic limit. Other times, the approximation arises when we assert that a many-particle system can be described as a spatially coherent condensate, as in superfluids or superconductivity.

In a generic ordered medium, there is an energy cost associated with spatial variation in the order parameter, such that the introduction of a discontinuity would require a formally infinite

amount of energy. Since transitions between states which are topologically inequivalent requires such a discontinuity, one might think that such transitions are energetically forbidden, but this reasoning holds only as long as our underlying assumptions do.

Once the introduction of discontinuities in an order parameter becomes physically allowed due to the breakdown of those assumptions, any topologically non-trivial state gains a viable pathway by which it can be annihilated to a trivial state. That pathway generally takes the form of a continuous deformation to a regime in which the assumption of continuity becomes questionable, followed by a transition to an energetically favorable trivial state. Therefore, the stability of a topological state depends on the presence of an energy barrier which prevents this pathway from being followed. With that in mind, the relationship between topological non-triviality and dynamical stability is something which must be examined on a case-by-case basis, as we will do in the following two examples.

3.2 Magnetic Skyrmions

3.2.1 Homotopic Preliminaries

If we model a d -dimensional magnetic medium with an order parameter $\mathbf{m} : \mathbb{R}^d \rightarrow \mathbb{R}^3$, then because $\mathcal{M} = \mathbb{R}^3$ is contractible, there are no topologically distinct configurations of $\mathbf{m}$. However, if we require that the magnetization be everywhere non-vanishing, then the parameter space becomes $\mathcal{M} = \mathbb{R}^3 - \{\vec{0}\}$, which is homotopic to S^2 (see Fig. 3.3). In less technical terms, if $\mathbf{m}(\mathbf{x}) \neq 0$ for all x , then we may consider the alternative order parameter $\mathbf{n} : \mathbb{R}^d \rightarrow S^2$ where $\mathbf{n}(\mathbf{x}) := \mathbf{m}(\mathbf{x})/|\mathbf{m}(\mathbf{x})|$ is a unit vector defined by a point on the unit sphere.

The first two homotopy groups of the 2-sphere are well-known and can be understood

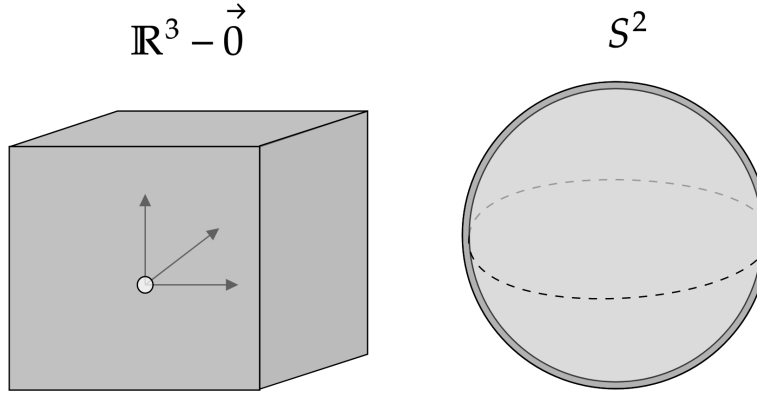


Figure 3.3: The punctured space $\mathbb{R}^3 - \vec{0}$ is homotopic to the 2-sphere S^2 . On this domain, point defects take the form of monopole configurations, and our classification scheme essentially counts the number of monopoles which are present at the location of the puncture.

intuitively:

- S^2 is simply-connected, which means that its fundamental group is $\pi_1(S^2) = \{0\}$, and all closed loops on S^2 can be deformed into one another².
- The second homotopy group is given by $\pi_2(S^2) = \mathbb{Z}$, suggesting that the wrapping number may be a viable classification tool.

This suggests that if we restrict our attention to materials with a non-vanishing local magnetization, then defects may be characterized by some kind of wrapping number. One example is point singularities in three dimensional systems; taking the singularity to occur at the coordinate origin, the region R on which the magnetization $\mathbf{m}$ is continuous and non-vanishing is given by $R = \mathbb{R}^3 - \{\mathbf{0}\}$, again homotopic to S^2 . This suggests surrounding the singularity with a spherical surface, defining a map $\mathcal{N} : S^2 \rightarrow S^2$ to be the restriction of n to the aforementioned surface, and computing the degree of $\mathcal{N}$. Physically, $\deg(\mathcal{N})$ would be the normalized flux of $\mathbf{m}$ through the

²This is often intuitively phrased as "you cannot lasso a basketball."

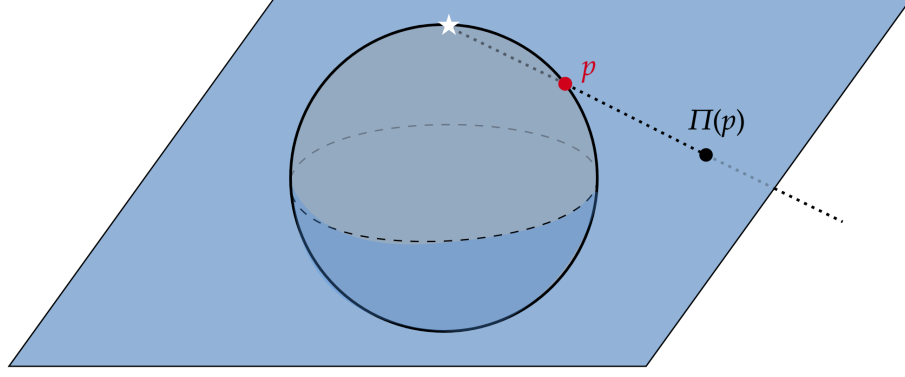


Figure 3.4: The stereographic projection map Π provides a continuous, one-to-one correspondence between points on S^2 and points in $\mathbb{R}^2 \cup \{\infty\}$, where the so-called *point at infinity* is associated to the north pole of the sphere.

surface, and could be called the *monopole number*.

3.2.2 2D Magnetic Skyrmions

Alternatively, consider a 2D magnetized film which is asymptotically uniform ($\lim_{|\mathbf{r}| \rightarrow \infty} \mathbf{n}(\mathbf{r})$ is well-defined and unique along all directions) along $\hat{z}$, and whose magnetization direction $\mathbf{n}$ is a continuously-varying function from $\mathbb{R}^2$ to S^2 . As shown in Fig. 3.4, we may define the stereographic projection $\Pi : S^2 \rightarrow \mathbb{R}^2 \cup \{\infty\}$ where the north pole of the sphere is mapped to the so-called "point at infinity"³ From there, the function $\mathcal{N} := \mathbf{n} \circ \Pi$ is a map from $S^2 \rightarrow S^2$; note that $\mathbf{n}(\infty)$ is well-defined because of the asymptotically uniform magnetization.

As per our classification scheme, the only non-trivial map between homotopy groups is $\pi_2(\mathbb{R}) \rightarrow \pi_2(\mathcal{M})$. A continuous state $\mathbf{n} : \mathbb{R}^2 \cup \{\infty\} \rightarrow S^2$ can then be classified precisely by the

³Because Π provides a homeomorphism between S^2 and $\mathbb{R}^2 \cup \{\infty\}$, the two spaces must share all topological properties - in particular, compactness. For this reason, $\mathbb{R}^2 \cup \{\infty\}$ is called a (*one-point*) *compactification* of $\mathbb{R}^2$.

degree of $\mathcal{N} \equiv \mathbf{n} \circ \Pi$, given explicitly by

$$\deg(\mathcal{N}) := \frac{1}{4\pi} \iint_{\mathbb{R}^2} dx dy \, \mathbf{n} \cdot (\partial_x \mathbf{n} \times \partial_y \mathbf{n}) \quad (3.4)$$

In the present context, we refer to this integer $\Sigma \equiv \deg(\mathcal{N})$ as the *skyrmion number*⁴ of the magnetic texture. As long as Σ is well-defined, it cannot be changed by continuous variation in $\mathcal{N}$. That is, we might parameterize the magnetization with a continuous variable λ such that for each $\lambda \in \mathbb{R}$, $\mathcal{N}_\lambda$ corresponds to a different function. If $\mathcal{N}_\lambda$ is a continuous function of λ , then we have that

$$\frac{d}{d\lambda} \Sigma_\lambda \equiv \frac{d}{d\lambda} \deg(\mathcal{N}_\lambda) = 0 \quad (3.5)$$

Physically, this implies that magnetic textures with different skyrmion numbers cannot be deformed into one another without introducing a discontinuity in the magnetization direction. As such, Σ constitutes a topological invariant, and so textures with $\Sigma = 0$ are referred to as (*topologically*) *trivial* while textures with $\Sigma \neq 0$ are referred to as (*topologically*) *non-trivial*. The topologically non-trivial magnetization textures obtained here are called *magnetic skyrmions*.

3.2.3 Skyrmion Stability

As mentioned in Sec. 3.1.2, fact that magnetic skyrmions cannot be created or deleted without introducing a discontinuity into the magnetization is often referred to as *topological protection*, and taken as an indication that these configurations are stable against perturbations and thermal noise. However, it should be noted that stability is really a statement about energetics, and topology alone is insufficient to determine whether a particular configuration is energetically

⁴The name is in deference to Tony Skyrme, who developed the relevant ideas in the context of nuclear physics.

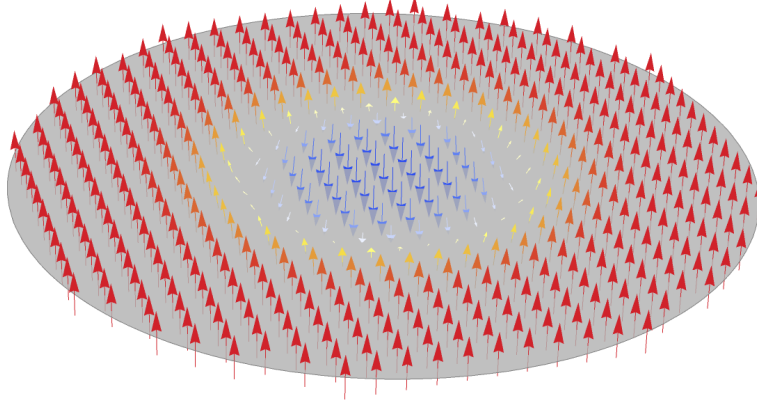


Figure 3.5: An example of a magnetic texture with skyrmion number $\Sigma = +1$.

stable.

3.2.3.1 Energy of Magnetic Structures

An axisymmetric magnetization texture $\mathbf{m}$ can be described using cylindrical coordinates (r, φ, z) as

$$\mathbf{m}(r, \varphi, z) = m_0 \begin{pmatrix} \sin(\theta(r)) \cos(\varphi + \psi) \\ \sin(\theta(r)) \sin(\varphi + \psi) \\ \cos(\theta(r)) \end{pmatrix} \quad (3.6)$$

where $\theta(r)$ is the polar angle (measured from $\hat{z}$) of the magnetization direction and ψ is the so-called *domain wall angle* which quantifies the direction of the in-plane component of $\mathbf{m}$ with respect to $\hat{r}$. For $\psi = 0$ (resp. π) $\mathbf{m}$ points radially outward (resp. inward), and the skyrmion is called *Néel-type*; if $\psi = \pi/2$ (resp. $-\pi/2$) $\mathbf{m}$ winds around counterclockwise (resp. clockwise) as observed from above and the skyrmion is called *Bloch-type*. In general, a skyrmion may have any ψ , but the energetics of a particular material may dictate that one of these two extremes is

preferred.

The energy of such a configuration can be expressed as the sum of several distinct contributions:

- The demagnetization energy E_d quantifies the interaction of the magnetization $\mathbf{m}$ with the demagnetizing field $\mathbf{H}$ which it produces, which typically acts to reduce the overall magnetic moment of the film. In the absence of other interactions, it favors the in-plane magnetization of a ferromagnetic film.
- The anisotropy energy E_k quantifies the preference toward out-of-plane magnetization for 2D films, and competes directly with the demagnetization energy. Electrons at interfaces (and in particular, in quasi-2D materials) can experience strong spin-orbit interactions which tend to align their spins along the out-of-plane direction.
- The exchange energy E_A quantifies the tendency of magnetic moments to align with their neighbors. If two adjacent sites host electrons with parallel spins, their spatial wavefunction is necessarily antisymmetric which leads to a reduced Coulomb interaction energy. When this effect dominates the dipole interaction energy (which favors antiparallel spins on adjacent sites) the net effect is a preference for aligned spins.
- The DMI energy E_D quantifies the tendency of adjacent magnetic moments to be orthogonal, and therefore competes directly with the exchange energy (favoring (anti)parallel alignment) to produce a gradual tilting in the magnetization of the film. This effect arises due to the Dzyaloshinskii-Moriya interaction, which is a higher-order effect of spin-orbit coupling, and has both an interfacial component and a bulk component.

- The Zeeman energy E_z quantifies the interaction between the magnetic configuration and an externally applied magnetic field $\mathbf{H}_{ext} = H_z \hat{z}$.

The relevant variables which determine the magnitudes of these contributions are the saturation magnetization of the film M_s , the externally-applied field H_z , the thickness of the film d , the uniaxial magnetic anisotropy K_u , the exchange constant A , and the interfacial and bulk DMI coefficients D_i and D_b . Given a choice of $\theta(r)$ and ψ , each of these five contributions to the total energy of the configuration can be evaluated for a magnetization of the form 3.6.

3.2.3.2 Skyrmion Model

A relatively simple model of skyrmion structure called the 360° Domain Wall model was first given in [19] and experimentally validated in [20]. The input parameters to the model are the characteristic radius R_0 , the domain wall width Δ , the domain wall angle ψ , and the skyrmion number (or topological charge) $\Sigma = \pm 1$. By defining the dimensionless parameter $\rho_0 \equiv R_0/\Delta$ and radial coordinate $\varrho = r/\Delta$, it takes the form

$$\theta(\varrho) = 2\left(\tan^{-1}(e^{\varrho-\rho_0}) + \tan^{-1}(e^{\varrho+\rho_0})\right) - \frac{(\Sigma + 1)\pi}{2} \quad (3.7)$$

which can subsequently be plugged into Eq. 3.6.

Note that when $R_0 < 1$, the parameter R_0 is not a particularly intuitive measure of the skyrmion size as shown in 3.6. It is more informative to define the radius R and reduced radius $\rho \equiv R/\Delta$ of the skyrmion to be the radius of the $m_z = 0$ contour, which (for $\Sigma = +1$) are related to ρ_0 via

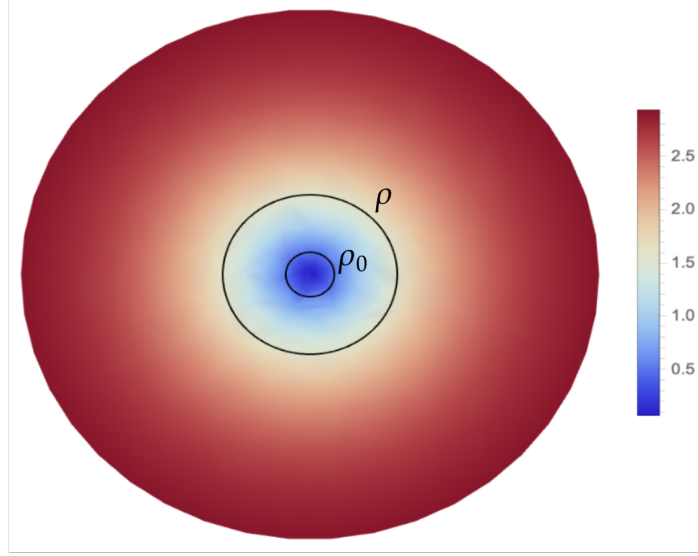


Figure 3.6: The profile of a magnetic skyrmion in the 360° domain wall model with $\rho_0 = 0.25$. The color scale denotes the polar angle θ , with the asymptotic magnetization pointing down. The inner circle is the parameter ρ_0 , while the outer circle is the more intuitive $\rho \equiv \sinh^{-1}(\cosh(\rho_0))$.

$$\rho = \sinh^{-1}(\cosh(\rho_0)) \quad (3.8)$$

Note that as $\rho_0 \rightarrow 0$, ρ approaches a minimum value $\rho_{min} = \sinh^{-1}(1) \approx 0.8813$.

3.2.3.3 Skyrmion Minimum Energy

An important observation can be made by computing the minimum exchange energy of a configuration with non-zero Skyrmion number. The exchange energy of an axisymmetric magnetic texture defined by a unit vector $\mathbf{n}$ is

$$E_A[\mathbf{n}] := A \int d^3r \sum (\nabla n^i(\mathbf{r}))^2 = Ad \int_{\mathbb{R}^2} dx dy \sum_i (\nabla n^i(x, y))^2 \quad (3.9)$$

Noting that

$$\left(\frac{\partial n^i}{\partial x} - \frac{1}{2} \left(\frac{\partial \mathbf{n}}{\partial y} \times \mathbf{n} \right)^i \right)^2 \geq 0 \implies \sum_i \left(\frac{\partial n^i}{\partial x} \right)^2 \geq \mathbf{n} \cdot \left(\frac{\partial \mathbf{n}}{\partial x} \times \frac{\partial \mathbf{n}}{\partial y} \right)$$

Similarly,

$$\left(\frac{\partial n^i}{\partial y} + \frac{1}{2} \left(\frac{\partial \mathbf{n}}{\partial x} \times \mathbf{n} \right)^i \right)^2 \geq 0 \implies \sum_i \left(\frac{\partial n^i}{\partial y} \right)^2 \geq \mathbf{n} \cdot \left(\frac{\partial \mathbf{n}}{\partial x} \times \frac{\partial \mathbf{n}}{\partial y} \right)$$

As a result, we find that⁵

$$E_A[\mathbf{n}] \geq 2Az \int dx dy \mathbf{n} \cdot \left(\frac{\partial \mathbf{n}}{\partial x} \times \frac{\partial \mathbf{n}}{\partial y} \right) = 8\pi Ad|\Sigma| \quad (3.10)$$

This is a direct link between the topology of a magnetic configuration (captured by the skyrmion number Σ) and its minimum energy. However, this gives us a sense of how much energy it costs to nucleate a magnetic skyrmion from the uniform ferromagnetic state, and is not directly related to the skyrmion's stability against collapse. For that, a more subtle study is needed.

3.2.3.4 Energy Analysis

Unsurprisingly, an analytic expression of the total energy of an arbitrary (or even axisymmetric) magnetic configuration is not available. However, a recent quasi-empirical model⁶ for the energy of an axisymmetric magnetic structure is both impressively accurate and computationally

⁵By flipping the minus signs in the preceding manipulations, one can obtain the analogous result for $\Sigma < 0$, hence the absolute value in Eq. 3.10

⁶ [21]

	Exact Expression	Approximate Expression
E_d	$-\frac{1}{2}\mu_0 \int d^3r (\mathbf{M} \cdot \mathbf{H}_s - M_s^2)$	$d \rightarrow 0 \implies K_u \mapsto K_u - \frac{1}{2}\mu_0 M_s^2$
E_k	$K_u \int d^3r \sin^2(\theta(r)) \equiv 2\pi K_u d \Delta^2 I_k$	$I_k \approx 2\rho - \frac{1}{3}e^{-\rho/\sqrt{3}}$
E_A	$A \int d^3r \nabla n ^2 = A \int d^3r (\partial_r \theta)^2 + \frac{1}{r^2} \sin^2(\theta))$ $\equiv 2\pi A d I_A$	$I_A \approx 2\rho + \frac{2}{\rho}$ $+1.93(\rho - 0.65)e^{-1.48(\rho - 0.65)}$
E_D	$-2\pi d \Delta (D_i \cos(\psi) + D_b \sin(\psi))$ $\times \int_0^\infty d\rho (\rho \partial_\rho \theta + \sin(\theta) \cos(\theta))$ $\equiv -2\pi d \Delta (D_i \cos(\psi) - D_b \sin(\psi)) I_D$	$I_D \approx \pi\rho + \frac{1}{2}e^{-\rho}$
E_z	$-\mu_0 \int d^3r \mathbf{M} \cdot \mathbf{H}_z$ $= -2\pi\mu_0 M_s H_z d \Delta^2 I_z$	$I_z \approx \rho^2 + \frac{\pi^2}{12} - 0.42 \exp(-\rho^2)$

Table 3.1: Exact and approximate expressions for the various contributions to the total energy of an isolated, axially-symmetric magnetic skyrmion in the 2D limit $d \rightarrow 0$. The I 's are numerical functions of the reduced skyrmion radius $\rho \equiv R/\Delta$, where R is the radius of the $m_z = 0$ contour and Δ is the domain wall thickness.

inexpensive, and uses the input configuration in Eq. 3.7 and some numerical approximations to otherwise complex integrals. The exact and numerically approximate expressions for the various contributions to the energy in the thin-film limit are given in Table 3.1.

The general route to collapse for a magnetic skyrmion is for its characteristic radius $R_0 \rightarrow 0$. This creates a singularity at the origin, and because topological invariants are *not* generally preserved in the presence of singularities, the configuration can revert to the uniform state. In the limit as $R_0 \rightarrow 0$, all of the contributions to the total energy go to zero (relative to the uniform state) with the exception of the exchange energy, which has a contribution from $\partial_r \theta$. Therefore, under the assumption of axisymmetry and the 360° DW model, we can numerically compute the universal skyrmion energy as $R \rightarrow 0$ to be

$$E_0 \approx 27.3Ad$$

which is quite close to our earlier estimate $E_0 \sim 8\pi Ad$. The energy barrier which stabilizes the skyrmion is therefore

$$\Delta E = E_0 - E_{tot}(R, \Delta)$$

The equilibrium values of R_0 and Δ are determined by the various contributions to the total energy. Without making the $d \rightarrow 0$ approximation to the stray field made in Table 3.1, the total energy can be written

$$\begin{aligned} \frac{E}{2\pi d} = & \Delta^2 (K_u I_k - \mu_0 M_s^2 I_s - \mu_0 H_z M_s I_z) \\ & - \Delta ([D_i \cos(\psi) - D_b \sin(\psi)] I_D - 2\mu_0 M_s^2 R d \cos^2(\psi) I_V) \\ & + A I_A \end{aligned} \quad (3.11)$$

where I_s and I_V correspond to the surface and volume stray fields. The approximate expressions for these integrals are numerically complex, but can be found in the supplementary material of [21], Section S2.7.

Observing that this expression is a quadratic function of Δ , it is clear that in order to have a stable equilibrium point, the first term must be positive, i.e.

$$K_u I_k - \mu_0 M_s^2 I_s - \mu_0 H_z M_s I_z > 0$$

The value of Δ at equilibrium is then given by

$$\begin{aligned} \Delta_{eq} = & \frac{[D_i \cos(\psi) - D_b \sin(\psi)] I_D - 2\mu_0 M_s^2 R d \cos^2(\psi) I_V}{2(K_u I_k - \mu_0 M_s^2 I_s - \mu_0 H_z M_s I_z)} > 0 \\ \implies & [D_i \cos(\psi) - D_b \sin(\psi)] I_D - 2\mu_0 M_s^2 R d \cos^2(\psi) I_V \end{aligned}$$

which must lie in the range $0 < \Delta < R/\sinh^{-1}(1)$ in order to have a physically meaningful solution. Plugging this in to our energy expression to eliminate Δ yields

$$\frac{E}{2\pi d} = AI_A - \frac{([D_i \cos(\psi) - D_b \sin(\psi)]I_D - 2\mu_0 M_s^2 R d \cos^2(\psi) I_V)^2}{4(K_u I_k - \mu_0 M_s^2 I_s - \mu_0 H_z M_s I_z)}$$

and an annihilation energy barrier

$$\frac{\Delta E}{2\pi d} = -(I_A - 4.34)A + \frac{([D_i \cos(\psi) - D_b \sin(\psi)]I_D - 2\mu_0 M_s^2 R d \cos^2(\psi) I_V)^2}{4(K_u I_k - \mu_0 M_s^2 I_s - \mu_0 H_z M_s I_z)} \quad (3.12)$$

We therefore arrive at the following qualitative conclusions based on this model:

- The perpendicular anisotropy term $K_u I_k$ must be strong enough to overcome the surface stray field and the externally applied H_z .
- A non-zero Dzyaloshinskii-Moriya interaction is essential - otherwise the energy could always be reduced by shrinking the domain wall thickness Δ .
- The exchange energy A serves to reduce the annihilation barrier for large ρ , and therefore favors compact skyrmions ($\rho \rightarrow \rho_{min}$) over bubble skyrmions ($\rho \gg 1$).

3.3 Abrikosov Vortices in Type-II Superconductors

3.3.1 Homotopic Preliminaries

Quantum theory is a rich source of topological phenomena, due in part to the ubiquity of the unitary groups $U(n)$, whose elements are the $n \times n$ matrices u which satisfy

$$u^\dagger u = uu^\dagger = \mathbf{1}_{n \times n} \quad (3.13)$$

where $u^\dagger$ is the conjugate transpose of u . The simplest example is the group $U(1)$, whose elements are simply the complex numbers with unit magnitude. This set traces out a unit circle in the complex plane, and so as one might expect we have that $U(1)$ and S^1 are homeomorphic. In particular, this implies that they are homotopically equivalent, which means that $\pi_d(U(1)) = \pi_d(S^1)$.

The homotopy groups of S^1 are given by

$$\pi_1(S^1) = \mathbb{Z} \quad \pi_d(S^1) = \{0\}, d \neq 1 \quad (3.14)$$

Therefore, if we are interested in a $U(1)$ -valued order parameter, we should be looking for winding numbers and families of inequivalent closed loops. If the domain R of our order parameter is $\mathbb{R}^3$ (or any other contractible space), then all configurations of the order parameter are homotopic to all others. If we allow the order parameter to be singular at isolated points $\mathbf{r}_i$, then $R = \mathbb{R}^3 - \cup_i \{\mathbf{r}_i\}$ which is no longer contractible. However, this space *is* simply-connected⁷, implying that the only map from $\pi_1(R) \rightarrow \pi_1(\mathcal{M})$ is the trivial map $0 \mapsto 0$ and again have no

⁷Recall that a space X is simply-connected if $\pi_1(X) = \{0\}$

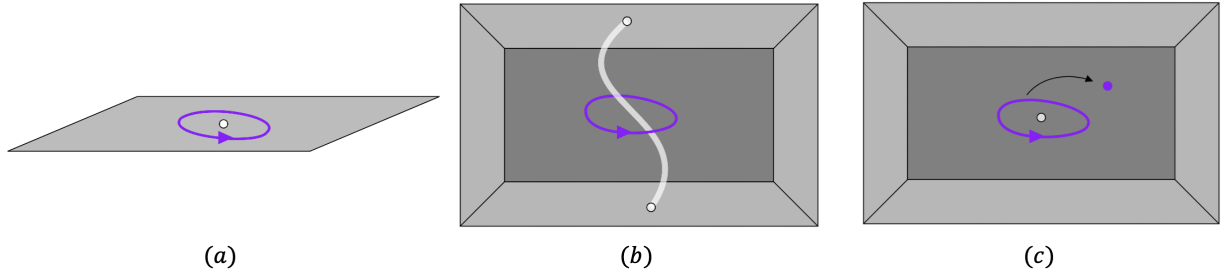


Figure 3.7: A $U(1)$ -valued order parameter exhibits non-trivial states only if $\pi_1(R) \neq \{0\}$. This applies to point defects in 2D (a) and line defects in 3D (b), but not point defects in 3D (c).

viable classification scheme.

As shown in Fig. 3.7, we are therefore led to the conclusion that if we are considering a $U(1)$ -valued order parameter on some region $R \subset \mathbb{R}^3$, we must have *line* defects - for example⁸, we might subtract the entire $\hat{z}$ -axis from $\mathbb{R}^3$ to let the domain of the order parameter be given by

$$R = \left\{ (a, b, c) \in \mathbb{R}^3 \mid a^2 + b^2 > 0 \right\}$$

Having done this, the possible configurations of the order parameter which are continuous on R decompose into families which are characterized by the winding numbers of the order parameter around the singular axis, which we will refer to as the *vorticity*; for a $U(1)$ -valued order parameter $\hat{\varphi}$ in cylindrical coordinates (r, θ, z) , the vorticity ν is given by

$$\nu[\varphi] = \frac{1}{2\pi i} \int d\theta \frac{d}{d\theta} [\log (\hat{\varphi}(1, \theta, 0))] \quad (3.15)$$

⁸An alternative and interesting possibility is that the line defects themselves form closed loops - see e.g. vortex rings.

3.3.2 A Note on Magnetic Free Energy

Before moving on to the Ginzburg-Landau theory, it is worthwhile to take a moment to discuss free energies in magnetic materials. In a material with polarization density $\mathbf{P}$ and magnetization density $\mathbf{M}$, the Maxwell-Heaviside⁹ equations are given by

$$\begin{aligned}\nabla \cdot \mathbf{D} &= \rho_f & \nabla \times \mathbf{E} &= -\frac{\partial}{\partial t} \mathbf{B} \\ \nabla \cdot \mathbf{B} &= 0 & \nabla \times \mathbf{H} &= \mathbf{J}_f + \frac{\partial}{\partial t} \mathbf{D}\end{aligned}$$

where $\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}$, $\mathbf{H} = \mathbf{B}/\mu_0 - \mathbf{M}$, and $\rho_f, \mathbf{J}_f$ are the free charge and current densities, respectively. The rate at which external power is delivered to a volume V by is given by¹⁰

$$\int_V d^3r \mathbf{E} \cdot \mathbf{J}_f = - \oint_{\partial V} d^2r \hat{n} \cdot (\mathbf{E} \times \mathbf{H}) + \int_V d^3r \left(\mathbf{D} \cdot \frac{\partial}{\partial t} \mathbf{E} + \mathbf{H} \cdot \frac{\partial}{\partial t} \mathbf{B} \right) \quad (3.16)$$

This invites us to interpret $\mathbf{D} \cdot d\mathbf{E} + \mathbf{H} \cdot d\mathbf{B}$ as the change in electromagnetic energy density within the volume. In particular, for a system with fixed volume immersed in a magnetic field, the first law of thermodynamics (for reversible, quasistatic processes) takes the form

$$dU = TdS - \delta w_{mag} = TdS + \int_V d^3r \mathbf{H} \cdot d\mathbf{B} \quad (3.17)$$

⁹Oliver Heaviside is often robbed of credit for a great many contributions to physics - in this case, inventing modern vector calculus notation and condensing Maxwell's twenty equations of electromagnetism down to four.

¹⁰This is Poynting's theorem.

3.3.2.1 The Helmholtz Free Energy

If S and $\mathbf{B}$ are held fixed by the experimenter, the allowed processes which can occur within the system are those for which no heat $\delta q = TdS$ is transferred to or from the system and for which the $\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M})$ is maintained at a constant value. Under these circumstances, the system will come to equilibrium in the state which minimizes U .

On the other hand, if the system is kept in thermal contact with a heat reservoir at temperature T , the entropy of the system is not held fixed because heat can be transferred back and forth between the system and reservoir. Therefore, the internal energy of the system is not minimized at equilibrium. However, because the system and reservoir are at the same temperature, we have that

$$dS + dS_R = \frac{\delta q}{T} - \frac{\delta q}{T} = 0 \implies d(S + S_R) = 0$$

Since the total entropy of system and reservoir is fixed, the total internal energy of the system and reservoir is minimized, and because $d(U + U_R) = dU - TdS = d(U - TS) = 0$ (where we have used that T is constant), we see that the so-called *Helmholtz free energy*

$$F = U - TS \tag{3.18}$$

is minimized at equilibrium. Observe the remarkable fact that because of the thermal equilibrium between the system and reservoir, we were able to reformulate an equilibrium condition for the combined "universe" (that $U + U_R$ is minimized) as an equilibrium condition about the system alone (that $F = U - TS$ is minimized), which holds as long as $dU_R = -TdS$.

3.3.2.2 The Gibbs Free Energy

Along precisely analogous lines, we often do not have direct experimental control of $\mathbf{B}$. If we have a sample within a solenoidal field coil, for example, what we really have control of is $\mathbf{H}$ - the thing which is related to the current in the coil. If there are processes or phase transitions which can change $\mathbf{M}$, holding $\mathbf{B}$ fixed implies dynamically modulating the current in the coil to compensate.

If instead we hold $\mathbf{H}$ fixed, then $\mathbf{B}$ is permitted to change. We have now introduced the possibility that the system may perform magnetic work $\delta w_{mag} = - \int_V d^3r \mathbf{H} \cdot d\mathbf{B}$ during a process. The change in energy of the environment becomes

$$dU_R = -T dS - \int_V d^3r \mathbf{H} \cdot d\mathbf{B}$$

where the latter term can be understood¹¹ by noting that a change in $\mathbf{B}$ induces a back-EMF in the solenoidal coil. As a result,

$$d(U + U_R) = dU - T dS - \int_V d^3r \mathbf{H} \cdot d\mathbf{B} = d \left(U - TS - \int_V d^3r \mathbf{H} \cdot \mathbf{B} \right) = 0$$

and so it is the *Gibbs free energy*

$$G = U - TS - \int_V d^3r \mathbf{H} \cdot \mathbf{B} \quad (3.19)$$

which is minimized at equilibrium under the constraint of fixed T and $\mathbf{H}$.

¹¹Working out the energetics of this is a straightforward but very illuminating exercise.

3.3.3 Ginzburg-Landau Theory of Superconductivity

In the Ginzburg-Landau theory of superconductivity, a superconductor is characterized by a $\mathbb{C}$ -valued order parameter ψ , whose squared magnitude $|\psi|^2$ gives the density of superconducting Cooper pairs. The Helmholtz free energy functional is taken by ansatz to be

$$F[\psi] - F_N := \int \left(\alpha |\psi|^2 + \frac{\beta}{2} |\psi|^4 + \frac{1}{2m^*} |(-i\hbar\nabla - e^*\mathbf{A})\psi|^2 + \frac{|\mathbf{B}|^2}{2\mu_0} \right) d^3r \quad (3.20)$$

where F_N is the free energy in the normal state with no electromagnetic fields, α and β are empirical parameters to be determined, e^* and m^* are the effective charge and mass of the Cooper pairs, $\mathbf{A}$ is the electromagnetic vector potential, and $\mathbf{B} = \nabla \times \mathbf{A}$ is the magnetic field. The supercurrent $\mathbf{J}$ is given by

$$\mathbf{J} \equiv \frac{e^*}{m^*} \text{Re} \{ \psi^* (-i\hbar\nabla - e^*\mathbf{A})\psi \} \quad (3.21)$$

It is extremely useful to non-dimensionalize this expression. In the absence of gradients or electromagnetic fields, and assuming that $\beta > 0$, the minimum of the free energy functional is found to be

$$|\psi|^2 = \begin{cases} 0 & \alpha \geq 0 \\ \frac{|\alpha|}{\beta} \equiv \psi_0^2 & \alpha < 0 \end{cases} \quad (3.22)$$

This suggests defining the normalized order parameter $\varphi \equiv \psi/\psi_0$. The coherence length - the scale which characterizes fluctuations in the order parameter - is found to be $\xi \equiv \sqrt{\hbar^2/4m|\alpha|}$ in the superconducting state¹². This scale can be absorbed into the gradient by defining normalized

¹²Observe the possibly unexpected factor 4 in the denominator, where we might expect a 2. To see this arise explicitly, set $\mathbf{A} = 0$, take ψ real, compute the Euler-Lagrange equations, and let $\psi = \psi_0 + \chi$, $\chi \ll \psi_0$ to obtain $\frac{\hbar^2}{4m|\alpha|} \nabla^2 \chi = \chi$

coordinates¹³ $\mathbf{r} \equiv \xi \mathbf{r}^*$. Defining the dimensionless vector potential $\mathbf{a} = \mathbf{A} \xi e^* / \hbar$, we finally obtain

$$\frac{F[\varphi] - F_N}{\beta \psi_0^2 \xi^3} = \int \left(-|\varphi|^2 + \frac{1}{2}|\varphi|^4 + 2|(\nabla - i\mathbf{a})\varphi|^2 + 4\left(\frac{\lambda}{\xi}\right)^2 (\nabla \times \mathbf{a})^2 \right) d^3 r^* \quad (3.23)$$

$$\mathbf{j} \equiv \frac{e^* \psi_0^2 \hbar}{m \xi} \mathbf{J} = \text{Im} \{ \varphi^* (\nabla - i\mathbf{a}) \varphi \} \quad (3.24)$$

The quantity $\lambda \equiv \sqrt{m^* \beta / \mu_0 e^{*2} |\alpha|}$ is the London penetration depth, which characterizes how far magnetic fields can penetrate into the superconductor before being screened out by the supercurrent which flows across the surface. The parameter $\kappa \equiv \lambda / \xi$ is called the Ginzburg-Landau parameter, which quantifies the energy cost of a non-zero magnetic field relative to the energy cost of density variations. To get a physical interpretation of κ , consider a magnetic field parallel to the surface of a superconductor. The magnitude of the order parameter rises to its bulk value over a characteristic distance ξ , while the magnetic field decays over a characteristic distance λ . In the $\kappa \ll 1$ limit, the magnetic field is screened out well before $|\varphi| = |\varphi_0|$, whereas in the $\kappa \gg 1$ limit the magnetic field persists for some distance into the bulk.

3.3.4 Abrikosov Vortices

Turning to topology, as long as the order parameter φ is non-vanishing, we may consider the alternative U(1)-valued order parameter $\hat{\varphi} \equiv \varphi / |\varphi|$. Referring to Sec. 3.3.1, we are led immediately to consider line defects characterized by non-zero vorticity. Such a state is called an *Abrikosov*¹⁴ *vortex*. To study them in greater depth, we adopt cylindrical coordinates (r, ϕ, z)

¹³We will suppress the asterisk on the normalized coordinates in what follows.

¹⁴Named for Alexei Abrikosov.

(appropriately normalized to the correlation length) and assume vanishing z -dependence. Writing the order parameter as $\varphi = \rho e^{i\vartheta}$ with $\rho \in \mathbb{R}$, we also assume that ρ depends on r alone. Finally, in order for φ to be continuous we must have that $\vartheta = m\phi$ for some $m \in \mathbb{Z}$ - it's easy to see that the vorticity is given by $\nu[\varphi] = m$. We also assume that $\mathbf{a} = a_\phi \hat{\phi}$, and that a_ϕ is a function of r alone. Under these conditions, the free energy functional takes the form

$$\frac{F[\rho] - F_N}{\beta\psi_0^2\xi^3} = \int \left(-\rho^2 + \frac{1}{2}\rho^4 + 2\dot{\rho}^2 + 2\left(a_\phi - \frac{m}{r}\right)^2 \rho^2 + 4\kappa^2 \left(\dot{a}_\phi + \frac{a_\phi}{r}\right)^2 \right) d^3r^* \quad (3.25)$$

The Euler-Lagrange equations for ρ and a_ϕ are

$$\begin{aligned} \ddot{\rho} + \frac{1}{r}\dot{\rho} &= -\frac{1}{2}(\rho - \rho^3) + \left(a_\phi - \frac{m}{r}\right)^2 \rho \\ \ddot{a}_\phi + \frac{1}{r}\dot{a}_\phi &= \frac{a_\phi}{r^2} + \frac{1}{2\kappa^2} \left(a_\phi - \frac{m}{r}\right) \rho^2 \end{aligned} \quad (3.26)$$

Far from the vortex core, we expect $a_\phi \rightarrow m/r$, yielding zero magnetic field and an asymptotic magnetic flux $\int r dr d\phi \mathbf{b} \cdot \hat{z} = 2\pi r a_\phi = 2\pi m$. For numerical reasons, it is useful to re-frame Eq. 3.26 in terms of a normalized flux variable $\Phi \equiv a_\phi r / m$, which yields

$$\begin{aligned} \ddot{\rho} + \frac{1}{r}\dot{\rho} &= -\frac{1}{2}(\rho - \rho^3) + \frac{m^2}{r^2}(1 - \Phi)^2 \rho \\ \ddot{\Phi} - \frac{1}{r}\dot{\Phi} &= -\frac{\rho^2}{2\kappa^2}(1 - \Phi) \end{aligned} \quad (3.27)$$

Solving these equations in the $r \ll 1$ limit yields the core solutions

$$\begin{aligned}\rho_{<}(r) &\sim r^m \\ \Phi_{<}(r) &\sim \frac{b_0 r^2}{2m}\end{aligned}\tag{3.28}$$

where b_0 is the value of the magnetic field at the core. In the opposite limit $r \gg 1$ we find approximate solutions

$$\begin{aligned}\rho_{>}(r) &\sim 1 - K_0(r) \\ \Phi_{>}(r) &\sim 1 - \frac{e^{-r/\sqrt{2k^2}}}{\sqrt{r/\sqrt{2k^2}}}\end{aligned}\tag{3.29}$$

where K_0 is the modified Bessel function. After some numerical experimentation, a fairly good approximation (shown in Fig. 3.8) for all r is provided by

$$\begin{aligned}\rho(r) &= \tanh(r)^m \\ \Phi(r) &= 1 - \left(\frac{\tanh(3r/4\sqrt{2k^2})}{3r/4\sqrt{2k^2}} \right)^3\end{aligned}\tag{3.30}$$

Recall that r is normalized to ξ , while $\kappa \equiv \lambda/\xi$. It makes good physical sense that the characteristic length scale over which the Cooper pair density ρ varies is given by the correlation length, while the flux and magnetic field vary over a length scale given by the London penetration depth.

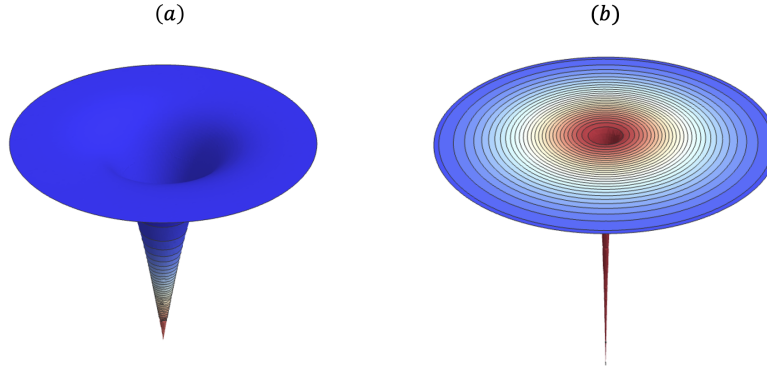


Figure 3.8: Abrikosov vortices and their associated magnetic fields for (a) $\kappa = 0.2$ and (b) $\kappa = 10$. The density ρ is reflected in the surface shape, while the color reflects the magnitude of the magnetic field. Note that in the $\kappa \ll 1$ limit, the flux is restricted to the very center of the core, whereas for $\kappa \gg 1$ the flux extends deep into the bulk.

3.3.4.1 The Energy of an Abrikosov Vortex

Plugging the Euler-Lagrange equation for ρ back into the free energy functional and integrating by parts yields

$$\frac{F[\rho, a_\phi] - F_N}{\beta\psi_0^2\xi^3} = 2\pi\ell \int_0^R \left(-\frac{1}{2}\rho^4 + \frac{4m^2\kappa^2}{r^2}\dot{\Phi}^2 \right) r dr$$

where R and ℓ are the extent of the sample in the $\hat{r}$ and $\hat{z}$ -directions¹⁵. Plugging in the approximations given in Eq. 3.30 yields the result

$$\frac{F[\rho, a_\phi] - F_N}{\beta\psi_0^2\xi^3} = -\frac{\pi R^2\ell}{2} + \epsilon(m, \kappa)\ell \quad \epsilon(m, \kappa) = \epsilon_\rho(m) + \epsilon_\Phi(m, \kappa) \quad (3.31)$$

¹⁵Normalized to ξ , of course.

where the respective energy terms take the approximate forms

$$\epsilon_\rho(m) = 0.537m^{0.359} + 0.178 \log(m) \quad \epsilon_\Phi(m, \kappa) = 1.448m^2\kappa^{0.568} |\log(\kappa)| \quad (3.32)$$

Once again, we find a direct link between the energy ΔF of the configuration and the vorticity m of the topological line defect. If $m = 0$, ρ is free to relax to the uniform value 1. However, if $m \neq 0$ then this is impossible - ρ *must* vanish at the singular axis.

Note that we assumed that the phase singularity arises because the order parameter vanishes at the singular axis, but this need not necessarily be so. We could consider $\varphi = e^{im\phi}$ with no density perturbation at all. However, in this case we find that $a_\phi = m/r \implies \mathbf{b} = 2\pi m\delta(r)\hat{z}$, which gives an infinite contribution $\propto \delta(0)$ to the magnetic energy density. More generally, a supercurrent $\mathbf{j} = (\frac{m}{r} - a_\phi)\rho^2\hat{\phi}$ is generated if $\rho \neq 1 \implies a_\phi \neq m/r$ which diverges (therefore requiring an infinite amount of energy) unless $\rho \rightarrow 0$ at the singular axis.

Chapter 4: Elementary Band Theory

4.1 Preliminary Remarks

We now turn our attention to the electronic states of a crystalline solid consisting of N_{el} electrons with mass m and charge $-e$ as well as N_{ion} ions with mass M and charges Ze . The corresponding $N_{el} + N_{ion}$ -particle Hamiltonian is [22], [23]

$$\begin{aligned} H = & \left(\sum_{i=1}^{N_{el}} -\frac{\hbar^2}{2m} \nabla_i^2 \right) + \left(\sum_{i=1}^{N_{ion}} -\frac{\hbar^2}{2M_i} \nabla_i^2 \right) \\ & - \sum_{i=1}^{N_{el}} \sum_{j=1}^{N_{ion}} \frac{Ze^2}{4\pi\epsilon_0 |\mathbf{x}_i - \mathbf{R}_j|} + \sum_{i \neq j} \frac{e^2}{4\pi\epsilon_0 |\mathbf{x}_i - \mathbf{x}_j|} + \sum_{i \neq j} \frac{Z^2 e^2}{4\pi\epsilon_0 |\mathbf{R}_i - \mathbf{R}_j|} \end{aligned} \quad (4.1)$$

whose respective terms correspond to the electron kinetic energy, ion kinetic energy, electron-ion interaction, electron-electron interaction, and ion-ion interaction.

At a glance the problem seems insurmountable, given some $N \sim 10^{24}$ electrons and an equal number of ions interacting among themselves and with each other. In order to make the problem even remotely tractable, some fairly drastic approximations need to be made.

- First, the ions are vastly heavier than the electrons, which means that their dynamics occur over much longer time scales. We can begin by treating them as stationary and then add their dynamics as an adiabatic perturbation. This is the so-called *Born-Oppenheimer approximation* and eliminates the second and fifth terms from 4.1.

- Next, we assume that the electron-electron interactions¹ can be approximated with a mean-field potential which is independent of the positions of the other electrons. This allows us to combine the third and fourth terms of 4.1 into a single term of the form $\sum_{i=1}^{N_{el}} V(\mathbf{x}_i)$, at which point the Hamiltonian factors into a direct sum of single-particle Hamiltonians of the form

$$H_0 = -\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{x}) \quad (4.2)$$

- Finally, we assume that we may treat V as a periodic function with the same periodicity as the crystal lattice. This ignores any disorder (crystalline defects, etc) and the fact that real crystals have edges - an issue we will return to at the end of the chapter.

These approximations can later be weakened or removed to varying degrees, but they provide us with a reasonable starting point. We will begin by trying to understand the states of a single-electron system subject to a periodic potential.

4.2 Bloch Theorem in the Continuum

4.2.1 Formulation of the Theorem

Consider the Hilbert space $L^2(\mathbb{R}^d)$, where $d \in \{1, 2, 3\}$, with a Schrödinger-type Hamiltonian $H = -\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{x})$. Furthermore, let there be d linearly-independent vectors $\{\mathbf{a}_1, \dots, \mathbf{a}_d\}$, called *primitive lattice vectors*, such that $V(\mathbf{x} + \mathbf{a}_i) = V(\mathbf{x})$. For any d -tuple of integers $(n_1, \dots, n_d)$, the linear combination $\sum_{i=1}^d n_i \mathbf{a}_i$ is called a *lattice vector* because it points from

¹Note that we are *not* assuming that the electrons are non-interacting, but rather that those interactions can be treated by a mean-field approximation. The potential which appears in the single-electron Hamiltonian therefore encodes this mean-field interaction as well as the interaction with the ionic lattice.

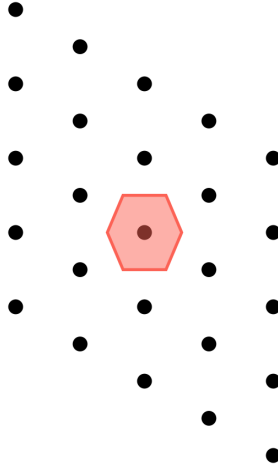


Figure 4.1: The Wigner-Seitz unit cell of a 2D triangular lattice, consisting of all of the points nearer to the origin than to any other lattice site.

one lattice site to another; let the space of all lattice vectors² $\text{span}_{\mathbb{Z}}(\mathbf{a}_1, \dots, \mathbf{a}_d)$ be denoted $\mathcal{L}$.

We can think of $\mathbb{R}^d$ as being tessellated by identical copies of a primitive unit cell $\mathcal{C}$ which have been translated to each lattice point. A good choice is the the *Wigner-Seitz* definition, wherein the unit cell consists of all of the points which are closer to the origin than to any other lattice point; an example is shown in Fig. 4.1 for a 2D triangular lattice.

We define the discrete translation operators T_i via $(T_i\psi)(\mathbf{x}) = \psi(\mathbf{x} - \mathbf{a}_i)$, and note that (i) each T_i is unitary, and (ii) the periodicity of the potential means that $T_i H T_i^\dagger = H \iff [H, T_i] = 0$. We therefore seek generalized eigenvectors of H which are simultaneous eigenvectors of each T_i . To do this, it is convenient to define the *primitive reciprocal lattice vectors* $\mathbf{b}_k$ such that $\mathbf{a}_j \cdot \mathbf{b}_k = 2\pi\delta_{jk}$, and the reciprocal lattice as $\mathcal{L}^{-1} = \text{span}_{\mathbb{Z}}(\{\mathbf{b}_1, \dots, \mathbf{b}_d\})$. Note that if ψ is an eigenvector of T_j with eigenvalue $e^{2\pi i q_j}$, $q_i \in [-\frac{1}{2}, \frac{1}{2}]$, then the function

² $\mathbb{Z}$ is the set of all integers, so $\text{span}_{\mathbb{Z}}(\mathbf{a}_1, \dots, \mathbf{a}_d)$ means the set of linear combinations of the $\mathbf{a}_i$'s with integer coefficients.

$u(\mathbf{x}) = e^{-iq_j \mathbf{b}_j \cdot \mathbf{x}} \psi(\mathbf{x})$ satisfies $u(\mathbf{x} + \mathbf{a}_i) = u(\mathbf{x})$. More generally, if we define $\mathbf{k} \equiv \sum_{i=1}^d q_i \mathbf{b}_i$, then $u_{\mathbf{k}}(\mathbf{x}) \equiv e^{-i\mathbf{k} \cdot \mathbf{x}} \psi_{\mathbf{k}}(\mathbf{x}) = u_{\mathbf{k}}(\mathbf{x} + \mathbf{a}_i)$ and $\psi(\mathbf{x}) = e^{i\mathbf{k} \cdot \mathbf{x}} u_{\mathbf{k}}(\mathbf{x})$ is an eigenvector of each T_j with respective eigenvalues $e^{2\pi i q_j}$. From this we conclude that if ψ is a simultaneous eigenvector of each T_j , then it can be expressed in the form

$$\psi(\mathbf{x}) = e^{i\mathbf{k} \cdot \mathbf{x}} u_{\mathbf{k}}(\mathbf{x}) \quad \mathbf{k} = \sum_i q_i \mathbf{b}_i$$

These ψ 's are the Bloch waves, and the cell-periodic functions $u_{\mathbf{k}}$ are called *Bloch functions*. The set of all $\mathbf{k}$'s is called the (*first*) *Brillouin zone*, which we will denote by **BZ**. Note that the Brillouin zone is the Wigner-Seitz cell of the reciprocal lattice $\mathcal{L}^{-1}$.

Having identified the form of simultaneous eigenvectors of the d translation operators, we turn to the Hamiltonian. Plugging Eq. 4.2.1 into the Schrödinger equation, we find

$$(H\psi)(\mathbf{x}) = e^{i\mathbf{k} \cdot \mathbf{x}} \left[\frac{(-i\hbar \nabla + \hbar \mathbf{k})^2}{2m} + V(\mathbf{x}) \right] u(\mathbf{x}) = E e^{i\mathbf{k} \cdot \mathbf{x}} u(\mathbf{x})$$

Observe that for each $\mathbf{k} \in \mathbf{BZ}$, the Bloch functions $u_{\mathbf{k}}$ can be understood as periodic eigenfunctions of the so-called *Bloch Hamiltonian*

$$H_{\mathbf{k}} = e^{-i\mathbf{k} \cdot \mathbf{x}} H e^{i\mathbf{k} \cdot \mathbf{x}} = \frac{(-i\hbar \nabla + \hbar \mathbf{k})^2}{2m} + V(\mathbf{x}) \quad (4.3)$$

with periodic boundary conditions. Each $H_{\mathbf{k}}$ is a self-adjoint operator with a discrete eigenvalue spectrum. We are therefore guaranteed that any square-integrable function f on $\mathcal{C}$ can be

expressed as a linear combination

$$f = \sum_n c_n u_{n\mathbf{k}} \quad (4.4)$$

where $u_{n\mathbf{k}}$ is the n^{th} eigenstate of $H_{\mathbf{k}}$.

Putting everything together, Bloch's theorem tells us that the eigenfunctions of a periodic Hamiltonian can be written in the form

$$\psi_{n\mathbf{k}}(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{x}} u_{n\mathbf{k}}(\mathbf{x}), \quad H\psi_{n\mathbf{k}} = E_{n\mathbf{k}}\psi_{n\mathbf{k}} \quad (4.5)$$

where $u_{n\mathbf{k}}$ is the periodic n^{th} eigenfunction of $H_{\mathbf{k}}$ with eigenvalue $E_{n\mathbf{k}}$. As an important note for later, there is some subtlety when we get to the edges of $\mathbf{BZ}$; naïvely taking the Brillouin zone to be the set $\mathbf{BZ} = \left\{ \sum_{i=1}^d q_i \mathbf{b}_i \right\}$ with $q_i \in [-1/2, 1/2]$ means that we will be counting all of the states on the edge of $\mathbf{BZ}$ twice, because for any reciprocal lattice vector $\mathbf{G} \in \mathcal{L}^{-1}$ and eigenvector $u_{\mathbf{k}}$ of $H_{\mathbf{k}}$, we have that $u_{\mathbf{k}}$ is also an eigenvector of $H_{\mathbf{k}+\mathbf{G}}$. At the moment, we may simply choose $q_i \in [-1/2, 1/2)$; in Chapter 5, we will take this a step further and identify opposite edges of the Brillouin zone with one another to give $\mathbf{BZ}$ the topology of a d -torus.

4.2.2 Example Models

4.2.2.1 The Free Electron Model

The simplest example is the 1D free electron model with $V(x) = 0$; this is trivially periodic³. Choosing the primitive lattice vector $\mathbf{a} = a\hat{x}$, we obtain $\mathbf{b} = \frac{2\pi}{a}\hat{x}$ and $\mathbf{BZ} = [-\frac{\pi}{a}, \frac{\pi}{a})$.

³Obviously if $V(x) = 0$, then $V(x+a) = V(x)$ for any a we choose. Picking a value for a out of thin air allows us to apply the machinery of Bloch's theorem, and it's a simple exercise to show that the set of all Bloch waves and their corresponding energies is independent of our choice.

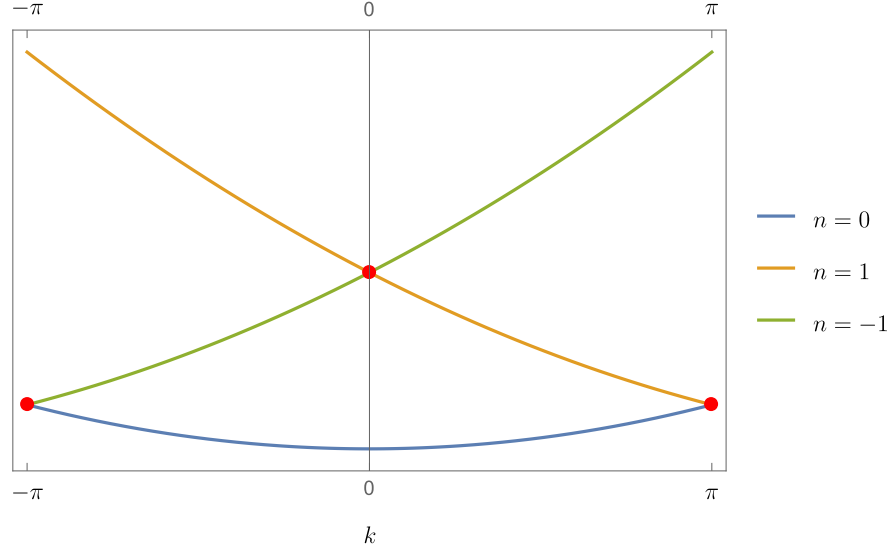


Figure 4.2: The free electron dispersion relation for the $n = 0$ and $n = \pm 1$ bands. The degenerate points where the bands intersect are denoted by red dots.

From there,

$$\frac{(-i\hbar\nabla + \hbar k)^2}{2m}u_k = E_k u_k \iff u_k \propto e^{iGx}, E_k = \frac{\hbar^2(G + k)^2}{2m}$$

Imposing periodicity yields $G = 2\pi n/a$ for some $n \in \mathbb{Z}$, and so we find that

$$u_{nk}(x) = \frac{1}{\sqrt{a}} e^{i\left(\frac{2\pi n}{a}\right)x} \quad E_{nk} = \frac{\hbar^2 \left(\frac{2\pi n}{a} + k\right)^2}{2m}$$

The free electron dispersion relation is shown in Fig. 4.2. Note that the bands intersect at $k = 0$ and $k = \pm\pi$; the corresponding two-fold degeneracies are generically lifted when interactions with a non-trivial periodic potential are taken into account, and the energy spectrum separates into bands.

4.2.2.2 The Kronig-Penney Model

One of the few non-trivial continuum systems with an exact analytic solution is the Kronig-Penney (or KP) model. The simplest 1D version is sometimes referred to as a *Dirac comb* because it features the Hamiltonian

$$H = -\frac{\hbar^2}{2m} \frac{d^2}{dx^2} + a\lambda \sum_{n \in \mathbb{Z}} \delta(x - na) \quad (4.6)$$

The δ -functions enforce a jump condition on the derivative of the Bloch waves. The result is the following eigenvalue equation for the periodic Bloch functions:

$$\left(\frac{d}{dx} + ik \right)^2 u_k = -q^2 u_k \quad (4.7a)$$

$$u_k(0) = u_k(a) \quad (4.7b)$$

$$u'_k(0) = u'_k(a) + \alpha u_k(0) \quad (4.7c)$$

where $q^2 \equiv \frac{2mE}{\hbar^2}$ and $\alpha \equiv \frac{2m\lambda a}{\hbar^2}$. The general solution to Eq. 4.7 takes the form

$$u_k(x) = e^{-ikx} (A \sin(qx) + B \cos(qx))$$

Enforcing the boundary conditions on u_k gives

$$u_k(x) = A e^{-ikx} (\sin(qx) + f(k, q) \cos(qx)), \quad f(k, q) = \frac{e^{-ika} \sin(qa)}{1 - e^{-ika} \cos(qa)} \quad (4.8)$$

while the jump condition yields

$$\cos(ka) = \cos(qa) + \alpha \frac{\sin(qa)}{qa} \quad (4.9)$$

Eq. 4.9 provides the Kronig-Penney dispersion relation (in the Dirac comb limit). Clearly this condition can be satisfied if and only if the right-hand side lies in the range $[-1, 1]$; for a generic nonzero value of α , the energy spectrum is partitioned up into allowed bands separated by forbidden gaps, as shown in Fig. 4.3.

4.3 Bloch Theorem on a Lattice

If the electrons in the crystal are tightly bound to their parent ions, then there is an alternative approach which is often computationally simpler than a continuum model. Let the full potential be decomposed as

$$V(\mathbf{x}) = V_0(\mathbf{x}) + \sum_{\mathbf{R} \neq \mathbf{0} \in \mathcal{L}} V_0(\mathbf{x} - \mathbf{R}) \equiv V_0(\mathbf{x}) + U(\mathbf{x}) \quad (4.10)$$

and imagine plucking a single unit cell out of the lattice. If we denote the (orthonormal) eigenfunctions of the single-cell Hamiltonian $H_{cell} = -\frac{\hbar^2}{2m}\nabla^2 + V_0(\mathbf{x})$ by ϕ_n , we loosely describe a system as tightly bound if, for all n and for all $\mathbf{R} \in \mathcal{L}$, the functions $\phi_n(\mathbf{x} - \mathbf{R})$ serve as a reasonable approximation to an orthonormal basis of the full Hamiltonian H . Operationally, this means that we can study the system by starting from the bound states of a single unit cell and then adding in the influence of the rest of the crystal as a perturbation.

In general, this is an approximation for three reasons:

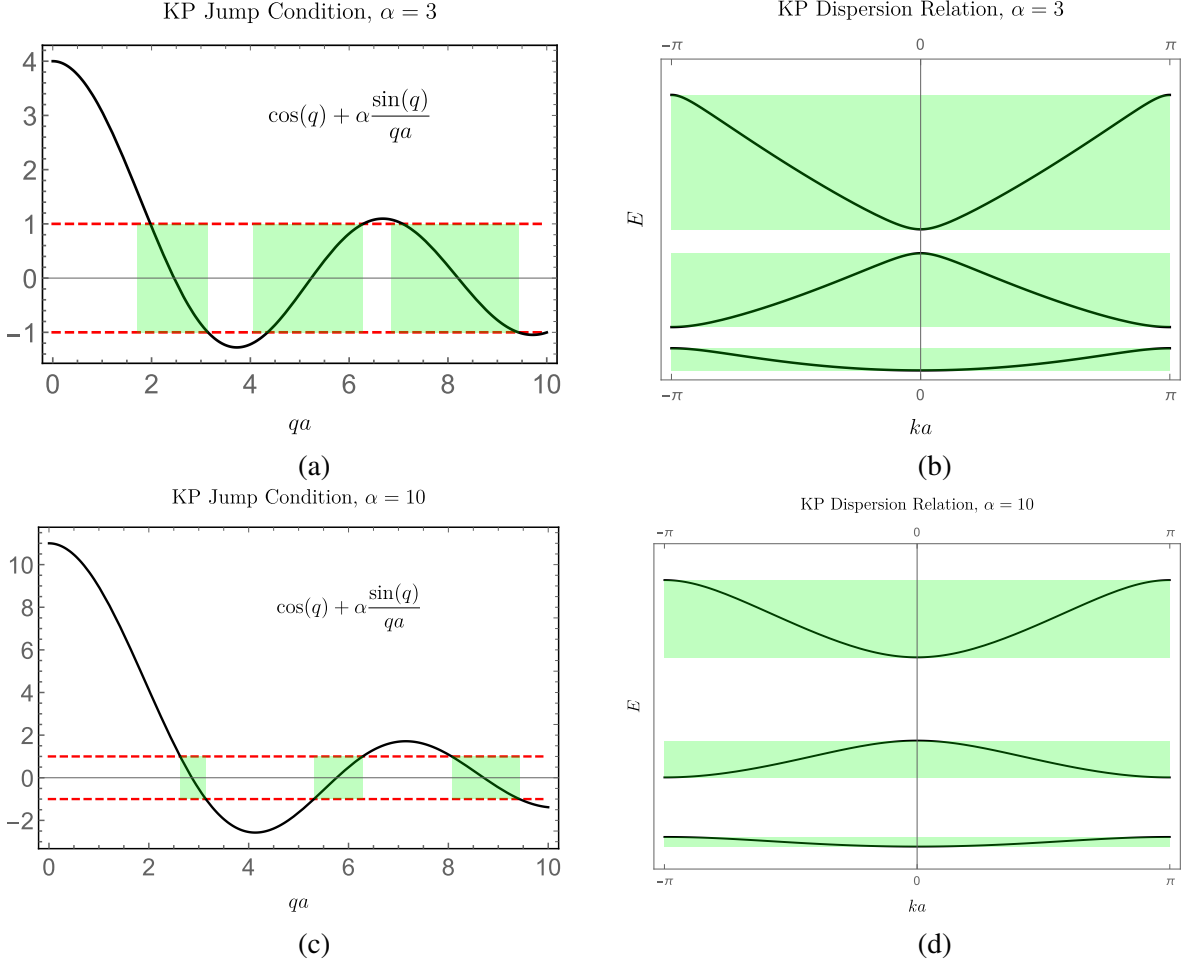


Figure 4.3: Illustrations of the allowed energies in the Kronig-Penney model for several values of the barrier strength. The allowed energies $E \equiv q^2 \hbar^2 / 2m$ are those for which $\cos(ka) = \cos(qa) + \alpha \sin(qa)/qa \in [-1, 1]$. When $\alpha = 3$ (Figs. *a* and *b*), the δ -barriers are fairly weak and the forbidden gaps are small. When $\alpha = 10$ (Figs. *c* and *d*), the δ -barriers are strong and the gaps are correspondingly large compared to the width of the bands.

1. $\phi_n(\mathbf{x})$ and $\phi_m(\mathbf{x}-\mathbf{R})$ are not orthogonal, so the overlap integrals⁴ $\alpha_{nm}(\mathbf{R}) \equiv \langle \phi_n, T_{\mathbf{R}}\phi_m \rangle \neq 0$ for $\mathbf{R} \neq \mathbf{0}$.
2. At a given lattice site, the presence of the off-site potential U generally couples eigenfunctions of H_{cell} to those with different energies so $\beta_{nm} \equiv \langle \phi_n, U\phi_m \rangle \neq 0$.
3. The presence U also couples eigenfunctions of H_{cell} to those localized to adjacent cells, so $\gamma_{nm}(\mathbf{R}) \equiv \langle \phi_n, UT_{\mathbf{R}}\phi_m \rangle \neq 0$.

If the single-cell potential V_0 is strong and well-localized, then the overlap integrals $\alpha_{nm}(\mathbf{R}) \approx 0$ for $\mathbf{R} \neq \mathbf{0}$. If U is sufficiently weak, then transitions between single-cell states with different energies are strongly suppressed and $\beta_{nm}(\mathbf{R}) \approx \beta_n \delta_{nm} \delta_{\mathbf{R},0}$ can be absorbed into the energies E_n . If both conditions are met, then the full Hamiltonian is well-approximated by the tight-binding Hamiltonian

$$H_{TB} = \sum_{n\mathbf{R}} (E_n + \beta_n) |n, \mathbf{R}\rangle \langle n, \mathbf{R}| + \frac{1}{2} \sum_{nm\mathbf{R}\mathbf{R}'} \gamma_{nm}(\mathbf{R} - \mathbf{R}') |n, \mathbf{R}\rangle \langle m, \mathbf{R}'| \quad (4.11)$$

where $|n, \mathbf{R}\rangle \equiv T_{\mathbf{R}}|\phi_n\rangle$.

4.3.1 Formulation of the Theorem

Consider now the Hilbert space $\ell^2(\mathbb{Z}^d)$ whose elements are the square-summable sequences indexed by d -tuples of integers. Each $\mathbf{n} \equiv (n_1, \dots, n_d)$ defines a particular lattice site $\mathbf{R}_{\mathbf{n}} = \sum_{j=1}^d n_j \mathbf{a}_j$, where $\{\mathbf{a}_1, \dots, \mathbf{a}_d\}$ are once again the primitive lattice vectors. A useful basis for this space is given by $|\mathbf{n}\rangle = \{\delta_{j\mathbf{n}}\}_{j \in \mathbb{Z}^d}$ i.e. the sequence whose $\mathbf{n}^{th}$ term is 1 and whose remaining

⁴Recall that $T_{\mathbf{R}}$ is the discrete translation operator, so $(T_{\mathbf{R}}\phi)(\mathbf{x}) = \phi(\mathbf{x} - \mathbf{R})$

terms are 0; note that $\mathbf{n} \equiv (n_1, \dots, n_d)$ and $\mathbf{j} \equiv (j_1, \dots, j_d)$. We interpret $|\mathbf{n}\rangle$ as a vector describing a state which is localized to the $\mathbf{n}^{th}$ lattice site. We'd also like to add a space $\mathcal{H}_{int}$ of internal excitations - these could be spin states, atomic energy levels, etc. Therefore, our full Hilbert space should be $\ell^2(\mathbb{Z}^d) \otimes \mathcal{H}_{int}$ which is spanned by the basis $|\mathbf{n}, \mu\rangle \equiv |\mathbf{n}\rangle \otimes |\mu\rangle$ where $\{|\mu\rangle\}$ is a basis for $\mathcal{H}_{int}$.

In this basis, the most generic Hamiltonian is given by Eq. 4.11. Translational invariance of the Hamiltonian implies that $h_{\mathbf{n}\mathbf{n}'+\mu\mu'}$ depends only on $\mathbf{n} - \mathbf{n}'$; as a result, it is convenient to rewrite this as

$$H = \sum_{\mathbf{n}\mathbf{j}\mu\mu'} h_{\mathbf{j}\mu\mu'} |\mathbf{n}, \mu\rangle \langle (\mathbf{n} + \mathbf{j})\mu'| = \sum_{\mathbf{n}\mathbf{j}\mu\mu'} h_{\mathbf{j}\mu\mu'} |\mathbf{n}\rangle \langle \mathbf{n} + \mathbf{j}| \otimes |\mu\rangle \langle \mu'| \quad (4.12)$$

We now perform a semi-discrete Fourier transform by defining

$$|\mathbf{k}\rangle = \frac{1}{(2\pi)^{d/2}} \sum_{\mathbf{n} \in \mathbb{Z}^d} e^{-i\mathbf{k} \cdot \mathbf{R}_\mathbf{n}} |\mathbf{n}\rangle \quad |\mathbf{n}\rangle = \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{BZ}} d^d k e^{i\mathbf{k} \cdot \mathbf{R}_\mathbf{n}} |\mathbf{k}\rangle \quad (4.13)$$

where, as before, $\mathbf{k}$ is an element of the Brillouin zone $\mathbf{BZ}$. Plugging this in to the previous expression for H yields

$$H = \int_{\mathbf{BZ}} d^d k |\mathbf{k}\rangle \langle \mathbf{k}| \otimes \left[\underbrace{\tilde{h}_{\mathbf{k}\mu\mu'} |\mu\rangle \langle \mu'|}_{\equiv H_{\mathbf{k}}} \right] \quad \tilde{h}_{\mathbf{k}\mu\mu'} \equiv \sum_{\mathbf{j} \in \mathbb{Z}^d} h_{\mathbf{j}\mu\mu'} e^{i\mathbf{k} \cdot \mathbf{R}_\mathbf{j}} \quad (4.14)$$

The eigenvectors for this Hamiltonian can be expressed as $|\mathbf{k}\rangle \otimes |u_{n\mathbf{k}}\rangle \equiv |\mathbf{k}, u_{n\mathbf{k}}\rangle$, where $|u_{n\mathbf{k}}\rangle$ is the n^{th} eigenvector of $H_{\mathbf{k}}$ - in precise analogy to the result in the continuum case.

4.3.2 Example Models

4.3.2.1 Free Nearest-Neighbor Hopping on a Lattice

Consider a 1D lattice with lattice spacing a and Hamiltonian

$$H = -t \sum_{n \in \mathbb{Z}} (|n\rangle\langle n+1| + h.c.) = \sum_{n, j \in \mathbb{Z}} \underbrace{(-t\delta_{j,1} + -t\delta_{j,-1})}_{\equiv h_j} |n\rangle\langle n+j| \quad (4.15)$$

The Brillouin zone is the circle $[-\pi/a, \pi/a]$ where we identify the endpoints as equivalent. The Bloch Hamiltonian becomes

$$H_k = \sum_{j \in \mathbb{Z}} h_j e^{ijka} = -(e^{ika} + e^{-ika}) = -2t \cos(ka) \quad (4.16)$$

and the energy eigenstates are simply the plane waves $|k\rangle \equiv \frac{1}{\sqrt{2\pi}} \sum_n e^{-ikna} |n\rangle$. Note that $|\pi/a\rangle$ and $|- \pi/a\rangle$ are equal to each other, providing some additional motivation for our impending identification of **BZ** with a circle rather than a half-open interval.

4.3.2.2 Hopping on a Bipartite Chain

A less trivial lattice example is obtained by allowing each lattice site to have two accessible states A and B with energies $-\epsilon$ and $+\epsilon$ respectively. On-site hopping between A and B occurs with amplitude v , and intercell hopping between $|n, B\rangle$ and $|n+1, A\rangle$ occurs with amplitude w . The Hamiltonian for this system is

$$\begin{aligned}
H = & \sum_{n \in \mathbb{Z}} \epsilon \left(|n, B\rangle \langle n, B| - |n, A\rangle \langle n, A| \right) \\
& + v \left(|n, A\rangle \langle n, B| + |n, B\rangle \langle n, A| \right) \\
& + w \left(|n, B\rangle \langle n+1, A| + |n+1, A\rangle \langle n, B| \right)
\end{aligned} \tag{4.17}$$

where we take $\epsilon, v, w \in \mathbb{R}$. In terms of Eq. 4.12, we have that

$$\begin{aligned}
h_{0AA} &= -h_{0BB} = -\epsilon \\
h_{0AB} &= h_{0BA} = v \\
h_{1BA} &= h_{-1AB} = w
\end{aligned} \tag{4.18}$$

and

$$\begin{aligned}
\tilde{h}_{kAA} &= -\tilde{h}_{kBB} = -\epsilon \\
\tilde{h}_{kAB} &= v + we^{ika} = \overline{\tilde{h}_{kBA}}
\end{aligned} \tag{4.19}$$

and so the Bloch Hamiltonian is given by

$$H_k = \begin{pmatrix} -\epsilon & v + we^{ika} \\ v + we^{-ika} & \epsilon \end{pmatrix} \tag{4.20}$$

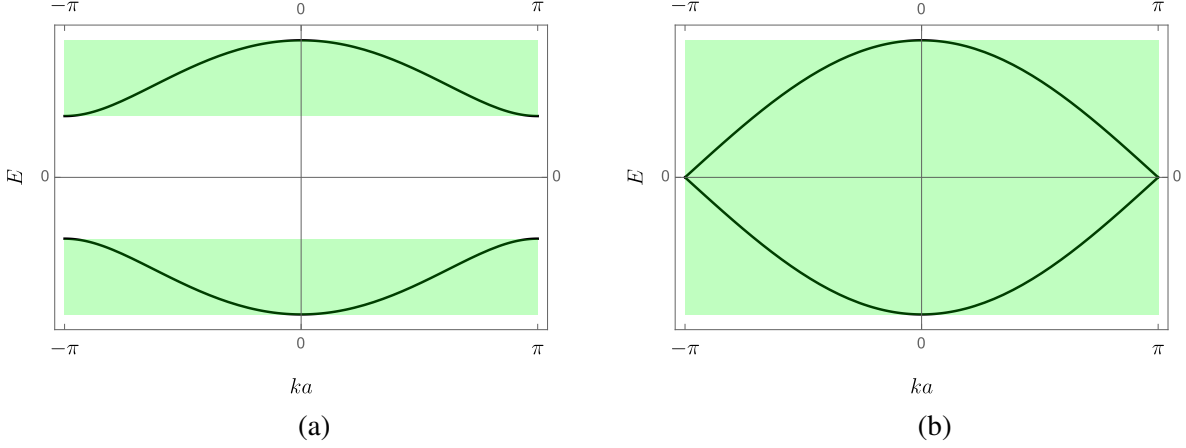


Figure 4.4: The dispersion relation for the bipartite chain with on-site energy $\pm\epsilon$ and intracell and intercell hopping amplitudes v and w . Allowed energy ranges are highlighted in green. In (a), $(\epsilon, v, w) = (0.5, 1, 1)$ and the system is gapped. In (b), $(\epsilon, v, w) = (0, 1, 1)$ and the gap closes, making the system a semimetal.

The eigenvalues of this Hamiltonian are given by $E_k = \pm\sqrt{\epsilon^2 + v^2 + w^2 + 2vw \cos(ka)}$, and so the two energy intervals are given by

$$\begin{aligned} I_- &= \left[-\sqrt{\epsilon^2 + (v+w)^2}, -\sqrt{\epsilon^2 + (v-w)^2} \right] \\ I_+ &= \left[\sqrt{\epsilon^2 + (v-w)^2}, \sqrt{\epsilon^2 + (v+w)^2} \right] \end{aligned} \quad (4.21)$$

with band gap $2\sqrt{\epsilon^2 + (v-w)^2}$. The band structure of this system is plotted for in Fig. 4.4. Note that if $\epsilon = 0$ and $v = w$, the bands will touch⁵ at the edges of the Brillouin zone and the band gap closes; we will return to this possibility in Chapter 5 when we meet the Su-Schrieffer-Heeger model of polyacetylene.

⁵When the gap closes at discrete points, the system is classified as a *semimetal*.

4.4 Finite Systems and the Thermodynamic Limit

Up to this point, we have formulated Bloch's theorem for infinite systems - that is, on the unbounded Hilbert spaces $L^2(\mathbb{R}^d)$ and $\ell^2(\mathbb{Z}^d)$, respectively. This is ostensibly necessary because it is obviously impossible for a system restricted to a compact region of $\mathbb{R}^3$ (or $\mathbb{Z}^3$) to exhibit translation invariance. However, it introduces a number of inconvenient technical issues, such as the non-normalizability of the Bloch waves and the formally infinite number of electrons needed to fill a given band.

An alternative approach is to consider a crystalline system on a *torus* - which we may understand as a d -dimensional box with side length L whose opposite faces are identified with one another. This is ultimately unphysical - real space is not (apparently) periodic after all - but in the limit as $L \rightarrow \infty$ then we can postulate on physical grounds that it will be a good physical model for the bulk of a crystal, far away from any edge effects.

The formulation of Bloch's theorem on a periodic domain mirrors the approaches taken above. The primary difference is that before, we defined the Brillouin zone to be the set of points

$$\mathbf{BZ} = \left\{ \sum_{i=1}^d q_i \mathbf{b}_i \mid q_i \in [-1/2, 1/2) \right\} \quad (4.22)$$

where $\mathbf{b}_i$ are the reciprocal lattice vectors. When the system is toroidal with side lengths $L_i \equiv N_i |\mathbf{a}_i|$, then q_i takes on discrete values $(-\frac{1}{2} + \frac{n}{N_i})$ for $n = 1, \dots, N_i$. As a result, the Brillouin zone becomes discrete (regardless of whether the real space is treated as a continuum or a discrete lattice) and the integrals over $\mathbf{BZ}$ must be replaced by discrete sums via

$$\int_{\Gamma} d^d k \longleftrightarrow \frac{(2\pi)^d}{\mathcal{V}} \sum_{\mathbf{k}}, \quad \mathcal{V} \equiv \prod_{i=1}^d L_i \quad (4.23)$$

The Bloch waves take the same form as before, as do the equations satisfied by the Bloch functions $u_{n\mathbf{k}}$ (or $|u_{n\mathbf{k}}\rangle$ in the discrete formulation). In particular, the energy eigenvalues can be obtained by plugging the discrete quasimomenta $\mathbf{k}_i$ into the continuous-band dispersion relation $E_{n\mathbf{k}}$ derived in the unbounded space formulation. In this sense, taking the thermodynamic limit $L_i \rightarrow \infty$ simply means adding points to the discrete Brillouin zone until it becomes a continuum. This also implies that the discrete density of states - which formally consists of a sum of delta functions - converges (in some sense) to the continuous density of states we obtain in the unbounded formulation, and so the use of the latter introduces an error which vanishes in the thermodynamic limit.

4.4.0.1 Hopping on a Circle

To see a concrete example, consider hopping on the finite chain with N sites, where we enforce periodic⁶ boundary conditions. The Hamiltonian becomes

$$H = -t \sum_{n=1}^{N-1} (|n\rangle\langle n+1| + h.c.) - t(|1\rangle\langle N| + |N\rangle\langle 1|) \quad (4.24)$$

We now perform a discrete Fourier transform

$$|k\rangle = \frac{1}{\sqrt{N}} \sum_{n=1}^N e^{-ikna} |n\rangle \longleftrightarrow |n\rangle = \frac{1}{\sqrt{N}} \sum_{k \in \mathbf{BZ}} e^{ikna} |k\rangle \quad (4.25)$$

⁶These boundary conditions are usually called *Born-von Karman* boundary conditions

where now

$$\mathbf{BZ} = \left\{ \frac{\pi}{a} \left(-1 + \frac{2j}{N} \right) \mid j = 0, \dots, N-1 \right\} \quad (4.26)$$

Substituting this transformation into the Hamiltonian yields the same result as the free hopping example from the previous section - namely that $H_k = -2t \cos(ka)$ - but now the k 's are restricted to a discrete set.

4.4.0.2 The Density of States

Now that everything is finite, we may consider the number of states in a given energy interval. Because the number of points in the Brillouin zone is the same as the number of lattice sites, we may expect that the number of states in a given interval should scale with the system size $L \equiv Na$. Therefore, it is reasonable to define the density of states $g(E)$ as the number of states *per unit volume* with energy between E and $E + dE$, with the expectation that this quantity should be finite when we take the limit $L \rightarrow \infty$.

The discrete density of states is formally given by

$$g(E) = \frac{1}{L} \sum_{k \in \mathbf{BZ}} \delta(E - 2t \cos(ka)) \quad (4.27)$$

Because the density of states is particularly useful when computing weighted averages of physical quantities, we might consider what happens when we integrate g against some well-behaved function f . The result is

$$\int f(E) g(E) dE = \frac{1}{L} \sum_{k \in \mathbf{BZ}} f(2t \cos(ka))$$

Noting that the spacing between k 's is $\Delta k \equiv 2\pi/L$, we may approximate the sum as an integral:

$$\frac{1}{L} \sum_{k \in \mathbf{BZ}} f(-2t \cos(ka)) = \frac{1}{2\pi} \sum_{k \in \mathbf{BZ}} f(-2t \cos(ka)) \Delta k \xrightarrow{N \rightarrow \infty} \frac{1}{2\pi} \int_{k \in \mathbf{BZ}} f(-2t \cos(ka)) dk$$

Finally, letting $dk = \left| \frac{dk}{dE} \right| dE = dE/[2ta\sqrt{1 - E^2/4t^2}]$ this becomes

$$\int f(E)g(E)dE \xrightarrow{L \rightarrow \infty} \int \frac{f(E)}{2t\sqrt{1 - \left(\frac{E}{2t}\right)^2}} dE \quad (4.28)$$

$$\implies g(E) \xrightarrow{L \rightarrow \infty} \frac{1}{2ta\sqrt{1 - \left(\frac{E}{2t}\right)^2}} = \left| \frac{dk}{dE} \right| \quad (4.29)$$

As anticipated, the density of states remains finite in the $L \rightarrow \infty$ limit, and is found by differentiating and inverting the dispersion relation. Though the DOS formally depends on L , this dependence vanishes as L becomes macroscopically large.

4.5 The Emergence of Forbidden Bands

Consider again Eq. 4.3. The $H_{\mathbf{k}}$'s constitute a *continuous family* of self-adjoint operators labeled by the pseudomomentum $\mathbf{k} \in \mathbf{BZ}$. The compactness of $\mathcal{C}$ and the periodic boundary conditions ensure that for each $\mathbf{k}$, the eigenvalue spectrum of $H_{\mathbf{k}}$ is discrete; at the same time, the fact that $H_{\mathbf{k}}$'s constitute a continuous family ensures that the eigenvalues $E_{\mathbf{k}}$ are continuous and periodic functions of $\mathbf{k}$. Therefore, as we sweep $\mathbf{k}$ across the Brillouin zone, the n^{th} eigenvalue $E_{n\mathbf{k}}$ traces out a compact (i.e. closed and bounded) interval $\mathcal{I}_n \subset \mathbb{R}$ called an (*allowed*) *energy*

band.

There is no *a priori* constraint on how these bands align with one another. Distinct bands may overlap, meet at discrete points, or be entirely disjoint. However, the emergence of forbidden *band gaps* at the boundaries of the Brillouin zone is a generic feature of a system with a periodic potential, as seen in the previous examples - a fact which can be explained in a number of different ways.

4.5.0.1 Nearly-free Electrons

Consider the case of electrons whose interaction with the lattice is sufficiently weak that the potential $V_c(\mathbf{x})$ can be thought of as a small perturbation to the free electron Hamiltonian $H = -\frac{\hbar^2}{2m}\nabla^2$. If $\mathbf{k}$ lies on the edge of the Brillouin zone, then there exists some reciprocal lattice vector $\mathbf{G}$ such that $|\mathbf{k}| = |\mathbf{k} + \mathbf{G}|$, and therefore $E_{\mathbf{k}}^{(0)} = E_{\mathbf{k}+\mathbf{G}}^{(0)}$ where $E^{(0)}$ is the unperturbed energy. As a result, we can express the eigenfunctions at this energy as standing waves of the form

$$\psi = e^{i\mathbf{k}_0 \cdot \mathbf{x}} \left[A \cos\left(\frac{\mathbf{G} \cdot \mathbf{x}}{2}\right) + B \sin\left(\frac{\mathbf{G} \cdot \mathbf{x}}{2}\right) \right], \quad \mathbf{k}_0 \equiv \mathbf{k} + \frac{1}{2}\mathbf{G} \quad (4.30)$$

The periodicity of the potential means that we can write it as

$$V_c(\mathbf{x}) = \sum_{\mathbf{G}'} \alpha_{\mathbf{G}'} \cos(\mathbf{G}' \cdot \mathbf{x}) + \beta_{\mathbf{G}'} \sin(\mathbf{G}' \cdot \mathbf{x}) \quad (4.31)$$

where the sum is over the reciprocal lattice vectors. Computing the expected value of the potential in our states yields

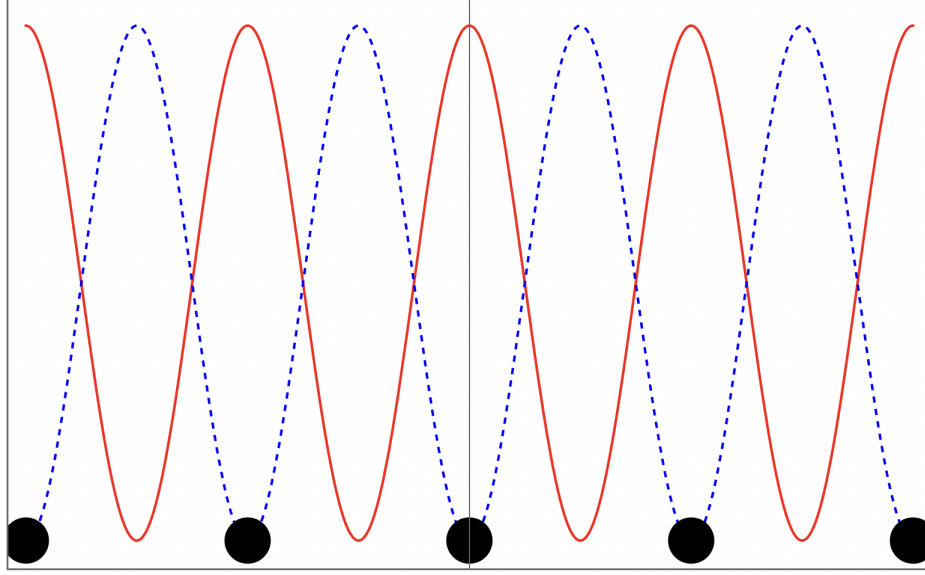


Figure 4.5: The origin of the band gap in nearly free electron models at the edge of the Brillouin zone. States at the edge of **BZ** can be written as (co)sinusoidal waves whose corresponding probability densities are maximized (red) or minimized (blue) on the lattice sites.

$$\langle V_c \rangle_\psi = \left(\frac{A^2 - B^2}{2} \right) \alpha_{\mathbf{G}} + AB\beta_{\mathbf{G}} \quad (4.32)$$

where $\mathcal{V}$ is the volume of the unit cell. If we assume that the potential is centrosymmetric for simplicity, then $\beta_{\mathbf{G}} = 0$ and we find that the (co)sinusoidal states are respectively pushed up and down in energy by an amount $|\alpha_{\mathbf{G}}|$. In essence, the (co)sinusoidal states achieve their maximum and minimum probability densities at the center of the ionic potential, leading to a splitting of the degeneracy and the opening of a forbidden energy gap as shown in 4.5

A closely related picture is to note that at the edges of the Brillouin zone, scattering off of the lattice plane defined $\mathbf{G} \cdot \mathbf{x} = 0$ for reciprocal lattice vector $\mathbf{G}$ satisfies the von Laue condition $2\mathbf{k} \cdot \mathbf{G} = |\mathbf{G}|^2$ for destructive interference; as a result, such waves cannot propagate through the crystal. This corresponds to the opening of a gap (though as we shall see in the next section, such states *can* be exponentially localized to the surfaces of a finite crystal).

4.5.0.2 Tightly Bound Electrons

Alternatively, we might consider the case of electrons which are so tightly bound to their parent ions that they may be considered nearly isolated. In this case, the Hamiltonian would be well-approximated by a tight-binding Hamiltonian of the form $H = \sum_{n \in \mathbb{Z}} |n\rangle\langle n| \otimes H_{atom}$, where n labels the lattice site and H_{atom} is the atomic Hamiltonian with eigenstates $|\phi_j\rangle$ and eigenvalues ϵ_j . The eigenvalues of this Hamiltonian are simply the discrete atomic energies ϵ_j , and the eigenvectors take the form

$$\psi = \sum_n c_n |n, \phi_j\rangle, \quad H\psi = \epsilon_j \psi \quad (4.33)$$

with the c_n 's arbitrary. As a result, the energy levels of the isolated Hamiltonian are massively degenerate. If we turn on interactions with neighboring ions via a hopping term

$$V = -t \sum_{n \in \mathbb{Z}} (|n\rangle\langle n+1| \otimes \mathbf{1} + |n+1\rangle\langle n| \otimes \mathbf{1}) \quad (4.34)$$

then this degeneracy is broken. Computing the expected value of the potential in the state given by Eq. 4.33 yields

$$\langle V \rangle_\psi = -t \sum_{n \in \mathbb{Z}} (\bar{c}_n c_{n+1} + \overline{c_{n+1}} c_n) = -2t \sum_{n \in \mathbb{Z}} \text{Re}(\bar{c}_n c_{n+1}) \quad (4.35)$$

Writing $c_n = r_n e^{i\theta_n}$, the expected value becomes

$$\langle V \rangle_\psi = -2t \sum_{n \in \mathbb{Z}} r_n r_{n+1} \cos(\Delta_n), \quad \Delta_n \equiv \theta_{n+1} - \theta_n \quad (4.36)$$

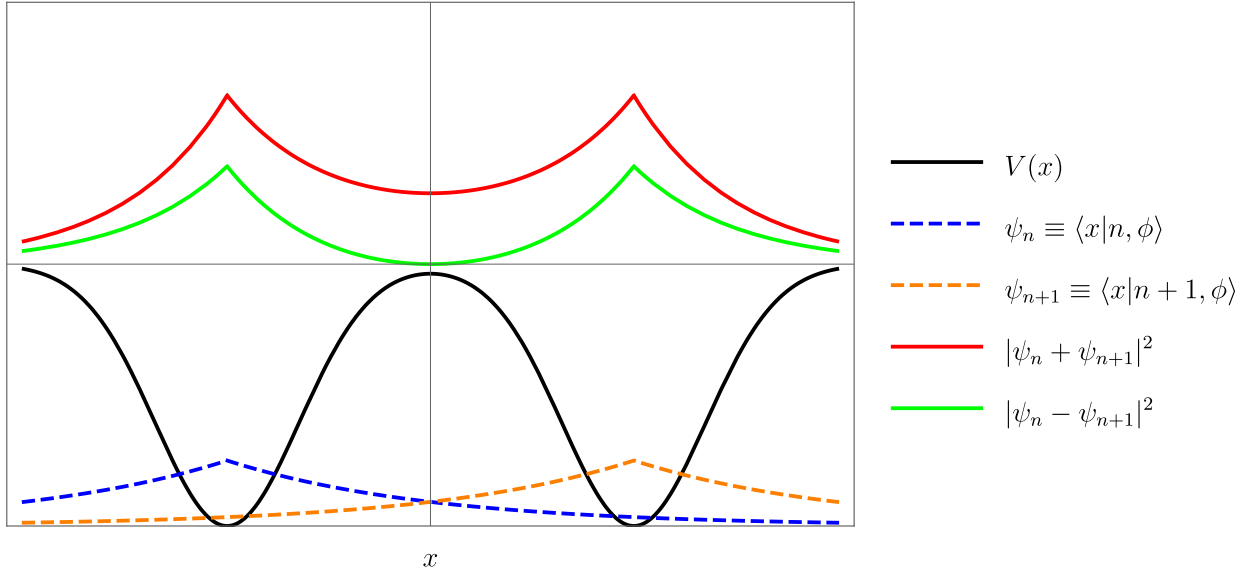


Figure 4.6: A cartoon illustrating the breaking of the isolated atom energy degeneracy by inclusion of nearest-neighbor interactions. The dashed lines correspond to the unperturbed wavefunctions centered at lattice sites n and $n + 1$. The symmetric linear combination (red) is larger than the antisymmetric combination (green) in between the lattice sites, and therefore feels the potential due to the neighboring sites more strongly.

which is negative if neighboring lattice site amplitudes c_n and c_{n+1} tend to be in-phase, and positive if they tend to be out-of-phase. The maximum energy decrease occurs when $\Delta_n = 2\pi\mathbb{Z}$, and the maximum increase occurs $\theta_{n+1} = \theta_n + (2n - 1)\pi\mathbb{Z}$. Expressing this in terms of our Bloch waves, the former case corresponds to the center of the Brillouin zone $k = 2\pi/a \sim 0$, while the latter corresponds to the edge of the Brillouin zone $k = \pm\pi/a$.

Physically, some intuition for this is obtained by identifying $|n, \phi_j\rangle$ as an atomic orbital wavefunction ϕ_j centered at the n^{th} lattice site. Turning on the perturbation amounts to allowing the energy of a state localized on site n to be influenced by the neighboring lattice sites $n \pm 1$. If $\Delta_n = 2\pi\mathbb{Z}$ then there is constructive interference in between sites, and the electrons spend more time in the region where the perturbation (i.e. the potential from the neighboring ions) is larger. On the other hand, if $\Delta_n = (2n - 1)\pi\mathbb{Z}$ then the opposite is true. This is illustrated schematically in Fig. 4.6.

4.6 The Wannier Representation

The Bloch states provide a very important tool for understanding the electronic states of a crystalline solid. However, the fact that they are extended throughout the entire crystal makes them both non-intuitive and somewhat unsuitable for certain applications. For example, it is not immediately clear how they respond to the presence of localized disorder, nor is it clear how best to visualize the distribution of charge within the unit cell.

The Wannier states $\mathcal{W}_{n,\mathbf{R}}$ are an exceedingly useful alternative representation [24]. They are defined as

$$\mathcal{W}_{n,\mathbf{R}}(\mathbf{x}) = \frac{1}{\sqrt{\Omega}} \int e^{-i\mathbf{k}\cdot\mathbf{R}} \psi_{n\mathbf{k}}(\mathbf{x}) = \frac{1}{\sqrt{\Omega}} \int e^{i\mathbf{k}\cdot(\mathbf{x}-\mathbf{R})} u_{n\mathbf{k}}(\mathbf{x}) \quad (4.37)$$

where $\mathbf{R}$ is a lattice vector, and Ω is the volume of the Brillouin zone⁷. Given the orthonormalization of the Bloch waves, it is a straightforward exercise to demonstrate that the Wannier functions are also orthonormal:

$$\langle \mathcal{W}_{n\mathbf{R}}, \mathcal{W}_{m\mathbf{R}'} \rangle = \delta_{nm} \delta_{\mathbf{R},\mathbf{R}'} \quad (4.38)$$

Additionally, though the Wannier states are not eigenstates of the Hamiltonian, they do constitute a complete basis for each band. Any filled band n can equally well be described as an electron occupying the extended Bloch wave $\psi_{n\mathbf{k}}$ for each $\mathbf{k} \in \mathbf{BZ}$ or as an electron occupying the localized Wannier orbital $\mathcal{W}_{n\mathbf{R}}$ for each lattice vector $\mathbf{R}$.

The Wannier states provide a real-space (as opposed to k -space) representation of the electronic states, and in that sense are much more physically intuitive than the Bloch states. $\mathcal{W}_{n\mathbf{R}}$ is

⁷If $\mathcal{V}$ is the volume of the d -dimensional lattice's unit cell, then $\Omega = (2\pi)^d/\mathcal{V}$.

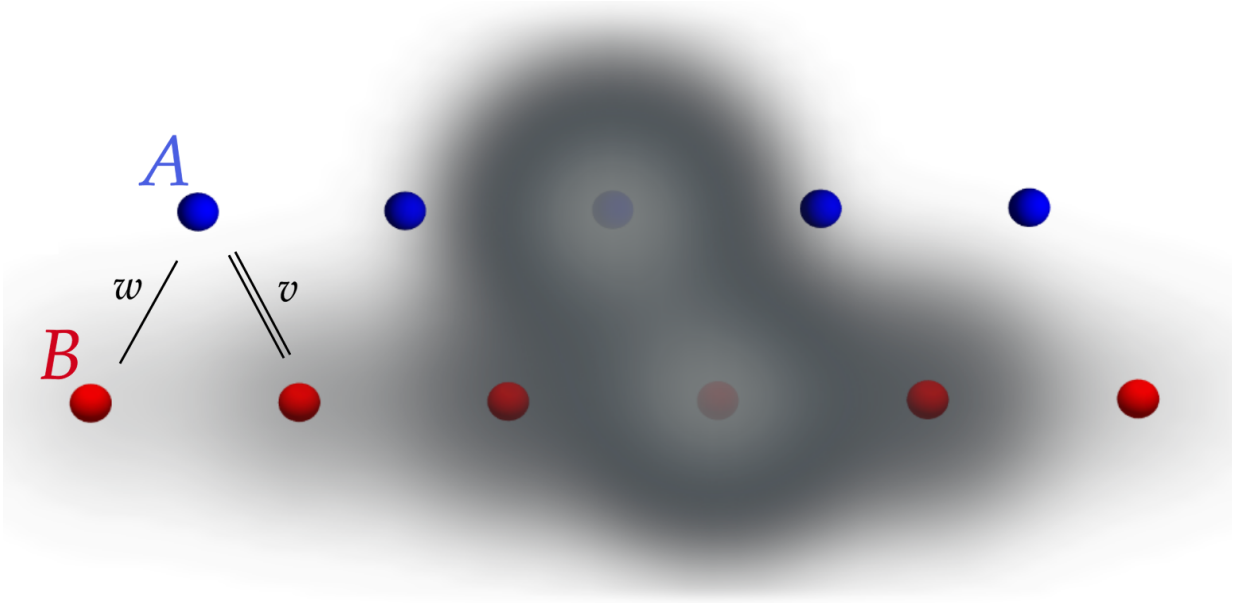


Figure 4.7: A representation of the magnitude of the conduction band Wannier function $\mathcal{W}_{1,0}$ for the bipartite lattice (see Sec. 4.3.2.2), where the intercell hopping amplitude w is half the value of the intracell hopping amplitude v . Whereas the Bloch waves $u_{n\mathbf{k}}$ have uniform probability density on the A and B sublattices, the Wannier representation clearly shows a strong intra-cell bond.

localized to the unit cell at $\mathbf{R}$, and provides a concrete sense of the spatial distribution of n -band electrons, as illustrated in Fig. 4.7. They are frequently used to justify the tight-binding approximation⁸ that led to the lattice formulation of Sec. 4.3, which in turn is usually the preferred context in which to discuss the effects of defects and disorder on the electronic states.

4.6.0.1 Maximally Localized Wannier States

With that being said, there is a substantial ambiguity hiding in our definition of the Wannier functions which is tied to our choice of phase for our Bloch functions $u_{n\mathbf{k}}$. Performing a k -dependent change of phase

$$u_{n\mathbf{k}} \mapsto e^{i\theta(\mathbf{k})} u_{n\mathbf{k}}$$

⁸We did this in a hand-wavy way by simply asserting the existence of a basis $|\mathbf{n}\rangle$ of states localized to the lattice site $\mathbf{n}$; we now recognize this as a Wannier basis.

can have a substantial effect on the shape and nature of the corresponding Wannier functions. If there are N occupied bands under consideration, then this ambiguity takes the form of $N \times N$ unitary matrices $U_{\mathbf{k}}$, since the system is invariant under the transformation

$$u_{n\mathbf{k}} \mapsto \sum_m (U_{\mathbf{k}})_{nm} u_{m\mathbf{k}}$$

As a result, the Wannier functions are highly non-unique. However, it is often desirable to choose the phase function $\theta(\mathbf{k})$ such that the Wannier functions are as localized as possible. It has been known for many years⁹ that exponentially localized Wannier states can always be found in 1D [26]. These *maximally localized Wannier states* are the eigenfunctions of the band-projected position operator

$$\mathbb{X} = \mathbb{P}X\mathbb{P}, \quad \mathbb{P} = \sum_{n=1}^M \int dk |\psi_{nk}\rangle \langle \psi_{nk}| \quad (4.39)$$

which measures the position of the electrons averaged over the M occupied bands. Using the maximally-localized Wannier functions, we have

$$\mathbb{X}|\mathcal{W}_{nR}\rangle = (R + \Delta_n)|\mathcal{W}_{nR}\rangle \quad (4.40)$$

where Δ_n is an R -independent constant called the *Wannier charge center*. Physically, the n -band electrons have an average position which is not generically in the center of the unit cell, and the displacement from the center is given by Δ_n .

⁹It was not until the 21st century that it was shown that exponential localization is always possible in 2D and 3D as long as time-reversal symmetry is obeyed - c.f. [25].

4.6.0.2 Hybrid Wannier States

In higher dimensions, it is generically not possible to find Wannier functions which are exponentially localized in each direction, a fact which can be traced back to the non-commutivity of the projected position operators $\mathbb{X}$, $\mathbb{Y}$, and $\mathbb{Z}$. However, we can define the *hybrid* Wannier functions to be exponentially localized along one direction while possessing a plane wave character in the orthogonal directions. With the appropriate choice of gauge, these states take the form

$$\mathcal{W}_{nR\mathbf{k}_{\parallel}} = \frac{1}{\sqrt{2\pi}} \int dk_{\perp} e^{ik_{\perp}R} |\psi_{n\mathbf{k}}\rangle \quad (4.41)$$

where $k_{\perp}$ defines the direction along which the states are exponentially localized and $\mathbf{k} \equiv (k_{\perp}, \mathbf{k}_{\parallel})$. These states are the eigenvectors of the corresponding projected position operators

$$\mathbb{X}_{\perp} = \mathbb{P}_{\perp} X_{\perp} \mathbb{P}_{\perp}, \quad \mathbb{P}_{\perp} = \sum_{n=1}^M \int dk_{\perp} |\psi_{n\mathbf{k}}\rangle \langle \psi_{n\mathbf{k}}| \quad (4.42)$$

with eigenvalues $R_{\perp} + \Delta_{n\perp}$.

Chapter 5: Physical Applications of Topology II: Topological Band Theory

Having reviewed the fundamentals of the band theory of crystalline solids, we now return to topology. In 1980, Klaus von Klitzing discovered the quantum Hall effect, in which the Hall conductance becomes quantized to spectacular precision in sufficiently strong magnetic fields. von Klitzing was awarded the Nobel prize five years later, and the experimental and theoretical developments since that time have revolutionized condensed matter physics.

There are many ways in which the tools of topology can be leveraged to study quantum matter. In Chapter 3, we saw how defects in classical ordered media can be predicted, studied, and classified through the lens of homotopy; in this chapter, we take a different focus. We first introduce the concept of a vector bundle as a way to reinterpret the space of electronic states. From there, we will study the topology of the ground states of gapped systems by defining a sense in which states can be transported through the bundle, at which point we can topologically distinguish between distinct phases if they cannot be continuously deformed into one another (possibly subject to certain constraints). We conclude this rather abstract mathematical chapter by making contact to the physical interpretations of these phases and to the bulk-boundary correspondence - one of the most beautiful and startling phenomena in physics.

5.1 The Bloch Bundle

5.1.1 The Nontrivial Topology of the Brillouin Zone

Recall from 4 that the eigenstates of a periodic Hamiltonian $H = -\frac{1}{2}\nabla^2 + V$ can be written in the form

$$\psi_{n\mathbf{k}}(\mathbf{x}) = e^{i\mathbf{k}\cdot\mathbf{x}} u_{n\mathbf{k}}(\mathbf{x}) \quad (5.1)$$

where $u_{n\mathbf{k}}$ is periodic on the lattice's unit cell and $\mathbf{k} \in \mathbf{BZ}$ is taken from the Wigner-Seitz cell of the reciprocal lattice, more commonly called the *first Brillouin zone*. Additionally, for any points $\mathbf{k}, \mathbf{k}' \in \mathbf{BZ}$, we were to identify $\mathbf{k} \sim \mathbf{k}' \iff \mathbf{k}' - \mathbf{k} = \mathbf{G}$ for some reciprocal lattice vector $\mathbf{G}$, which implies that the edges of $\mathbf{BZ}$ are to be identified with one another.

This point is subtle, and is worth taking some time to consider carefully. When we derived the Bloch theorem we found that the lattice translation operators T_i were unitary and commuted pairwise with one another, implying that we could find a basis of simultaneous eigenvectors of them all. For any such eigenvector ψ ,

$$(T_j\psi)(\mathbf{x}) = \psi(\mathbf{x} - \mathbf{a}_j) = \Phi_j\psi(\mathbf{x}) \implies \psi(\mathbf{x} - n\mathbf{a}_j) = \Phi_j^n\psi(\mathbf{x}) \quad (5.2)$$

where $\Phi \in \mathbf{U}(1)$ is a phase factor¹. In d spatial dimensions, the set of all such phase factors

¹Previously, we wrote $\Phi_j = e^{i\theta_j}$, but we will see shortly why this form may be more illuminating in the present context.

is

$$\underbrace{U(1) \times \dots \times U(1)}_{d \text{ times}} \simeq (S^1)^d \simeq \mathbb{T}^d \quad (5.3)$$

where $\mathbb{T}^d$ is the d -torus.

A point $(\Phi_1, \dots, \Phi_d) \in \mathbb{T}^d$ specifies an eigenspace of the set of lattice translation operators.

Because the Hamiltonian commutes with all of them, we may specialize to a particular set $\{\Phi_j\}$ and attempt to diagonalize the Hamiltonian there.

Recalling the definition of the primitive reciprocal lattice vectors $\mathbf{b}_j$, we observe that the function

$$u(\mathbf{x}) \equiv \left(\prod_j \Phi_j^{-\mathbf{b}_j \cdot \mathbf{x} / 2\pi} \right) \psi(\mathbf{x}) \quad \mathbf{b}_j \cdot \mathbf{a}_k = 2\pi \delta_{jk} \quad (5.4)$$

is lattice-periodic, because

$$u(\mathbf{x} - \mathbf{a}_j) = \left(\prod_k \Phi_k^{-\mathbf{b}_k \cdot (\mathbf{x} - \mathbf{a}_j) / 2\pi} \right) \psi(\mathbf{x} - \mathbf{a}_j) = \Phi_j^{-1} \Phi_j u(\mathbf{x}) = u(\mathbf{x})$$

It follows that we may write

$$\psi(\mathbf{x}) = \left(\prod_j \Phi_j^{-\mathbf{b}_j \cdot \mathbf{x} / 2\pi} \right) u(\mathbf{x}) \quad (5.5)$$

for some phase factors Φ_j and some lattice-periodic function u .

As it turns out, this seemingly simple step has a substantial hole in it. The problem arises when we attempt to define the functions $\Phi_j^{-\mathbf{b}_j \cdot \mathbf{x} / 2\pi}$. By definition,

$$\Phi_j^{-\mathbf{b}_j \cdot \mathbf{x} / 2\pi} = \exp \left[-\frac{\mathbf{b}_j \cdot \mathbf{x}}{2\pi} \log(\Phi_j) \right] \quad (5.6)$$

and herein lies the problem - the complex logarithm is not single-valued. It is obvious that $\Phi_j = \Phi_j e^{2\pi i}$, but plugging that into Eq. 5.6 leads to the obvious contradiction $e^{-\mathbf{b}_j \cdot \mathbf{x}} = 1$. In order for Eq. 5.6 to be well-defined, we must choose a branch cut in the complex plane. By (arbitrary) convention, we choose the branch cut to be along the negative real axis, and so $\Phi_j = e^{i\theta_j}$, $\theta_j \in [-\pi, \pi)$. Having made that choice, we are justified in writing

$$\psi(\mathbf{x}) = e^{i\mathbf{k} \cdot \mathbf{x}} u(\mathbf{x}) \quad \mathbf{k} \equiv \sum_j \frac{\theta_j \mathbf{b}_j}{2\pi} \quad (5.7)$$

where we identify the set of all such $\mathbf{k}$'s as the familiar **BZ**. Plugging this into the Hamiltonian leads to the familiar Bloch Hamiltonian

$$H_{\mathbf{k}} = e^{-i\mathbf{k} \cdot \mathbf{x}} H e^{i\mathbf{k} \cdot \mathbf{x}} = \frac{1}{2}(-i\nabla + \mathbf{k})^2 + V \quad (5.8)$$

To summarize, each simultaneous eigenspace² of the T_j 's can be labeled by a set of phase factors $\{\Phi_j\}$. The set of such phase factors in d dimensions is $U(1)^d \simeq \mathbb{T}^d$. Given a point in this parameter space **and a choice of branch cut**, we may express an element in the corresponding eigenspace of the translation operators as a Bloch wave. The reason two points $\mathbf{k}, \mathbf{k}'$ which differ by a reciprocal lattice vector are identified as equivalent is because we are using them to specify a particular eigenspace of the T_j 's, and if $\mathbf{k} - \mathbf{k}' = \mathbf{G}$ for some reciprocal lattice vector $\mathbf{G}$ then the eigenspaces they correspond to are the same.

²Because the T_j 's are not a complete set of commuting observables, their eigenspaces are degenerate. That is, the eigenvalues Φ_j do not correspond to a single eigenvector (up to scalar multiples) but rather to entire *eigenspaces*.

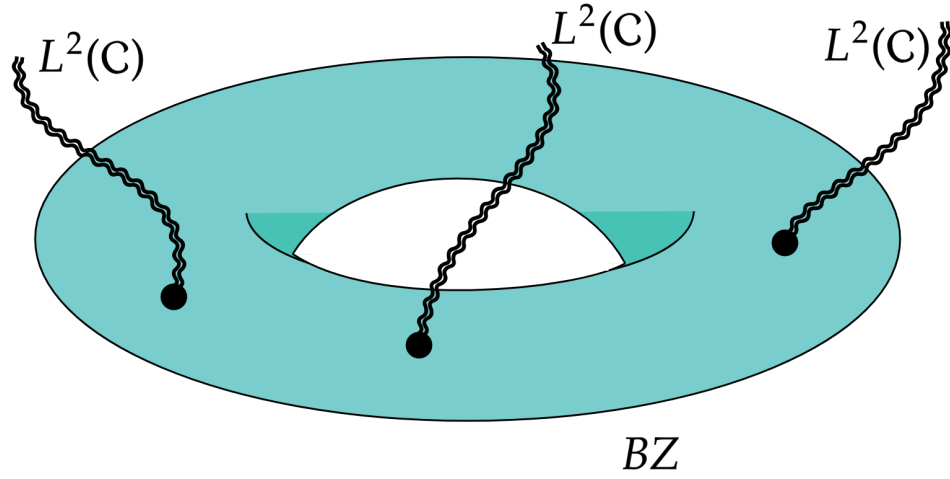


Figure 5.1: A cartoon schematic of the Bloch bundle. The Brillouin zone $\mathbf{BZ}$ is called the *base space*, and the copies of $L^2(\mathcal{C})$ which are attached to each point of $\mathbf{BZ}$ are called the *fibers*. This type of structure is called a fiber bundle; because $L^2(\mathcal{C})$ is a vector space, the Bloch bundle is a special type of fiber bundle called a *vector bundle*.

5.1.2 The Bundle Picture

In this way, the Bloch theorem provides a decomposition of the Hilbert space. An energy eigenvector can be identified by a point $\mathbf{k} \in \mathbf{BZ}$ and a vector $u_{n\mathbf{k}}$ which is an eigenvector of the Bloch Hamiltonian $H_{\mathbf{k}}$. One way to think about this is to imagine that a copy of the Hilbert space³ $L^2(\mathcal{C})$ is attached to each point $\mathbf{k} \in \mathbf{BZ}$ - see Fig. 5.1 for an intuitive cartoon. Note that in an n -band lattice model, $L^2(\mathcal{C})$ can be replaced by $\mathbb{C}^n$.

This construction is called a *fiber bundle* - more specifically, a *vector bundle* [15], because $L^2(\mathcal{C})$ is a vector space⁴.

It's important to emphasize that the copies of $L^2(\mathcal{C})$ attached to each point of $\mathbf{BZ}$ are fully distinct - in precisely the same way that the tangent plane to the north pole of S^2 is distinct

³Recall that $L^2(\mathcal{C})$ is the Hilbert space of square integrable functions on the unit cell .

⁴Even more specifically, a *Hilbert bundle*, because $L^2(\mathcal{C})$ is a Hilbert space.

from the tangent plane attached to a point on the equator⁵. To the copy of $L^2(\mathcal{C})$ attached to a point $\mathbf{k} \in \mathbf{BZ}$, we assign the Hamiltonian $H_{\mathbf{k}}$ - so every fiber is a different copy of $L^2(\mathcal{C})$ equipped with a different Hamiltonian operator, which possesses different eigenvectors $u_{n\mathbf{k}}$ and eigenvalues $E_{n\mathbf{k}}$.

5.1.3 States in the Bloch Bundle

A single-particle state in the full Hilbert space can be identified with a square-integrable function from $\mathbb{R}^d \rightarrow \mathbb{C}$. Expanding a generic state φ in terms of Bloch waves,

$$\begin{aligned}\varphi(\mathbf{x}) &= \frac{1}{(2\pi)^{d/2}} \int d^d k \sum_n f_n(\mathbf{k}) e^{i\mathbf{k} \cdot \mathbf{x}} u_{n\mathbf{k}}(\mathbf{x}) \\ \|\varphi\|^2 &= \int d^d k \sum_n |f_n(\mathbf{k})|^2\end{aligned}\tag{5.9}$$

In the Bloch decomposition, we can interpret this as a map $\mathbf{k} \mapsto \sum_n f_n(\mathbf{k}) u_{n\mathbf{k}}$. This appears to be a vector-valued function of $\mathbf{k}$, but we must be careful. Different points in $\mathbf{BZ}$ are mapped to *different spaces*! This object is not a map from $\mathbf{k} \in \mathbf{BZ}$ to $L^2(\mathcal{C})$ - it is a map from $\mathbf{k} \in \mathbf{BZ}$ to the *copy* of $L^2(\mathcal{C})$ which is attached to that point. Such an object is called a *section* of a fiber bundle, and is diagrammed in Fig. 5.2.

5.1.4 The Valence Bundle

When studying thermodynamically large systems of fermions at temperatures well below the Fermi temperature, to very good approximation we can assume that it is essentially occupying its ground state, in which all states below the Fermi level are occupied and all states above the

⁵This analogy is quite precise - the set of all tangent spaces to a manifold is the *tangent bundle*, another example of a vector bundle.

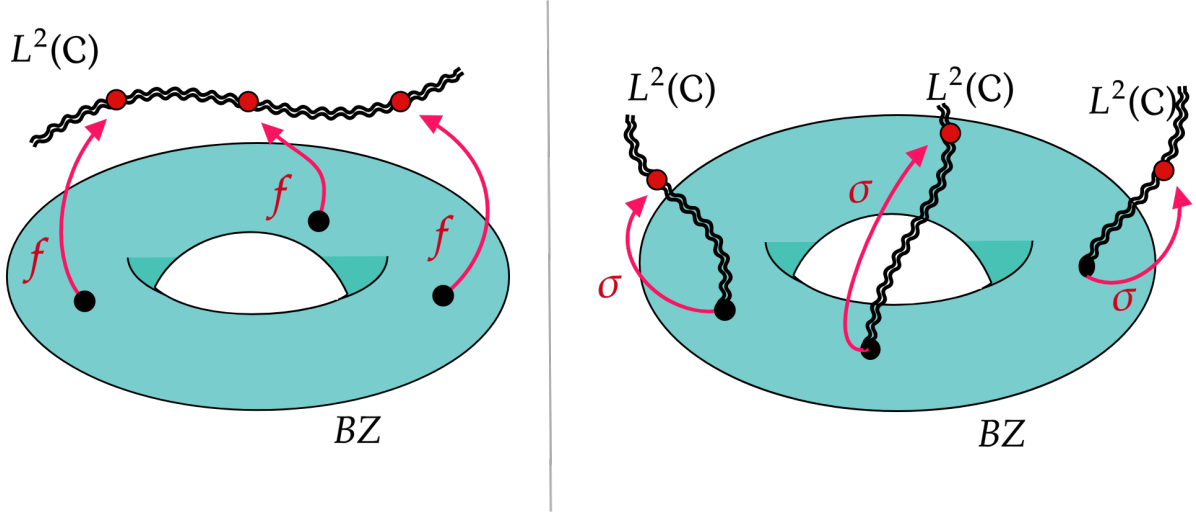


Figure 5.2: A genuine function $f : \mathbf{BZ} \rightarrow L^2(\mathcal{C})$ (left) vs. a section σ of the Bloch bundle (right). f sends all points in $\mathbf{BZ}$ to the same vector space, while σ does not. This has critical implications for differentiation, which fundamentally relies on expressions like $f(\mathbf{k} + d\mathbf{k}) - f(\mathbf{k})$ which don't make any sense if $f(\mathbf{k} + d\mathbf{k})$ and $f(\mathbf{k})$ live in different vector spaces.

Fermi level are vacant. If the system is *gapped* - by which we mean there is a nonzero lower bound Δ to the energy required to excite it from the ground state - then we can unambiguously define the projector

$$\mathbb{P}_{\mathbf{k}} = \sum_{n \text{ occupied}} |u_{n\mathbf{k}}\rangle \langle u_{n\mathbf{k}}| \quad (5.10)$$

onto the space of occupied states at each $\mathbf{k}$. This object contains all of the physical information about the system's ground state - notice that it is in particular the $T \rightarrow 0$ limit of the density matrix⁶

$$\rho_{\mathbf{k}} = e^{-(H - \mu N)/T} / \mathcal{Z} \quad \mathcal{Z} = \text{Tr} [e^{-(H - \mu N)/T}] \quad (5.11)$$

which describes thermodynamic⁷ systems of fermions.

⁶Actually, our $\mathbb{P}_{\mathbf{k}}$ is normalized such that $\text{Tr}[\rho_{\mathbf{k}}]$ is the number of particles with pseudomomentum $\mathbf{k}$, and so physical quantities scale with system size.

⁷I absorb the Boltzmann constant into the temperature in Eq. 5.11.

At each $\mathbf{k} \in \mathbf{BZ}$, the N occupied states span a subspace $\text{Val}_{\mathbf{k}} \subset L^2(\mathcal{C})$. If we attach this *subspace* to each point in $\mathbf{BZ}$, we obtain the *valence bundle*. Formally, the valence bundle is called a *subbundle* of the Bloch bundle.

5.1.4.1 A Note on Gapped Systems

In the present context, the reason we specialize to gapped systems is that for systems without a gap, the projectors $\mathbb{P}_{\mathbf{k}}$ cannot be unambiguously. If there is some $\mathbf{k} \in \mathbf{BZ}$ where the gap vanishes, $\mathbb{P}_{\mathbf{k}}$ becomes discontinuous there - which can be seen by the simple observation that the trace of the projector is the number of occupied states, and the trace is a continuous function. If the number of occupied states changes, that indicates a discontinuity of $\mathbb{P}_{\mathbf{k}}$.

Topological classification of non-gapped systems is still possible [27], but such classification schemes generally take a different approach than the one presented here. In the case of Weyl and Dirac semimetals, the points at which the gap closes are typically excised from the Brillouin zone, in much the same way as we excised points from the spatial domains of the ordered media in Chapter 3. However, for the time being we will restrict our attention to systems with nonzero gaps.

Finally, it is frequently useful from a calculational perspective to assume that the gap is centered at $E = 0$ for all $\mathbf{k} \in \mathbf{BZ}$. For example, this allows us to identify points at which the gap closes with points at which the energy eigenvalues go to zero. This is not a generic feature - c.f. 5.3 - but we can always subtract the mid-gap energy at each $\mathbf{k} \in \mathbf{BZ}$ in a continuous way so that the topological properties of the ground state remain unchanged. For convenience, we will universally assume this has been done.

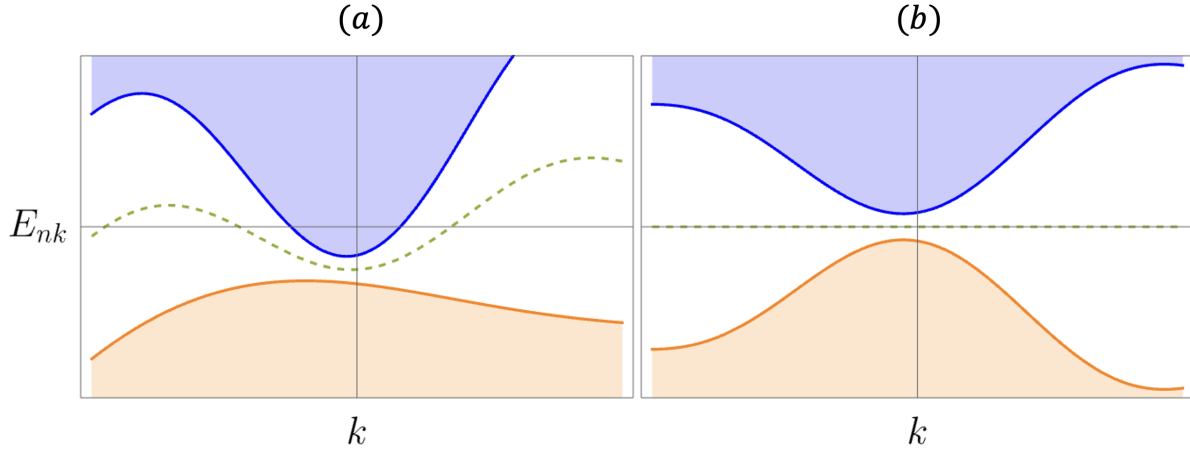


Figure 5.3: A generic band structure before (a) and after (b) subtracting the center of the gap at each point in the Brillouin zone. This can always be done to a gapped system without changing its topology, so we will always assume that the gap is centered at 0.

5.1.5 Transport of States

To explore the structure of a particular ground state, consider the following question. Say you start with some vector ϕ which lives in the space of occupied states at a particular $\mathbf{k}_i \in \mathbf{BZ}$, and would like to transport it to a second point $\mathbf{k}_f$. How do we do this? In fact it is not so simple.

Consider the one-dimensional two-band Hamiltonian

$$H_k = \begin{pmatrix} 0 & \epsilon e^{-ik} \\ \epsilon e^{ik} & 0 \end{pmatrix} \quad E_{\pm k} = \pm \epsilon \quad u_{\pm, k} = \frac{1}{\sqrt{2}} \begin{pmatrix} \pm 1 \\ e^{ik} \end{pmatrix} \quad (5.12)$$

At $k = 0$, consider the state $u_{-,0} = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}$. If we naively move this to a different k point, it will *not* remain in the lower energy subspace. It must be somehow modified as we move it through the Brillouin zone. You could say that we should just follow the eigenstate prescribed as a function of k in 5.12, but this was an arbitrary choice - multiplying $u_{-,k}$ by $e^{i\theta(k)}$ for any

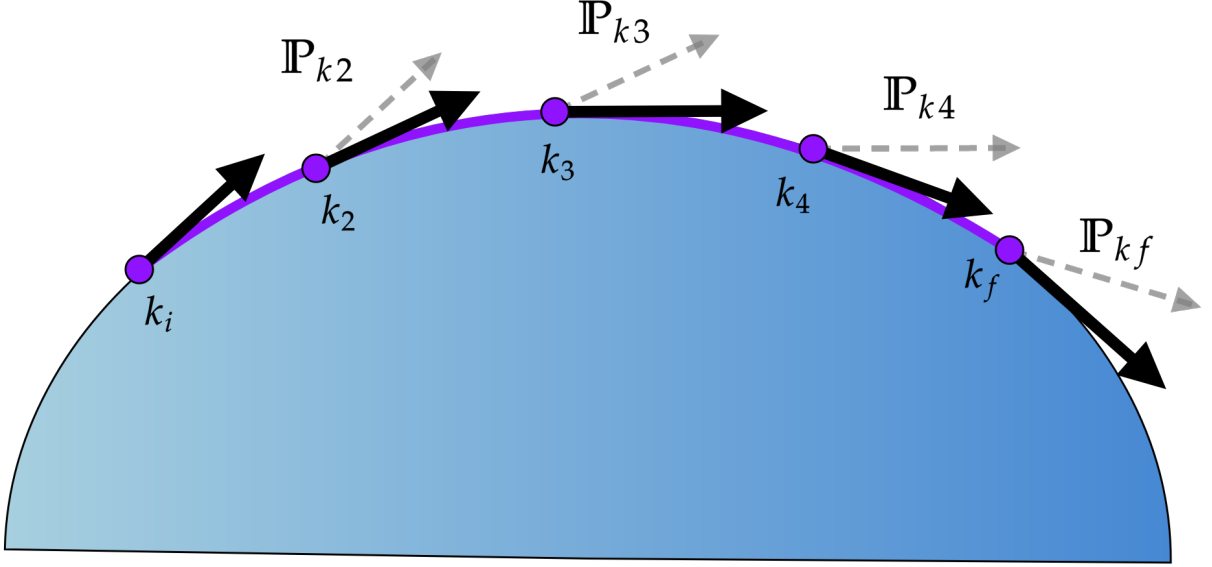


Figure 5.4: The transport procedure defined by the projectors $\mathbb{P}_{\mathbf{k}}$. As the initial vector is moved in infinitesimal steps along the curve, it generically leaves the subspace of occupied states, but the relevant projection operator acts to project it back. In the limit as $|\delta \mathbf{k}_i| \rightarrow 0$, this process preserves the normalization of the initial vector, and the operator $W[\gamma]$ will be seen to be unitary.

2π -periodic function $\theta(k)$ would result in a perfectly good alternative. Therefore, we see that transport through the valence band is not unambiguous, even in the simplest case.

Luckily, there is a prescription we can follow which arises naturally from the fact that the valence bundle is a subbundle of the Bloch bundle. All we need to do is transport the ϕ in the naive way along an infinitesimal step $d\mathbf{k}$, and then use $\mathbb{P}_{\mathbf{k}+d\mathbf{k}}$ to project the result back into the space of occupied states at $\mathbf{k} + d\mathbf{k}$. Repeating this procedure along a curve γ connecting $\mathbf{k}_i$ and $\mathbf{k}_f$ allows us to define a transport operator $W[\gamma]$ which guides our state through the space of occupied states along the curve γ . This process is diagrammed in Fig. 5.4

Before diving into the calculation of this object, note that the result will generically be dependent on which curve we use to connect $\mathbf{k}_i$ and $\mathbf{k}_f$. Whether $W[\gamma]$ is (or is not) independent of path will depend on the system - and will in fact be one of the crucial topological properties of

a particular ground state.

5.1.6 Wilson Lines

Consider first a discretized path consisting of N points $\{\mathbf{k}_1, \dots, \mathbf{k}_N\}$ such that $\mathbf{k}_1 = \mathbf{k}_i$, $\mathbf{k}_N = \mathbf{k}_f$, and $\mathbf{k}_{n+1} = \mathbf{k}_n + \delta\mathbf{k}_n$. The result of the transport process is then given by

$$\begin{aligned}
W[\gamma]|\phi\rangle &= \prod_{n=1}^N \mathbb{P}_{\mathbf{k}_n} |\phi\rangle \\
&= \sum_{m_1} \dots \sum_{m_N} \left(|u_{m_N, \mathbf{k}_N}\rangle \langle u_{m_N, \mathbf{k}_N} | u_{m_{N-1}, \mathbf{k}_{N-1}} \rangle \langle u_{m_{N-1}, \mathbf{k}_{N-1}} | \dots \right. \\
&\quad \left. \dots |u_{m_3, \mathbf{k}_3}\rangle \langle u_{m_3, \mathbf{k}_3} | u_{m_2, \mathbf{k}_2} \rangle \langle u_{m_2, \mathbf{k}_2} | \right) |\phi\rangle
\end{aligned} \tag{5.13}$$

Let us take a closer look at those interior terms. Observe that

$$\langle u_{m, \mathbf{k}} | u_{\ell, \mathbf{k} - \delta\mathbf{k}} \rangle = \langle u_{m, \mathbf{k}} | u_{\ell, \mathbf{k}} \rangle - \delta\mathbf{k} \cdot \langle u_{m, \mathbf{k}} | \nabla_{\mathbf{k}} u_{\ell, \mathbf{k}} \rangle \equiv \delta_{m\ell} - \delta\mathbf{k} \cdot \mathcal{A}_{m\ell}(\mathbf{k}) \tag{5.14}$$

The quantity $\mathcal{A}_{m\ell}$ is called the *Berry connection*. For each m, ℓ , $\mathcal{A}_{m\ell}(\mathbf{k})$ is a d -dimensional vector, where d is the dimensionality of the Brillouin zone (and therefore of the physical space we are working in). Since m and ℓ both range from 1 to the number of occupied bands N_{oc} , $\mathcal{A}$ can be thought of as a d -dimensional vector of $N_{oc} \times N_{oc}$ matrices.

Having made this definition, our transportation operator becomes

$$\begin{aligned}
W[\gamma] &= \left(\mathbb{I} - \delta \mathbf{k}_{N-1} \cdot \mathcal{A}(\mathbf{k}_{N-1}) \right) \dots \left(\mathbb{I} - \delta \mathbf{k}_1 \cdot \mathcal{A}(\mathbf{k}_1) \right) \\
&= \mathcal{P} \left[\prod_{n=1}^{N-1} \left(\mathbb{I} - \delta \mathbf{k}_n \cdot \mathcal{A}(\mathbf{k}_n) \right) \right]
\end{aligned} \tag{5.15}$$

The symbol $\mathcal{P}$ is the so-called *path-ordering* symbol, which reminds us because the terms in the product are matrices, their multiplication generally does not commute. Therefore, it is essential that the terms are multiplied in the correct order, with the points which come later in the path to the left.

In the limit as $N \rightarrow \infty$, $|\delta \mathbf{k}_n| \rightarrow 0$, the transport operator can be formally expressed as a *path-ordered exponential*:

$$W[\gamma] = \mathcal{P} \left[\exp \left(- \int_{\gamma} \mathcal{A}(\mathbf{k}) \cdot d\mathbf{k} \right) \right] \tag{5.16}$$

This object is called a *Wilson line*. The exponential notation is, in some sense, purely formal; naively integrating $\mathcal{A}$ component-wise along γ and then exponentiating the result fails to give the correct answer because the path-ordering prescription was not followed. For short paths or small $\mathcal{A}$'s, one may approximate it as

$$\begin{aligned}
W[\gamma] &\approx \mathbb{I} - \int_0^1 dt \mathcal{A}(\gamma(t)) \cdot \gamma'(t) dt \\
&\quad + \frac{1}{2} \int_0^1 dt_1 \int_0^1 dt_2 \left[\mathcal{A}(\gamma(t_1)) \cdot \gamma'(t_1) \right] \left[\mathcal{A}(\gamma(t_2)) \cdot \gamma'(t_2) \right] \\
&\quad + \dots
\end{aligned} \tag{5.17}$$

where we have parameterized the path as $\gamma : [0, 1] \ni t \mapsto \gamma(t) \in \mathbf{BZ}$ where $\gamma(0) = \mathbf{k}_i, \gamma(1) = \mathbf{k}_f$, and once again we remember that the terms in the expansion are matrices and we must keep the order straight. A somewhat more useful fact is that if we can compute the unique solution to the matrix differential equation

$$\mathcal{M}'(t) = -\left[\mathcal{A}(\gamma(t)) \cdot \gamma'(t)\right]\mathcal{M}(t), \quad \mathcal{M}(0) = \mathbb{I} \quad (5.18)$$

then $W[\gamma] = \mathcal{M}(1)$. It's important to note that if γ_r is the reversed path through the Brillouin zone (i.e. $\gamma_r(t) = \gamma(1 - t)$) then $W[\gamma_r] = W[\gamma]^{-1} = W[\gamma]^\dagger$.

5.1.7 The Role of Gauge Invariance

The underlying idea of this section was to define a consistent notion of the transport of a vector through the space of occupied states. The Wilson line provides such a structure, but in defining it we made use of an arbitrary choice of basis $|u_{n,\mathbf{k}}\rangle$ spanning the space of occupied states at each $\mathbf{k} \in \mathbf{BZ}$. We have not shown that $W[\gamma]$ is independent of this choice of basis - and unfortunately, it is not.

To see this, consider an alternative basis given by $|\tilde{u}_{n,\mathbf{k}}\rangle \equiv U_{\mathbf{k}}|u_{n,\mathbf{k}}\rangle$, where each $U_{\mathbf{k}}$ is a unitary matrix. The Berry connection $\tilde{\mathcal{A}}$ associated to the new basis is related to $\mathcal{A}$ via

$$\tilde{\mathcal{A}} = U_{\mathbf{k}}^{-1}\mathcal{A}U_{\mathbf{k}} + U_{\mathbf{k}}^{-1}\nabla_{\mathbf{k}}U_{\mathbf{k}} \quad (5.19)$$

We may hope that even though the Berry connection depends on our choice of basis, this dependence drops out when computing the Wilson loop. This turns out not to be the case.

Plugging Eq. 5.19 into the differential equation 5.18, we find after a bit of algebra that Wilson line associated to the new basis is

$$\widetilde{W}[\gamma] = U_{\mathbf{k}_f}^{-1} W[\gamma] U_{\mathbf{k}_i} \quad (5.20)$$

The arbitrariness of the choice of basis for the space of occupied states at each $\mathbf{k} \in \mathbf{BZ}$ is an example of a *gauge ambiguity*. It is precisely analogous to the gauge ambiguity in electromagnetism, except for the fact that the set of all possible gauge transformations is merely the set of phase transformations - multiplication by $\mathbb{C}$ -valued functions with unit magnitude.

In contrast, the possible transformations we are presently considering take their values in the unitary group $U(N_{oc})$, which are not $\mathbb{C}$ -valued but rather *matrix*-valued. The fact that they do not commute - as illustrated in our previous calculations - is a profoundly rich source of complexity.

5.1.8 Wilson Loops

Since a choice of gauge is ultimately arbitrary and unphysical, no quantity of direct physical meaning can depend on it. One could begin to worry that our focus on transportation and Wilson lines was all for nothing. However, there *is* a way to construct a gauge invariant physical quantity from the Berry connection, and that is to construct a Wilson *loop* by taking γ to be closed. If the curve γ starts and ends at the same point $\mathbf{k}_0 \in \mathbf{BZ}$, then under a gauge transformation we have that

$$\widetilde{W}[\gamma] = U_{\mathbf{k}_0}^{-1} W[\gamma] U_{\mathbf{k}_0} \quad (5.21)$$

Note that now the matrices on either side of $W[\gamma]$ are inverses of one another, unlike the case of an open Wilson line. This kind of behavior under a gauge transformation is called *gauge covariance*. It is not *invariant* under gauge transformations, but it is getting close - for example, if $W[\gamma] = 0$ in one gauge, then it must vanish in every gauge.

5.1.8.1 The Spectrum of a Wilson Loop

Though $W[\gamma]$ is generally only gauge covariant, its eigenvalue spectrum is gauge *invariant*, since a change of gauge corresponds to conjugation by a unitary matrix. As a result, it is possible that these eigenvalues contain physical information. Being unitary itself, $W[\gamma]$ has eigenvalues which are pure phase factors $z_i \in \text{U}(1), i = 1, \dots, N_{oc}$. Our experience with homotopy suggests that these phases may be a source of topological invariants, which turns out to be true, as we will see shortly.

5.1.8.2 The Trace and Determinant of a Wilson Loop

In particular, gauge-invariant information can be extracted by recalling the cyclic property of the trace operation. Note that

$$\text{Tr}(\widetilde{W}[\gamma]) = \text{Tr}(U_{\mathbf{k}_0}^{-1} W[\gamma] U_{\mathbf{k}_0}) = \text{Tr}(W[\gamma] U_{\mathbf{k}_0} U_{\mathbf{k}_0}^{-1}) = \text{Tr}(W[\gamma]) \quad (5.22)$$

Similarly,

$$\det(\widetilde{W}[\gamma]) = \det(U_{\mathbf{k}_0}^{-1}) \det(W[\gamma]) \det(U_{\mathbf{k}_0}) = \det(W[\gamma]) \quad (5.23)$$

These facts follow from the gauge invariance of the spectrum. Denoting the eigenvalues by

$z_j \simeq e^{i\theta_j}$ for $\theta_j \in [-\pi, \pi)$,

$$\text{Tr}(W[\gamma]) = \sum_{j=1}^{N_{oc}} z_j \quad \det(W[\gamma]) = \prod_{j=1}^{N_{oc}} z_j = \exp \left[i \sum_{j=1}^{N_{oc}} \theta_j \right] \quad (5.24)$$

5.1.9 The Berry Curvature

Before spending a large amount of time attempting to compute holonomies and Wilson loops, we might first ask a basic question - namely, whether there's a way to check if the holonomy on a space is non-trivial in the first place. If, for example, it is possible to choose our gauge in such a way that $\mathcal{A} = 0$, then every holonomy is trivial in that gauge - but because holonomies are gauge-invariant, they must be trivial in *every* gauge.

For concreteness, we work in $d = 3$ spatial dimensions. Let us assume that, given a Berry connection $\mathcal{A}$, there exists a gauge transformation U such that $U^{-1}\mathcal{A}U + U^{-1}\nabla U = 0$. In that case, we have that

$$\mathcal{A} = -(\nabla U)U^{-1}$$

Taking the curl of this expression (being careful to remember that these terms are matrices) yields

$$\begin{aligned} (\nabla \times \mathcal{A})_n &= \epsilon_{n\ell m} \partial_\ell [(\partial_m U)U^{-1}] = \epsilon_{n\ell m} [\partial_\ell \partial_m U]U^{-1} + \epsilon_{n\ell m} [\partial_\ell U]U^{-1}[\partial_m U]U^{-1} \\ &= -\epsilon_{n\ell m} A_\ell A_m = -(\mathcal{A} \times \mathcal{A})_n \end{aligned} \quad (5.25)$$

This motivates us to define the gauge covariant *Berry Curvature* to be

$$\mathcal{F}_{\mathcal{A}} = \nabla \times \mathcal{A} + \mathcal{A} \times \mathcal{A} \quad (5.26)$$

Once again, note that the components in these expressions are matrices. This is why expressions like $\mathcal{A} \times \mathcal{A}$ - which features terms $\mathcal{A}_2\mathcal{A}_3 - \mathcal{A}_3\mathcal{A}_2$ - do not generically vanish.

It is a straightforward exercise to show that in $d = 2$ dimensions, the Berry curvature takes the form $\mathcal{F}_{\mathcal{A}} = \partial_1\mathcal{A}_2 - \partial_2\mathcal{A}_1 + [\mathcal{A}_1, \mathcal{A}_2]$ - essentially just the third component of the expression in three dimensions. In $d = 1$ dimensions, we take $\mathcal{F}_{\mathcal{A}} \equiv 0$.

$\mathcal{F}_{\mathcal{A}}$ can be viewed as a kind of obstruction against locally setting the Berry connection to zero via a gauge transformation. If $\mathcal{F}_{\mathcal{A}} \neq 0$ in some region R , it is not possible to find a gauge transformation that sets $\mathcal{A} = 0$ everywhere on R . Conversely, $\mathcal{F}_{\mathcal{A}} = 0$ implies that there *does* exist a gauge transformation which makes $\mathcal{A} = 0$ everywhere on R , *provided that R is contractible*. On a non-contractible domain, it is possible that $\mathcal{F}_{\mathcal{A}} = 0$ everywhere but $\mathcal{A}$ is necessarily non-zero - and we must bear this in mind, because $\mathbf{BZ} \simeq \mathbb{T}^d$ is not contractible by definition.

5.1.9.1 The Trace and Determinant of the Berry Curvature

Because $\mathcal{F}_{\mathcal{A}}$ is gauge-covariant, we see that just as in the case of the Wilson loop, its eigenvalue spectrum - and therefore both $\text{Tr}(\mathcal{F}_{\mathcal{A}})$ and $\det(\mathcal{F}_{\mathcal{A}})$ - are gauge-invariant. Note that in d spatial dimensions, $\mathcal{F}_{\mathcal{A}}$ is a d -component vector of $N_{oc} \times N_{oc}$. When we talk about the trace

and determinant in this context, we are referring to the N_{oc} degrees of freedom, so that e.g.

$$[\text{Tr}(\mathcal{F}_A)]_{\mathbf{j}} = \sum_{n=1}^{N_{oc}} \left\langle u_{n,\mathbf{k}} \left| \frac{\partial}{\partial k^{\mathbf{j}}} u_{n,\mathbf{k}} \right. \right\rangle \quad (5.27)$$

and similarly with $\det(\mathcal{F}_A)$.

5.1.10 The Abelian Simplification

If we happen to be studying a system in which there is only one occupied band, the Berry connection and curvature reduce to 1×1 matrices - that is, complex numbers. We refer to the Berry connection and curvature as *Abelian*, in reference to the fact that their multiplication operations commute. This drastically simplifies the mathematics. For example, the Berry curvature becomes gauge *invariant*, not merely *covariant*, and it can be given a physical interpretation in its own right. The Wilson line no longer requires a path-ordering symbol, and is the genuine exponential of the integral of the Berry connection along a curve. And so on.

The same is true if there are multiple occupied bands, but we are able to choose a gauge in which the matrix-valued components of the Berry connection and curvature are diagonal. Since the multiplication of diagonal matrices is also commutative, we can simply treat such situations as having an Abelian Berry connection and curvature assigned to each 1×1 .

In contrast, the non-Abelian case is necessary when we cannot guarantee that occupied bands are non-degenerate [28]. In such systems, transport via Wilson lines generally mixes together degenerate eigenspaces. In many realistic materials - certainly those in higher than 1D - the structure of the occupied bands near the Fermi level is simply too complicated for an Abelian model to be accurate. This is a generic feature of many systems, and is absolutely *guaranteed* in

others such as time-reversal invariant systems which feature the Kramers degeneracy. Though it adds to the operational complexity, the conceptual benefits to the non-Abelian generalization are worth the effort.

5.2 The Physicality of the Wilson Loop

In the preceding sections, the Berry connection $\mathcal{A}$ and curvature $\mathcal{F}_{\mathcal{A}}$ and the associated Wannier loop $W[\gamma]$ were introduced from a purely mathematical perspective, where our motivation came from a desire to formulate an unambiguous way to transport a vector through the space of occupied states. While this process does plausibly lead to topological invariants by which we can classify different phases of quantum matter, one could be forgiven for feeling that the aforementioned objects are merely tools for accomplishing that goal, with no particularly physical significance in their own right. As we will see in this section, that is far from the truth. We will focus in particular on the eigenvalue spectra of non-contractible Wilson loops.

5.2.1 The Zak Phase in 1D

The first piece of the puzzle emerges already in the simplest possible model - a two-band Hamiltonian in one dimension. Assume that we've diagonalized our Hamiltonian to obtain the Bloch functions $u_{n\mathbf{k}}$ and their corresponding energies $E_{n\mathbf{k}}$. As mentioned in Sec. 4.6, the Bloch waves are not a particularly useful basis for visualizing the real-space electron probability density. Therefore, we switch to the maximally localized Wannier basis

$$\mathcal{W}_{nR}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} dk e^{ik(x-R)} u_{nk}(x) \quad (5.28)$$

and compute the location of the Wannier charge center

$$\langle \mathcal{W}_{nR}, \mathbb{X} \mathcal{W}_{nR} \rangle \equiv \frac{1}{2\pi} \int_{-1/2}^{1/2} dx \int_{-\pi}^{\pi} dk \int_{-\pi}^{\pi} dq e^{i(k-q)(x-R)} \overline{u_{nq}(x)} u_{nk}(x) \quad (5.29)$$

After some simplification, this takes the remarkable form

$$\langle \mathcal{W}_{nR}, \mathbb{X} \mathcal{W}_{nR} \rangle = R + \int_{-\pi}^{\pi} dk \left\langle u_{nk}, i \frac{d}{dk} u_{nk} \right\rangle \equiv R + i \int_{-\pi}^{\pi} dk \mathcal{A}(k) \quad (5.30)$$

where we recognize $\mathcal{A}_n$ as the Berry connection of the n^{th} band. With this choice of basis, the integral of the Berry connection over the 1D Brillouin zone (called the *Zak phase*) can be interpreted physically as 2π times the Wannier charge center.

5.2.1.1 Gauge Invariance of the Wannier Charge Center

Note that while the Wannier functions themselves are not gauge-invariant, the location of the charge center is - at least up to the lattice constant. Performing the continuous gauge transformation $u_{nk} \mapsto e^{-i\theta(k)} u_{nk}$ (where $e^{i\theta(k)}$ is 2π -periodic) shifts the location of the charge center by

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} dk \theta'(k) \equiv \ell \in \mathbb{Z}$$

where ℓ is the number of times $e^{-i\theta(k)}$ winds around the unit circle in the complex plane. In other words, the effect of the gauge transformation is - at most - to shift all of the charge centers over by some number of lattice vectors. But since *all* of the electrons are being shifted by the same amount, this does not lead to a physically meaningful change.

5.2.1.2 The Wilson Loop

In the language of the previous sections, the Wilson loop

$$W[\gamma] = \mathcal{P} \left\{ \exp \left[\int_{\gamma} dk \mathcal{A}(k) \right] \right\} = e^{-2\pi i \Delta} \quad (5.31)$$

is a 1×1 matrix, whose (eigen)value is the exponential of the Wannier charge center. $W[\gamma]$ is a 1×1 matrix because we only considered a single band, which simplifies the picture substantially (e.g. we don't need to worry about path-ordering or non-commutivity). Also, note that the spectrum of $W[\gamma]$ (which is just $W[\gamma]$ itself) is gauge-invariant because a change of gauge shifts δ by an integer.

Finally, note that in general the Wilson loop can take any value on the complex unit circle. However, if we impose a symmetry constraint - say, inversion symmetry - then this is no longer the case, which we will see in Sec. 5.3.

5.2.2 The Bulk-Boundary Correspondence

The intuition from the 1D case extends readily to higher dimensions. Choosing a direction and splitting $\mathbf{k} = (k_{\perp}, \mathbf{k}_{\parallel})$, we may consider the Wilson loop

$$W(\mathbf{k}_{\parallel}) = \exp \left[\int_{-\pi}^{\pi} dk_{\perp} \mathcal{A}(k_{\perp}, \mathbf{k}_{\parallel}) \right] \equiv e^{2\pi i \Delta_{\perp}(\mathbf{k}_{\parallel})} \quad (5.32)$$

For each value of $\mathbf{k}_{\parallel}$ in the $(d-1)$ -dimensional Brillouin zone, $\Delta_{\perp}$ (the charge center along the $\perp$ direction) takes a generically different value. The single-valuedness of the Wilson loop requires that $\Delta(\mathbf{k}_{\parallel})$ be periodic modulo 1; the charge center increasing by 1 as k is swept across

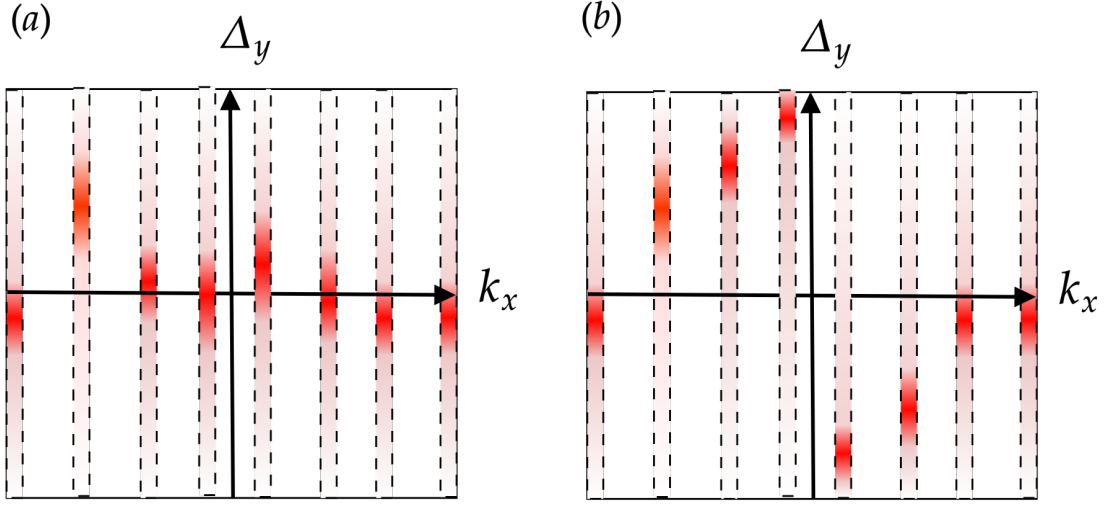


Figure 5.5: The periodic evolution of the Wannier charge center Δ_y as k_x is swept from $-\pi$ to π . The trivial phase (a) can be continuously deformed to $\Delta_y = 0$, but the topological phase (b) cannot, as it undergoes non-zero winding.

the Brillouin zone should be interpreted physically as the charge center entering the next unit cell in the $\perp$ direction while being replaced by the charge center of the preceding cell, as shown in Fig. 5.5.

This winding of the charge center as $\mathbf{k}_{\parallel}$ varies over the $(d-1)$ -dimensional Brillouin zone has dramatic implications for the states which exist at the boundary of the crystal. In the case of non-zero winding, it implies the existence surface modes which cross the Fermi level, and therefore that the surface becomes metallic. This can be argued intuitively by considering the adiabatic transport of states through the Brillouin zone; non-zero winding ℓ implies that during this transport, the physical charge is "pumped" a distance ℓ along the $\perp$ direction. This results in the accumulation of charge on the surface of the crystal - however, the final state is precisely the same as the initial state. This is not possible unless the surface state crosses the Fermi level, at which point an electron is removed from the system and charge neutrality is maintained.

5.2.3 The Surface Dispersion

A more quantitative argument was given in [29]. Given a bulk Hamiltonian

$$H = \sum_n \int d^3k E_{n\mathbf{k}} |\psi_{n\mathbf{k}}\rangle \langle \psi_{n\mathbf{k}}|$$

where the energy is measured with respect to the Fermi level, we can define

$$\begin{aligned} H_\lambda &= \sum_n \int d^3k E_{n\mathbf{k}}(\lambda) |\psi_{n\mathbf{k}}\rangle \langle \psi_{n\mathbf{k}}| \\ &= \underbrace{\sum_{n=1}^M \int d^3k E_{n\mathbf{k}}(\lambda) |\psi_{n\mathbf{k}}\rangle \langle \psi_{n\mathbf{k}}|}_{\text{occupied, } E < 0} + \underbrace{\sum_{n=M+1}^{\infty} \int d^3k E_{n\mathbf{k}}(\lambda) |\psi_{n\mathbf{k}}\rangle \langle \psi_{n\mathbf{k}}|}_{\text{empty, } E > 0} \end{aligned} \quad (5.33)$$

where

$$E_{n\mathbf{k}}(\lambda) = (1 - \lambda)E_{n\mathbf{k}} + \lambda \frac{E_{n\mathbf{k}}}{|E_{n\mathbf{k}}|} \quad (5.34)$$

and $\lambda \in [0, 1]$. When $\lambda = 0$, we have our original Hamiltonian; when $\lambda = 1$, all occupied (resp. empty) states have energy -1 (resp. +1). Therefore, we have that

$$H_1 = 1 - 2\mathbb{P}, \quad \mathbb{P} = \sum_{n=1}^M \int d^3k |\psi_{n\mathbf{k}}\rangle \langle \psi_{n\mathbf{k}}| \quad (5.35)$$

where $\mathbb{P}$ is the projector onto the occupied states. This constitutes a continuous deformation into a *flat-band* Hamiltonian without closing the gap. In a similar way, the flat-band Hamiltonian of the vacuum (in which no states are occupied) takes the trivial form $H_1^{vac} = 1$. This provides us with a convenient way to introduce an edge into the system - an edge perpendicular to $\hat{z}$ can be

modeled as

$$H = 1 - \mathbb{P} + \mathbb{P}V(Z)\mathbb{P} \quad (5.36)$$

where V is a function which approaches -1 as $z \rightarrow -\infty$ and $+1$ as $z \rightarrow \infty$. This models an edge which interpolates between the vacuum in the $z > 0$ region ($H = 1 - \mathbb{P} + \mathbb{P} = 1$) and the bulk in the $z < 0$ region ($H = 1 - \mathbb{P} - \mathbb{P} = 1 - 2\mathbb{P}$).

As a coarse model, one could simply take $V(z) = z$ on the region $z \in [-1, 1]$, at which point we recognize the Hamiltonian in the edge region as

$$H = 1 - \mathbb{P} + \mathbb{P}Z\mathbb{P} \equiv 1 - \mathbb{P} + \mathbb{Z} \quad (5.37)$$

where $\mathbb{Z}$ is the projected position operator from Sec. 4.6. Noting that

$$\mathbb{P} = \sum_{n=1}^M \int d^3k |\psi_{n\mathbf{k}}\rangle \langle \psi_{n\mathbf{k}}| = \sum_{n=1}^M \sum_R \int d^2k_{\parallel} |\chi_{nR\mathbf{k}_{\parallel}}\rangle \langle \chi_{nR\mathbf{k}_{\parallel}}| \quad (5.38)$$

we see that the eigenvectors of the edge Hamiltonian are the hybrid Wannier states (exponentially localized along $\hat{z}$) with eigenvalues, and the spectrum of the edge Hamiltonian is given by the spectrum of $\mathbb{Z}$. The "squashing" of the full spectrum of $\mathbb{Z}$ into a compact interval is shown in Fig. 5.6 for a trivial and topological phase; the latter is distinguished by the presence of winding.

Of course, the flat band Hamiltonian is a continuous (if somewhat dramatic) deformation of our realistic Hamiltonian, and so this is not a quantitatively accurate picture of the edge state dispersion relation. However, its topological features (such as discreteness for a given $\mathbf{k}_{\parallel}$ and continuity as $\mathbf{k}_{\parallel}$ is varied) should be retained as the Hamiltonian is deformed back into a realistic

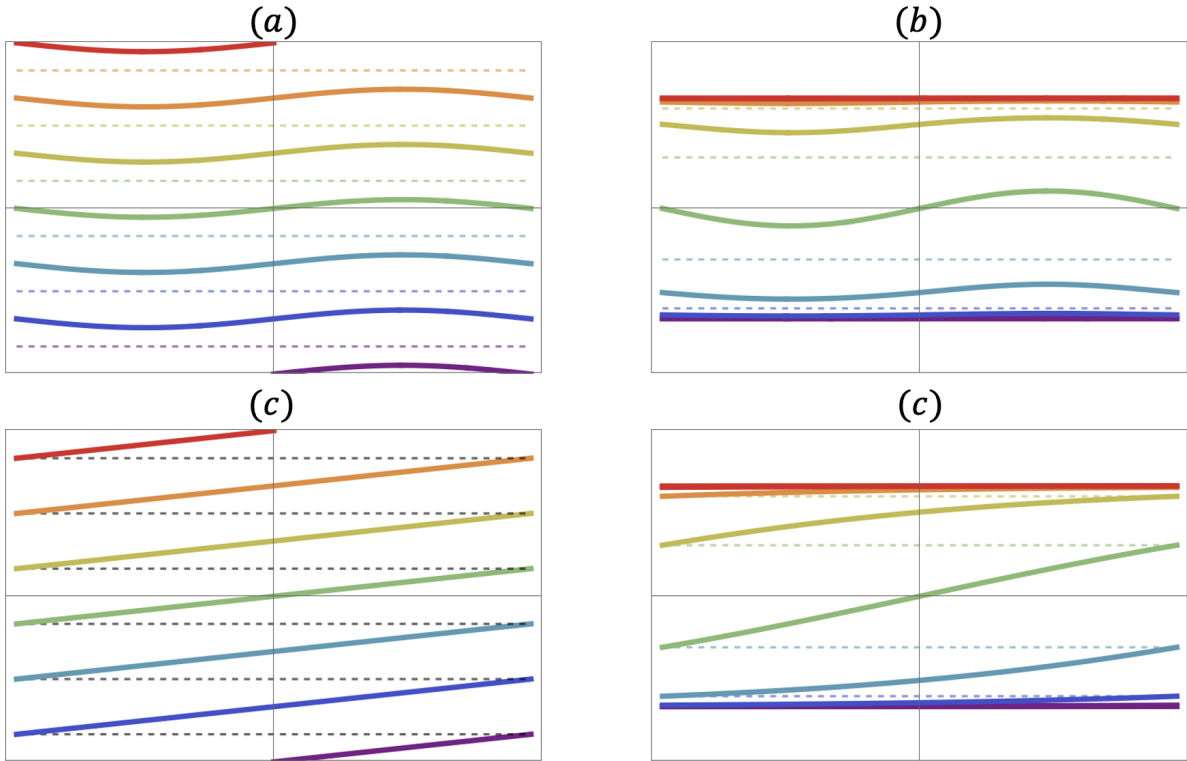


Figure 5.6: (a) In the trivial case, the spectrum of $\mathbb{Z}$ is periodic in $k_{\parallel}$ and exhibits no winding, so when it is squashed into a compact interval (b) there is no guarantee that it will cross the Fermi level. (c) In the trivial case, the spectrum of $\mathbb{Z}$ exhibits winding as $k_{\parallel}$ varies across the **BZ**, and so the squashed edge spectrum (d) *must* intersect the Fermi level.

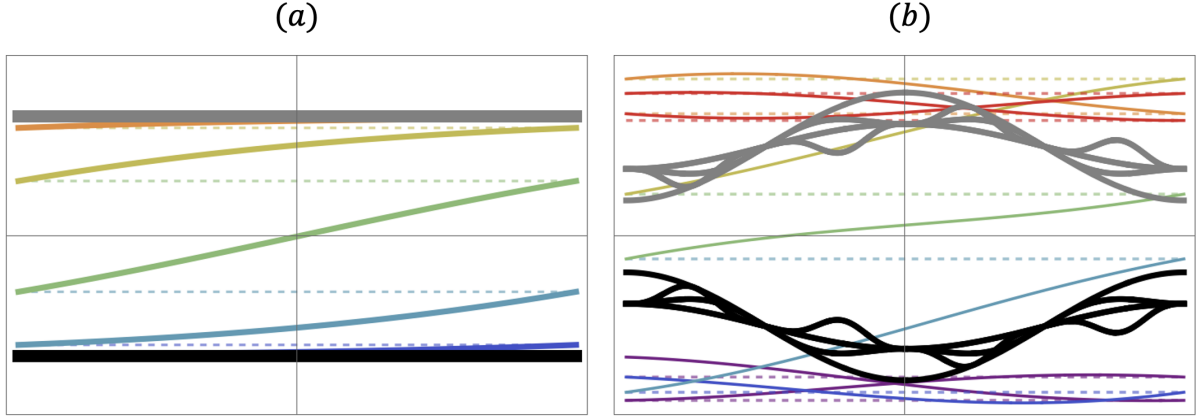


Figure 5.7: As the flat-band Hamiltonian is deformed back into a realistic one, the edge state spectrum should change continuously, in particular retaining its topological properties (e.g. whether or not it exhibits winding).

one, as shown in Fig. 5.7. In particular, if the charge center winds around the unit cell ℓ times as $\mathbf{k}_{\parallel}$ is varied along any line through the $(d-1)$ -dimensional Brillouin zone, then the corresponding surface dispersion along that line should intersect the Fermi level with a signed⁸ intersection number of ℓ .

5.3 Symmetry-Protected Topological Phases

We now come to the impact of symmetries on the existence and character of different topological phases. To recapitulate, we recognize two systems as being topologically distinct if it is not possible to deform one Hamiltonian into the other without closing the (bulk) gap. For systems which obey a particular symmetry, we can refine this question and ask whether one Hamiltonian can be deformed into the other *while preserving the symmetry*. In many cases, it is possible to deform one into the other without closing the gap, but to do so requires violating a

⁸This counts intersections as positive if the curve passes through the Fermi level from below, and negative if it passes through from above. The (absolute) number of intersections is not guaranteed, but their signed total is a topological invariant.

symmetry that they both exhibit. In such cases, we regard them as different, *symmetry-protected* topological phases.

Which symmetry is in play determines the physical character of the phases - in particular, the properties of the edge modes (including whether they exist at all). The connections between a symmetry, the phases it can protect, and the characteristics of the corresponding edge states will now be demonstrated by way of example. An exhaustive classification of symmetry-protected topological phases referred to as the *tenfold way* is presented in [30] - here we seek only to provide several illuminating examples.

5.3.1 The Chern Insulator in 2D

As a warm-up, we first consider a system whose topological properties exist in the absence of any symmetries at all - the 2D Chern insulator. Recall that in a 2D system with a single occupied band, the Berry curvature $\mathcal{F}_A$ is gauge-invariant and given by

$$\mathcal{F}_A = \frac{\partial}{\partial k^1} \mathcal{A}_2 - \frac{\partial}{\partial k^2} \mathcal{A}_1 \quad (5.39)$$

The (*first*) *Chern number* Ch is defined to be

$$\text{Ch} = \frac{1}{2\pi i} \iint_{\text{BZ}} d^2k \mathcal{F}_A(\mathbf{k}) \quad (5.40)$$

Given that the Brillouin zone is boundary-less, an appeal to Stokes theorem⁹, which states that

$$\iint_R \nabla \times \mathcal{A} \cdot d\mathbf{S} = \oint_{\partial R} \mathcal{A} \cdot d\mathbf{S} \quad (5.41)$$

⁹Or more specifically, the 2D variant known as Green's theorem.

might suggest that this quantity is always zero (since the right-hand side trivially reduces to 0 if $\partial R = \emptyset$). The reason it *isn't* (or rather, need not be) is due to the fact that while $\mathcal{F}_A$ is physical and non-singular, $\mathcal{A}$ is gauge-dependent and generally possesses (unphysical) singularities¹⁰. Since this violates the conditions under which Stokes' theorem is valid, the Chern number need not vanish¹¹.

If we consider a 2-band Hamiltonian of the form

$$H_{\mathbf{k}} = \mathbf{d}(\mathbf{k}) \cdot \boldsymbol{\sigma} \quad (5.42)$$

then it is straightforward to see that the eigenvalues and eigenvectors of $H_{\mathbf{k}}$ are given by

$$E_{\pm} = \pm d \quad u_{\pm} = \frac{1}{\sqrt{2d(d-d_3)}} \begin{pmatrix} d_1 \pm id_2 \\ d - d_3 \end{pmatrix} \quad (5.43)$$

where $d \equiv |\mathbf{d}|$. Therefore, the system is gapped as long as $d \neq 0$. The Berry connection and curvature are given via straightforward (but tedious) computation to be

$$\mathcal{A}_{\mu}^{\pm} = i \frac{d_2 \partial_{\mu} d_1 - d_1 \partial_{\mu} d_2}{2d(d \mp d_3)} \quad \mathcal{F}^{\pm} = \pm \frac{i}{2d^2} \mathbf{d} \cdot (\partial_1 \mathbf{d} \times \partial_2 \mathbf{d}) \quad (5.44)$$

From there, we compute the Chern number to be

$$\text{Ch}^{\pm} = \frac{1}{4\pi} \hat{d} \cdot (\partial_1 \hat{d} \times \partial_2 \hat{d}) \quad \hat{d} \equiv \frac{\mathbf{d}}{|\mathbf{d}|} \quad (5.45)$$

¹⁰These singularities are akin to coordinate singularities - e.g. the polar coordinates (r, θ) are singular at $r = 0$, because θ becomes undefined there.

¹¹From that perspective, the Chern number serves as a *topological obstruction* to the existence of a globally smooth gauge; if such a gauge existed, then the Chern number would be trivially zero, and so the non-triviality of Ch is an indicator that no globally smooth gauge is possible.

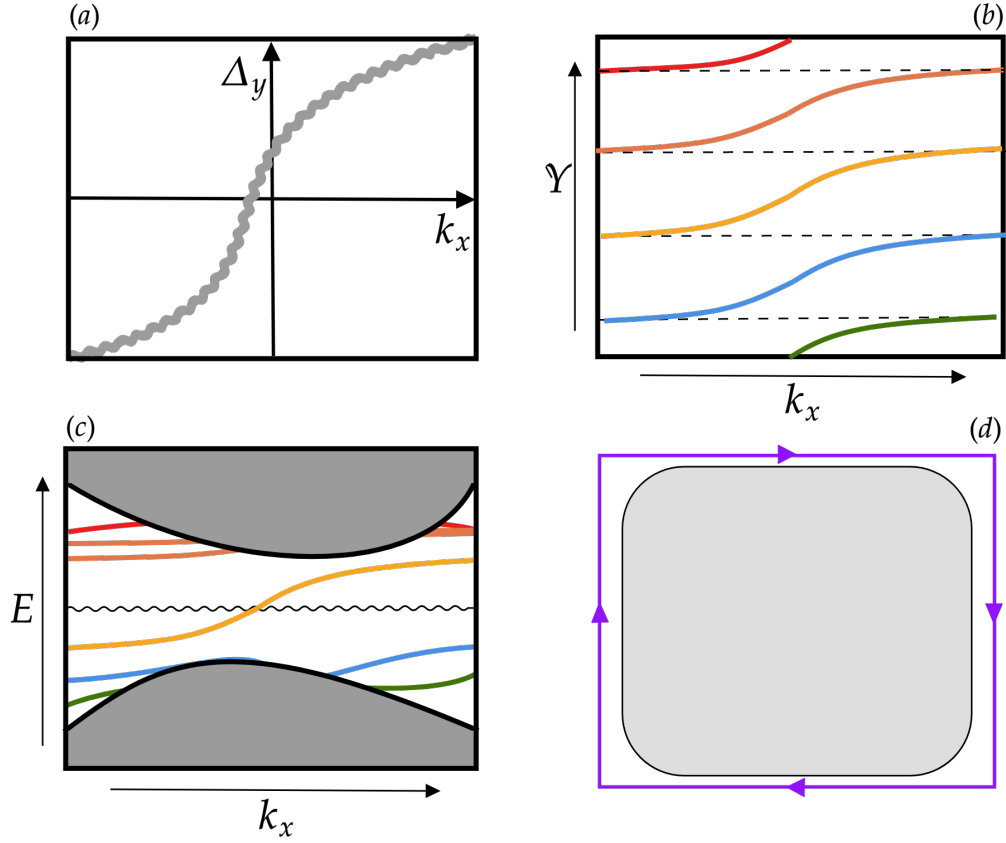


Figure 5.8: The Wannier charge center evolution (a), the spectrum of the projected position operator along y (b), the top edge state dispersion (c), and the helical edge mode (d) of a 2D Chern insulator with Chern number $\text{Ch} = +1$. In (c), the bulk bands are shown in gray, and the Fermi level is denoted by a wavy line.

which we happily recognize as some kind of Skyrmin number (c.f. Sec. 3.2.2). In fact, the vector $\mathbf{d}$ is a map from the 2-torus $\mathbf{BZ}$ to the punctured space $\mathbb{R}^3 - \{0\}$, with the puncture coming from the fact that we forbid the gap to close, and the Chern number is simply the topological degree¹² of this map.

5.3.1.1 Helical Edge States

For the Chern insulator, the bulk-boundary correspondence is manifested in the form of *helical edge states*, which propagate unidirectionally around the edge of the system. By choosing a gauge (or separate gauges) in which the Berry connection is non-singular along the k_x and k_y axes, we see that

$$\begin{aligned} \int_{-\pi}^{\pi} dk_x \mathcal{A}_x(\mathbf{k}) &= \iint_{\mathbf{BZ}} d^2k \mathcal{F}_{\mathcal{A}}(\mathbf{k}) = \int_{-\pi}^{\pi} dk_y \mathcal{A}_y(\mathbf{k}) \\ &= 2\pi i \text{Ch} \end{aligned} \tag{5.46}$$

and so we conclude that at each edge, the edge state dispersion relation intersects the Fermi surface with an intersection number of Ch - indicating Ch edge states propagating clockwise around the edge as seen from above.

5.3.2 Inversion Symmetry

The inversion operator $\mathcal{I}$ is a unitary, idempotent ($\mathcal{I}^2 = 1$) operator formally defined by its action on the position and momentum operators

$$\mathcal{I}\mathbf{X}\mathcal{I} = -\mathbf{X} \quad \mathcal{I}\mathbf{P}\mathcal{I} = -\mathbf{P} \quad \Longleftrightarrow \quad \mathcal{I}|\mathbf{x}\rangle = |-\mathbf{x}\rangle \quad \mathcal{I}|\mathbf{p}\rangle = |-\mathbf{p}\rangle \tag{5.47}$$

¹²See Sec. 2.3.1.3.

This subsequently induces a transformation on the Bloch Hamiltonian:

$$\begin{aligned}
H_{\mathbf{k}} &= U_{\mathbf{k}}^\dagger H U_{\mathbf{k}} \quad \langle x|U|x'\rangle = e^{i\mathbf{k}\cdot\mathbf{x}}\delta(\mathbf{x} - \mathbf{x}') \\
\mathcal{I}H_{\mathbf{k}}\mathcal{I} &= U_{-\mathbf{k}}^\dagger (\mathcal{I}H\mathcal{I})U_{-\mathbf{k}}
\end{aligned} \tag{5.48}$$

If the Hamiltonian is inversion-symmetric about some chosen point $\mathbf{p}$, then we call $\mathbf{p}$ an inversion center. In real space, this implies that $V(\mathbf{p} - \mathbf{x}) = V(\mathbf{x})$. It is convenient to choose coordinates such that $\mathbf{p} = 0$, in which case $V(-x) = V(x)$, $\mathcal{I}H\mathcal{I} = H$, and so $\mathcal{I}H_{\mathbf{k}}\mathcal{I} = H_{-\mathbf{k}}$. In such cases, in the bundle picture of Sec. 5.1, $\mathcal{I}$ constitutes a map from the fiber at $\mathbf{k}$ to the fiber at $-\mathbf{k}$. Additionally, we have that if $H_{\mathbf{k}}|u_{n\mathbf{k}}\rangle = E_{n\mathbf{k}}|u_{n\mathbf{k}}\rangle$, then $\mathcal{I}|u_{n\mathbf{k}}\rangle$ is an eigenvector of $\mathcal{I}H_{\mathbf{k}}\mathcal{I} = H_{-\mathbf{k}}$ with the same energy. Allowing for degeneracies between bands, we have most generally that

$$\mathcal{I}|u_{n\mathbf{k}}\rangle = (\varrho_{-\mathbf{k}})_{nm}|u_{m,-\mathbf{k}}\rangle \tag{5.49}$$

for some unitary matrix $\varrho_{\mathbf{k}}$. Inserting $\mathcal{I}^2$ in the matrix element between adjacent $\mathbf{k}$ points yields

$$\begin{aligned}
1 - \delta\mathbf{k} \cdot \mathcal{A}(\mathbf{k}) &= \langle u_{n\mathbf{k}}|u_{m\mathbf{k}-\delta\mathbf{k}}\rangle = \langle u_{n\mathbf{k}}|\mathcal{I}^2|u_{m\mathbf{k}-\delta\mathbf{k}}\rangle \\
&= \left[\langle u_{n,-\mathbf{k}}|\varrho_{-\mathbf{k}}^\dagger \right] \left[\varrho_{-(\mathbf{k}-\delta\mathbf{k})}|u_{m,-(\mathbf{k}-\delta\mathbf{k})}\rangle \right] = 1 + \delta\mathbf{k} \cdot \tilde{\mathcal{A}}(-\mathbf{k})
\end{aligned} \tag{5.50}$$

where $\tilde{\mathcal{A}}$ is the gauge-transformed Berry connection

$$\tilde{\mathcal{A}}(\mathbf{k}) = \mathcal{A}(\mathbf{k}) + (\nabla_{\mathbf{k}}\varrho_{\mathbf{k}})\varrho_{\mathbf{k}}^\dagger \tag{5.51}$$

This implies that for inversion-symmetric systems we have that $W[\gamma] = \tilde{W}[-\gamma]$ where $\tilde{W}$

is related to W via a gauge transformation. Furthermore, if $-\gamma$ is a curve such that $-\gamma$ corresponds to traversing the path in reverse, then $\widetilde{W}[-\gamma] = \widetilde{W}[\gamma]^\dagger$. Though the Wilson loop is not itself gauge-invariant, its *spectrum* is gauge-invariant, which means that for the aforementioned loops, its eigenvalues must be real (and therefore invariant under conjugation) *or* appear in complex conjugate pairs (which are exchanged under conjugation).

This can be physically understood in terms of the Wannier charge centers. An eigenvalue $1 = e^{i0}$ corresponds to the charge centers exactly on the inversion centers, an eigenvalue $-1 = e^{i\pi}$ corresponds to charge centers exactly in between inversion centers, and eigenvalue pairs $e^{\pm 2\pi i \Delta}$ correspond to symmetrically distributed charge centers at positions $\pm \Delta$. If the system exhibits inversion symmetry, then its charge centers must sit in inversion-symmetric arrangements within the unit cell.

5.3.2.1 Inversion-protected Phases in 1D

In summary, for a system with M occupied bands, the set of Wannier charge centers $\Delta_n, n = 1, \dots, M$ corresponds to the set of M eigenvalues $e^{2\pi i \Delta_n}$ of the Brillouin zone Wilson loop. In an inversion symmetric system, these Δ_n 's are constrained to take the values 0, $1/2$, or to appear in symmetric pairs $\pm \Delta_0$, corresponding to Wilson loop eigenvalues 1, -1 , and $(z, \bar{z}) = e^{\pm 2\pi i \Delta}$, respectively.

The topologically trivial phase is characterized by the charge centers all sitting at the inversion centers, which means that the eigenvalues of W are all $+1$. A topological phase - which cannot be connected to the trivial phase without closing the gap or breaking inversion symmetry - is characterized by a nonzero number of -1 's in the Wilson loop spectrum. To be more pre-

cise, we can resolve the spectrum into $N_{(+)}$ appearances of $+1$, N_p conjugate pairs, and $N_{(-)}$ appearances of -1 ; distinct topological phases are labeled by their value¹³ of $N_{(-)}$.

Observe how critical the inversion symmetry is to the topological protection for these phases. It is intuitively clear that wherever the location of the charges are, one could adiabatically move them to the inversion centers without closing the gap. However, to do so while maintaining inversion symmetry requires that there are no "unbalanced" charge centers sitting in between inversion sites.

5.3.3 Time Reversal Symmetry

Another symmetry of extreme practical importance is time reversal symmetry. The time reversal operator is particularly subtle because it is implemented by an *antiunitary*¹⁴ operator Θ , defined by

$$\Theta \mathbf{X} \Theta^{-1} = \mathbf{X} \quad \Theta \mathbf{P} \Theta^{-1} = -\mathbf{P} \quad \Longleftrightarrow \quad \Theta |\mathbf{x}\rangle = |\mathbf{x}\rangle \quad \Theta |\mathbf{p}\rangle = -|\mathbf{p}\rangle \quad (5.52)$$

as well as the fact that it flips the spin of a particle and - crucially - that $\Theta^2 = -1$ in spin-1/2 systems. The Bloch Hamiltonian transforms as

$$\Theta H_{\mathbf{k}} \Theta^{-1} = \left(\Theta U_{\mathbf{k}}^{\dagger} \Theta^{-1} \right) (\Theta H \Theta^{-1}) (\Theta U_{\mathbf{k}} \Theta^{-1}) = U_{-\mathbf{k}}^{\dagger} (\Theta H \Theta^{-1}) U_{-\mathbf{k}} \quad (5.53)$$

¹³Note that its possible that a conjugate pair could happen to sit at $(z, \bar{z}) = e^{\pm i\pi} = (-1, -1)$ - such pairs must be identified and excluded from $N_{(-)}$.

¹⁴An antiunitary operator has the property that $\langle \Theta \psi, \Theta \phi \rangle = \overline{\langle \psi, \phi \rangle} = \langle \phi, \psi \rangle$. This requires in particular that Θ includes complex conjugation. Additionally, defining the adjoint of an antilinear operator is non-trivial, and so we will not attempt to do so.

where we've made use of the antilinearity of Θ in showing that $\Theta U_{\mathbf{k}} \Theta^{-1} = U_{-\mathbf{k}}$. If $\Theta H \Theta^{-1} = H$, then $\Theta H_{\mathbf{k}} \Theta^{-1} = H_{-\mathbf{k}}$ just as we found for the inversion operator. As a result, Θ can also be understood as a map from the fiber at $\mathbf{k}$ to the fiber at $-\mathbf{k}$. Additionally, we have that if $H_{\mathbf{k}}|u_{n\mathbf{k}}\rangle = E_{n\mathbf{k}}|u_{n\mathbf{k}}\rangle$, then $(\Theta H_{\mathbf{k}} \Theta^{-1})\Theta|u_{n\mathbf{k}}\rangle = E_{n\mathbf{k}}\Theta|u_{n\mathbf{k}}\rangle$. Once again,

$$\Theta|u_{n\mathbf{k}}\rangle = (\tau_{\mathbf{k}})_{nm}|\overline{u_{m,-\mathbf{k}}}\rangle \quad (5.54)$$

where we note the complex conjugation in the latter ket is needed to satisfy antilinearity.

5.3.3.1 Time-Reversal Invariant Momenta

For some special values of $\mathbf{k}$ - those for which $\mathbf{k} = -\mathbf{k}$ - the time-reversal operator maps the fibers to themselves. In 1D, this occurs when $k = 0$ and when $k = \pi$ (since $-\pi \sim \pi$, as the Brillouin zone is a torus). In 2D, this occurs at four distinct points - namely $(k_x, k_y) = (0, 0), (\pi, 0), (0, \pi)$, and (π, π) . At these points, Θ enforces a special degeneracy called the *Kramers degeneracy*.

In the case of inversion symmetry, at the inversion-symmetric momenta we have that $|u_{n\mathbf{k}}\rangle$ and $\mathcal{I}|u_{n\mathbf{k}}\rangle$ have the same energy. However, this does not enforce degeneracy because it is entirely possible that $\mathcal{I}|u_{n\mathbf{k}}\rangle = c|u_{n\mathbf{k}}\rangle$ for some constant c . This is where the antilinearity of the time-reversal operator becomes critical - this is *not true* for Θ , because if $\Theta|u_{n\mathbf{k}}\rangle = c|u_{n\mathbf{k}}\rangle$, then

$$-|u_{n\mathbf{k}}\rangle = \Theta^2|u_{n\mathbf{k}}\rangle = \Theta c|u_{n\mathbf{k}}\rangle = |c|^2|u_{n\mathbf{k}}\rangle \quad (5.55)$$

which is not possible. Therefore, states at time-reversal invariant momenta *must* be (at least)

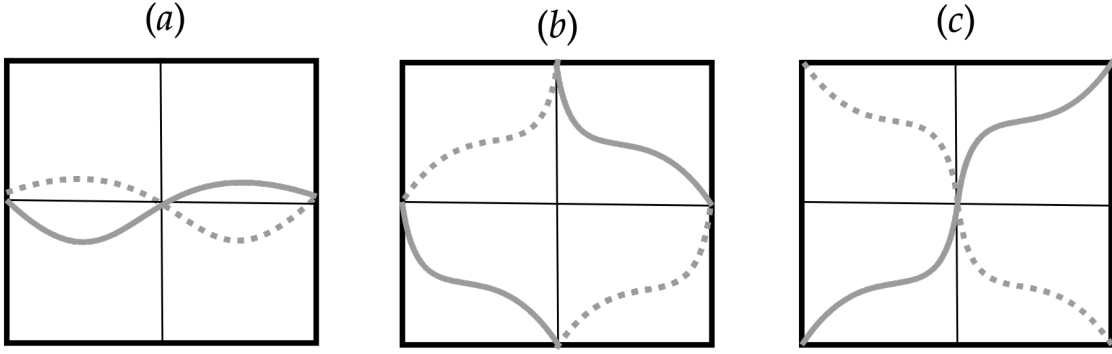


Figure 5.9: Several possible Wilson loop spectra which obey the constraints of time-reversal symmetry - namely, symmetry under $k_x \mapsto -k_x$ and double degeneracy at $k_x = 0, \pi$. The trivial phase (a) can be continuously deformed to $\Delta_y = 0$, but the nontrivial phases (b) and (c) cannot. Note that the non-trivial phases can nonetheless be deformed into one another, which corresponds to the Fu-Kane invariant being an element of $\mathbb{Z}_2$.

doubly-degenerate.

5.3.3.2 The Wilson Loop Spectra

Computing the effect of this degeneracy on the Wilson loop follows the same trajectory as we saw with inversion symmetry. The result is that

$$W[\gamma] = \Theta W[\gamma]^T \Theta^{-1} \quad (5.56)$$

For a Wilson loop integrated along k_y , this implies that the loop at k_x has the same eigenvalue spectrum as the loop at $-k_x$. It also implies that the eigenvalue spectra at $k_x = 0, \pi$ are (at least) doubly-degenerate, using an argument along the same lines as the Kramers degeneracy.

Several possible ways that the Wannier charge centers could evolve are shown schematically in 5.9, assuming a single pair of Kramers-degenerate bands. The requirements imposed

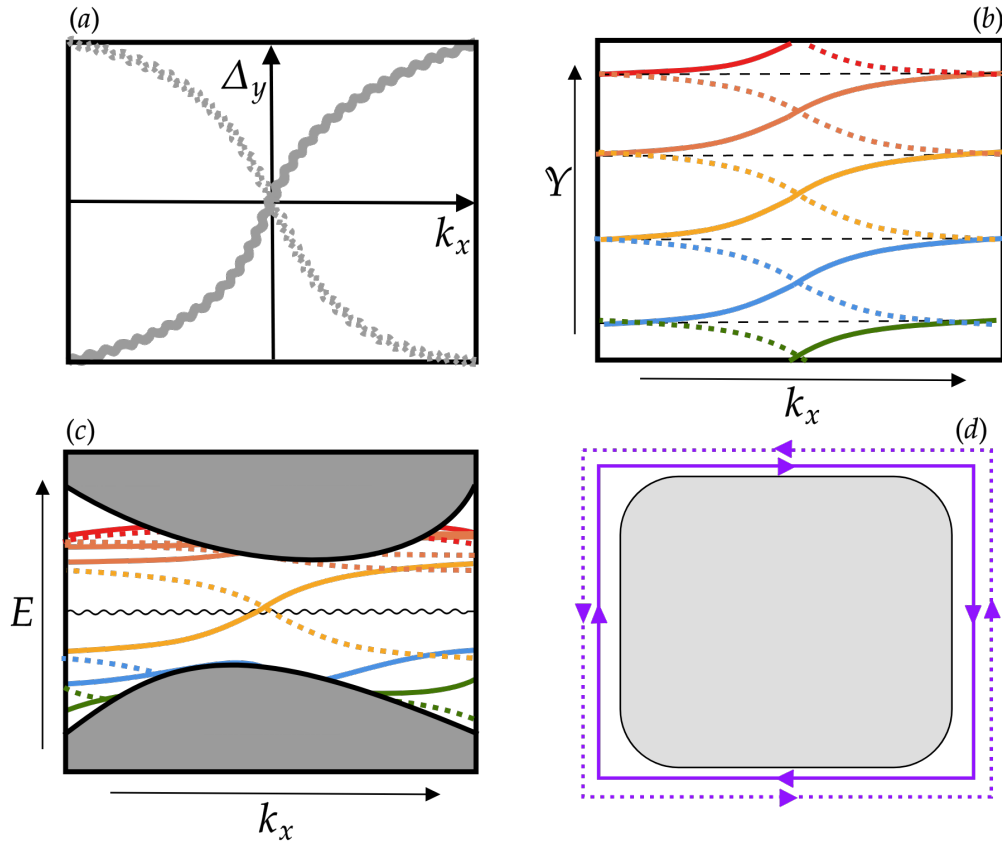


Figure 5.10: The Wannier charge center evolution (a), the spectrum of the projected position operator along y (b), the top edge state dispersion (c), and the chiral edge modes (d) of a 2D time-reversal symmetric topological insulator. The bulk states in (c) are shown in gray, and the Fermi level is denoted by a wavy line. Solid and dashed curves reflect Kramers partners with opposite spin.

by time reversal symmetry imply that there are two possibilities. In the trivial case, a generic horizontal line does not intersect the curves at all, or does so with an even intersection number. In the topological case, it does so with an odd intersection number. It's a good exercise to convince oneself that intersections can appear and disappear under continuous deformation, but only pairwise - the *parity* is a topological invariant, which is referred to as the *Fu-Kane invariant* [31].

The bulk-boundary correspondence for a 2D time-reversal invariant topological insulator requires the existence of gapless, counterpropagating chiral edge states with opposite spin as

sketched in Fig. 5.10. This is due to the fact that the curves which appear in the sketch represent the respective charge centers of Kramers pairs, which are necessarily opposite in spin. In this sense, the 2D topological insulator can be seen as two Chern insulators of opposite Chern number and spin, superimposed on top of one another.

Chapter 6: Scanning Tunneling Microscopy in Theory and Practice

In this chapter, I will discuss the fundamentals of scanning tunneling microscopy from both a theoretical and practical perspective. I will begin by deriving expressions for the tunneling current I and differential conductance dI/dV using a simplified 1D model due to Bardeen [32] and then a quasi-3D model due to Tersoff and Hamman [33]. I will also derive the impact of spin-polarization in the tip and sample. Having established a basic understanding of the theoretical underpinnings of STM, I will conclude this chapter with a more pragmatic discussion of some important aspects of spin-polarized STM in cryogenic conditions.

6.1 Principles of Operation

All scanning tunneling microscopes operate by bringing a sharp¹ metallic probe to within ~ 1 nm of the surface of a sample and applying a voltage bias V_0 between the tip and sample. At this range, electrons are able to tunnel from the tip to the sample or vice-versa, provided that the electronic density of states in the tip and sample have sufficient overlap. The tunneling current $I_t \propto \exp[-\ell/z_0]$, where ℓ is the tip-sample distance and z_0 is a characteristic attenuation length typically on the order of 0.1 nm. As a result, a small change in z corresponds to an exponentially

¹Ideally, the curvature radius of an STM tip is ~ 10 nm or less. However, because of the exponential dependence of the tunneling current on the tip-sample distance, only the nearest few atoms are relevant, and so STM can be performed with surprisingly blunt tips. The issue with this is that unless the surface is atomically flat over distances larger than the tip radius, precisely *which* atoms are closest can suddenly change, leading to misleading image artifacts which can be tricky to diagnose.

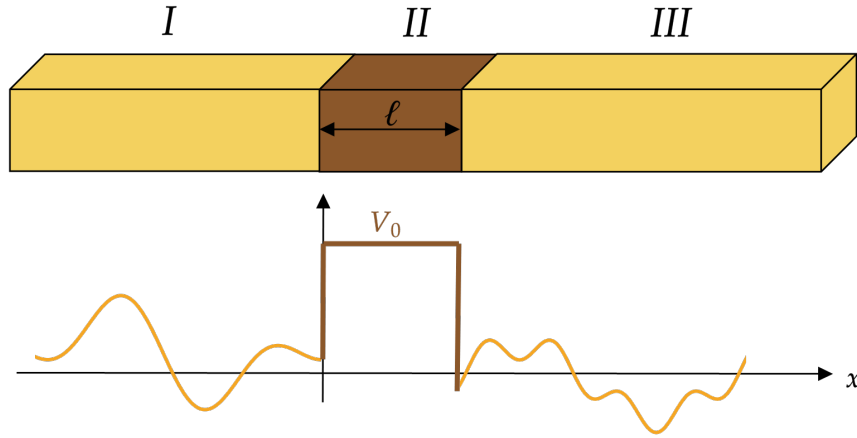


Figure 6.1: A simple model of tunneling due to Bardeen, in which we treat the two electrodes as quasi-independent systems weakly coupled by a barrier potential.

larger change in I_t , providing the feedback signal necessary to control the position of the tip with sub-nanometer precision.

6.1.1 The Bardeen Tunneling Model

A simple but powerful model for understanding the theoretical underpinnings of STM is provided by Bardeen [32], [34]. In this framework, we consider the tip and sample to be quasi-independent systems which are weakly coupled by a tunneling barrier. States which are localized completely within to the tip or the sample are not eigenstates of the full Hamiltonian, so we can use time dependent perturbation theory and Fermi's golden rule to compute the tunneling rate between states.

Starting from 1D for simplicity, we divide the real line into three regions as shown in Fig. 6.1. We interpret regions *I* and *III* as the tip and sample (or more generally, two electrodes) and region *II* as the barrier through which electrons will have to tunnel. The Hamiltonian takes the familiar form $H := -\frac{1}{2} \frac{d^2}{dx^2} + V(x)$.

In order to essentially decouple the electrodes, we make the following definitions:

$$H_A = -\frac{1}{2} \frac{d^2}{dx^2} + V(x)\theta(-x) + V_0\theta(x) \quad (6.1)$$

$$H_B = -\frac{1}{2} \frac{d^2}{dx^2} + V_0\theta(\ell - x) + V(x)\theta(x - \ell) \quad (6.2)$$

$$\Delta = -\frac{1}{2} \frac{d^2}{dx^2} + (V(x) - 2V_0)\theta(x)\theta(\ell - x) \quad (6.3)$$

where we make extensive use of the Heaviside step function

$$\theta(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

The potential functions for the three operators are shown in Fig. 6.2

We interpret H_A and H_B as the Hamiltonians of the two isolated electrodes, and Δ as the barrier Hamiltonian which couples them together. Being self-adjoint operators, H_A and H_B have complete sets of eigenfunctions $\{\psi_n\}$ and $\{\phi_n\}$, with eigenvalues $\{E_n^A\}$ and $\{E_n^B\}$, respectively. These quasi-isolated states decay exponentially outside of their respective electrodes, so we may make the approximation that the ψ 's vanish in region *III* while the ϕ 's vanish in region *I*.

From here, the key idea is to define

$$H_0 \equiv H_A\theta(-x) + H_B\theta(x - \ell) \quad h \equiv \Delta\theta(x)\theta(\ell - x)$$

$$\implies H = H_0 + h$$

The aforementioned approximation implies that both sets of functions $\{\psi_n\}$ and $\{\phi_n\}$ are

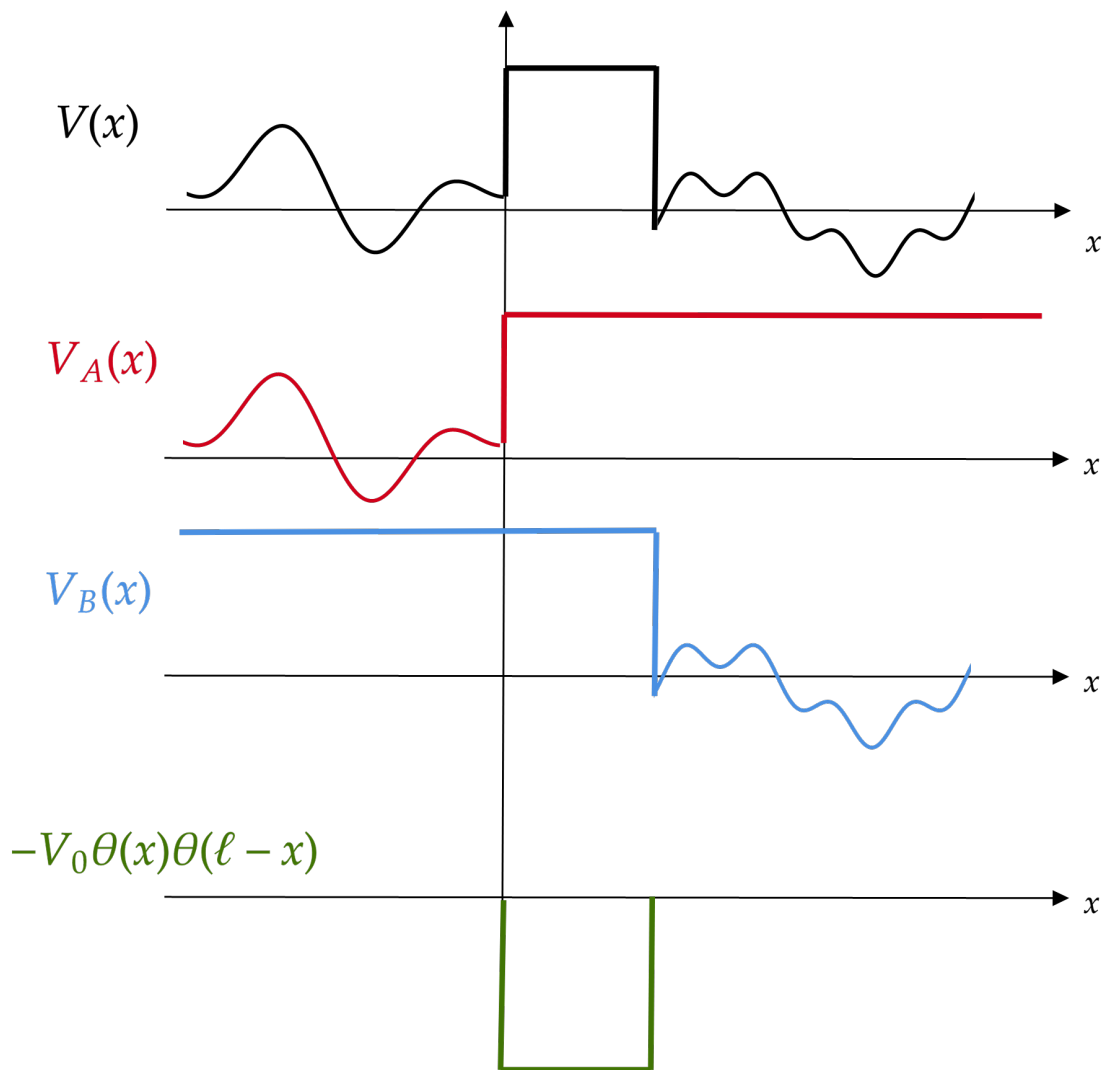


Figure 6.2: The decomposition of the potential $V(x)$ as described in Eq. 6.1. Note that the potentials do not sum to $V(x)$, but $H_A + H_B + \Delta$ does sum to H .

(approximate) eigenfunctions of H_0 . Our goal now is to compute the tunneling rate between states ψ_n and ϕ_m . It is convenient to reformulate the problem in the so-called *interaction picture*. The Schrödinger equation can be rearranged to yield

$$\left(i \frac{d}{dt} - H_0\right) \Psi(t) = h \Psi(t)$$

If we define $\Psi^{(I)}$ via $\Psi(t) = U_0(t) \Psi^{(I)}(t)$ where $U_0(t) \equiv \exp[-iH_0 t]$, then we find that

$$i \frac{d}{dt} \Psi^{(I)}(t) = h_I(t) \psi^{(I)}(t) \quad h_I(t) \equiv e^{iH_0 t} h e^{-iH_0 t} \quad (6.4)$$

By introducing the interaction picture propagator $\mathcal{U}(t)$ such that $\Psi^{(I)}(t) = \mathcal{U}(t) \Psi^{(I)}(0) = \mathcal{U}(t) \Psi(0)$, Eq. 6.4 becomes the operator equation

$$i \frac{d}{dt} \mathcal{U}(t) = h_I(t) \mathcal{U}(t) \quad (6.5)$$

whose formal solution takes the form of a Dyson series

$$\mathcal{U}(t) = \mathbb{I} - i \int_0^t dt' h_I(t') + (-i)^2 \int_0^t dt' \int_0^{t'} dt'' h_I(t') h_I(t'') + \dots \quad (6.6)$$

For our purposes, we will truncate the series after the second term to yield

$$\mathcal{U}(t) \approx \mathbb{I} - i \int_0^t dt' h_I(t') \quad (6.7)$$

By ansatz, let the state vector of the system be given by

$$\Psi(t) = a(t)\psi_n + \sum_k c_k(t)\phi_k \quad a(0) = 1, c_k(0) = 0$$

It follows that

$$c_m(t) = \langle \phi_m, \Psi(t) \rangle - a(t)\langle \phi_m, \psi_n \rangle \approx \langle \phi_m, (\mathcal{U}(t) - 1)\psi_n \rangle \quad (6.8)$$

for small t . Plugging in our approximate expression for the propagator yields²

$$\begin{aligned} c_m(t) &= -i \int_0^t dt' \langle \phi_m, h_I(t')\psi_n \rangle = -i \int_0^t dt' e^{-i(E_n^A - E_m^B)t'} \langle \phi_m, h\psi_n \rangle \\ &= \left(\frac{e^{i(E_n^A - E_m^B)t} - 1}{E_n^A - E_m^B} \right) \langle \phi_m, h\psi_n \rangle = 2ie^{i(E_n^A - E_m^B)t/2} \frac{\sin\left(\frac{E_n^A - E_m^B}{2}t\right)}{E_n^A - E_m^B} \langle \phi_m, h\psi_n \rangle \end{aligned}$$

The corresponding transition rate is

$$\begin{aligned} \Gamma_{n \rightarrow m} &= \frac{d}{dt} |c_m(t)|^2 = 2 \frac{\sin([E_n^A - E_m^B]t)}{E_n^A - E_m^B} |\langle \phi_m, h\psi_n \rangle|^2 \\ &\equiv P_{n \rightarrow m}(t) |\langle \phi_m, h\psi_n \rangle|^2 \end{aligned} \quad (6.9)$$

When computing the total tunneling rate, we will ultimately be integrating over all possible final energies.

²This assumes that $\mathcal{E} \neq E_i$; if $\mathcal{E} = E_i$, then we see straightforwardly that $c_m(t) = -it\langle \phi_m, h\psi_n \rangle$.

6.1.1.1 The Total Tunneling Rate

The rate $\Gamma_{n \rightarrow m}$ can be roughly interpreted as the characteristic rate at which an electron prepared in the state ψ_n of electrode A tunnels into the state ϕ_m of electrode B . In practical terms, however, we are interested in the total current flowing between the electrodes. We must therefore sum (integrate) over all final states. The total rate at which electrons tunnel from the state ψ_n to *any* state in electrode B is given by $\sum_m \Gamma_{n \rightarrow m}$.

To make additional progress, we observe that $P_{n \rightarrow m}(t)$ is sharply peaked around $E_n^A - E_m^B \equiv \Delta E_{nm} = 0$ for large t . More precisely, when we perform the sum over states of electrode B (the "final" states) we find that the primary contribution comes from the energy range $|\Delta E_{nm}| < \pi/t$. For sufficiently large t , the states of electrode B are uniformly distributed over the interval $[E_n^A - \pi/t, E_n^A + \pi/t]$ with density $\rho_B(E_n^A)$, and so we may define an average squared matrix element $\mathcal{M}^2(\psi_n)$ such that

$$\mathcal{M}^2(\psi_n) = \frac{1}{\rho_B(E_n^A) \frac{2\pi}{t}} \sum_{|\Delta E_{nm}| < \pi/t} |\langle \phi_m, h\psi_n \rangle|^2$$

and so

$$\begin{aligned} \sum_m \Gamma_{n \rightarrow m} &= \sum_m P_{n \rightarrow m}(t) |\langle \phi_m, h\psi_n \rangle|^2 \\ &\approx \mathcal{M}^2(\psi_n) \rho_B(E_n^A) \frac{2\pi}{t} \sum_{|\Delta E_{nm}| < \pi/t} P_{n \rightarrow m}(t) \\ &\approx \mathcal{M}^2(\psi_n) \rho_B(E_n^A) \int_{-\infty}^{\infty} 2 \frac{\sin(\epsilon t)}{\epsilon} d\epsilon \\ &= 2\pi \mathcal{M}^2(\psi_n) \rho_B(E_n^A) \end{aligned} \tag{6.10}$$

This is the tunneling rate of an electron prepared in state ψ_n of electrode A assuming that all states in electrode B are accessible. However, the Pauli exclusion principle tells us that tunneling is only permitted from $\psi_n \rightarrow \phi_n$ if ϕ_n is vacant. As a result, we must weight this expression by $(1 - \mathcal{F}(E; \mu_B))$ where $\mathcal{F}(E; \mu_B) = (\exp[(E - \mu_B)/kT] + 1)^{-1}$ is the Fermi-Dirac distribution function and μ_B is the chemical potential of electrode B .

6.1.1.2 The Total Current

The total current due to tunneling from $A \rightarrow B$ is obtained by summing over all occupied initial states, multiplying by 2 to account for spin degeneracy, and multiplying by the electron charge $-e$, which finally yields

$$-4\pi e \sum_n \mathcal{M}^2(\psi_n) \rho_B(E_n^A) \mathcal{F}(E_n^A, \mu_A) (1 - \mathcal{F}(E_n^A, \mu_B))$$

Converting this sum to an integral, we have

$$I_{A \rightarrow B} = -4\pi e \int dE \rho_A(E) \rho_B(E) \mathcal{M}_E^2 \mathcal{F}(E, \mu_A) (1 - \mathcal{F}(E, \mu_B))$$

where as before we define

$$\mathcal{M}_E^2 = \frac{1}{\rho_A(E) \Delta E} \sum_{|E_n^A - E| < \Delta E} \mathcal{M}^2(\psi_n)$$

Similarly, the current due to tunneling from $B \rightarrow A$ is seen to be

$$I_{B \rightarrow A} = 4\pi e \int dE \rho_A(E) \rho_B(E) \mathcal{M}_E^2 (1 - \mathcal{F}(E, \mu_A)) \mathcal{F}(E, \mu_B)$$

and so the total current is given by

$$I = -4\pi e \int dE \rho_A(E) \rho_B(E) \mathcal{M}_E^2 \times \left[\mathcal{F}(E, \mu_A) (1 - \mathcal{F}(E, \mu_B)) - (1 - \mathcal{F}(E, \mu_A)) \mathcal{F}(E, \mu_B) \right] \quad (6.11)$$

Note that if $\mu_A = \mu_B$, the total current 6.11 vanishes (as should be expected). The application of a small bias voltage V serves to shift the chemical potential of the electrodes such that $\mu_A - \mu_B = eV$. Letting $\mu_A \equiv 0$, we obtain the voltage-dependent tunneling current

$$I(V) = -4\pi e \int dE \rho_A(E) \rho_B(E + eV) \mathcal{M}_E^2 \times \left[\mathcal{F}(E) (1 - \mathcal{F}(E + eV)) - (1 - \mathcal{F}(E)) \mathcal{F}(E + eV) \right] \quad (6.12)$$

To further simplify this expression, we may take the zero temperature limit $\mathcal{F}(E) \rightarrow \theta(-E)$ and assume that the (positive) applied voltage is sufficiently small³ that ρ_B is constant over the range $[0, eV]$. In that case, we obtain

$$I(V) \approx -4\pi e \rho_B(0) \overline{\mathcal{M}^2} \int_{-eV}^0 dE \rho_A(E) \quad (6.13)$$

The differential conductance can be computed by differentiating Eq. 6.13 to yield

$$\frac{dI}{dV} \approx -2\pi e^2 \rho_B(0) \overline{\mathcal{M}^2} \rho_A(-eV) \quad (6.14)$$

Unlike the static conductance $G = I/V$, the differential conductance is directly proportional to

³One aspect of a good STM tip is that the density of states of the tip is effectively flat near the Fermi level.

the local density of states ρ_B . As a result, features such as 2-dimensional surface states, quantum wells, and Friedel oscillations appear sharply in dI/dV maps.

6.1.1.3 The Matrix Element

Up until now we have neglected the actual calculation of the matrix element $\langle \phi_m, h\psi_n \rangle$ or any subsequent averages. Recalling the definition of h , we have that

$$\langle \phi_m, h\psi_n \rangle = \int_0^\ell \phi_m^*(x) V_{II}(x) \psi_n(x) dx$$

Applying the WKB approximation yields

$$\psi_n(x) = \psi_n(0) \frac{\sqrt{V_{II}(0) - E}}{\sqrt{V_{II}(x) - E}} \exp \left[- \int_0^x dx' \sqrt{2(V_{II}(x') - E)} \right]$$

$$\phi_m(x) = \phi_m(\ell) \frac{\sqrt{V_{II}(\ell) - E}}{\sqrt{V_{II}(x) - E}} \exp \left[- \int_x^\ell dx' \sqrt{2(V_{II}(x') - E)} \right]$$

As a result, we find that

$$\langle \phi_m, h\psi_n \rangle \propto \exp \left[- \int_0^\ell dx' \sqrt{2(V_{II}(x') - E)} \right]$$

To get an order of magnitude estimate, we may approximate $V_{II}(x') - E \approx \phi$, the work function of the electrodes (which we take to be the same). Restoring SI units, we find that

$$\langle \phi_m, h\psi_n \rangle \sim 2z_0 \phi_m^*(\ell) \psi_n(0) e^{-\ell/2z_0}, \quad z_0 = [8m\phi/\hbar^2]^{-1/2} \approx 0.976 \text{\AA} \left[\frac{\phi}{1 \text{ eV}} \right]^{-1/2} \quad (6.15)$$

Since the total current is proportional to the square of the matrix element, we find at last that $I \propto e^{-\ell/z_0}$ as promised.

6.1.2 Beyond Bardeen: The Tersoff-Hamman Formula

The Bardeen tunneling model is conceptually straightforward, but it lacks some details which are important for practical STM. In particular, it models the tunnel junction as a 1D barrier. The Tersoff-Hamman model [33] extends this analysis to a more realistic model for the tip, yielding an analytic expression for the tunneling matrix element and an approximate measure of the STMs lateral resolution.

In the region outside of the tip, the tip states can be expanded as spherical waves of the form $\phi_n(\mathbf{x}) \approx A_n e^{-k_n r}/r$. Since the Hamiltonian in the exterior is just the free-particle Hamiltonian, it follows that $\hbar^2 k_n^2/2m = E_n$. The function $\Phi_n(r) = e^{-k_n r}/r$ is an eigenfunction of $-\frac{\hbar^2}{2m}\nabla^2$ on the entire region $\mathbb{R}^3 - \{\vec{0}\}$, but is not well-defined at the origin; nevertheless, one can show without excessive trouble that

$$(H_0 - E_n)\Phi_n = \left(-\frac{\hbar^2}{2m}\nabla^2 - E_n\right)\Phi_n = \frac{2\pi\hbar^2}{m}\delta(\mathbf{r}) \quad (6.16)$$

i.e. Φ_n is a Green's function of the operator $H_0 - E_n$. As a result, if we take $\psi_n \approx A\Phi_n$ within the tip region (not just outside it), we find that

$$\langle\phi_n, h\psi_m\rangle = \langle\phi, (H - H_0)\psi_m\rangle \approx \langle\phi, (E_m - H_0)A_m\Phi_m\rangle \quad (6.17)$$

Because Φ_m is a Green's function of $H_0 - E_m$, it follows that the matrix element is simply

$$\langle \phi_n, h\psi_\ell \rangle = \int d^3r \bar{A}_n \delta(\mathbf{r}) \left(-\frac{2\pi\hbar^2}{m} \right) \psi_\ell(\mathbf{r}) = -\frac{2\pi\hbar^2}{m} \bar{A}_n \psi_\ell(0) \quad (6.18)$$

and so finally

$$|\langle \phi_n, h\psi_\ell \rangle|^2 = \left(\frac{2\pi\hbar^2}{m} \right)^2 |A_n|^2 |\psi_\ell(0)|^2 \implies \mathcal{M}^2(\psi_n) = \left(\frac{2\pi\hbar^2}{m} \right)^2 |A|^2 |\psi_\ell(0)|^2 \quad (6.19)$$

where $|A|^2$ is the average of $|A_n|^2$ over a small range of energies near E_n . The conclusion is that if the tip can be modeled locally as a sphere of radius R , then the matrix element and tunneling current are proportional to the square of the wavefunction at the center of the sphere⁴

6.2 Spin-Polarized STM

The tunneling rate derived in Sec. 6.1.1 assumes that the respective densities of states of the tip and sample are independent of spin. However, in general this is not true; in particular, the density of states of a magnetic material should be written $\rho(E) = \rho_\uparrow(E) + \rho_\downarrow(E)$, where $\uparrow / \downarrow$ refers to the component of the electron spin along the axis normal to the interface.

Operating under the assumption that the electron spin does not change during the tunneling process, the tunneling rate depends strongly on the relative magnetization of the tip and sample.

As a result, a magnetic tip (whose magnetization we assume to be fixed) is sensitive to spatially-

⁴Of course, $\psi(0)$ may not actually actually *be* the wavefunction at the center of the sphere - the tip may not be spherical at all. The key is that it can be approximated as a sphere at the junction with the surface; the dependence on $\psi(0)$ is not because the junction is actually sensitive to the wavefunction within the tip, but rather results from an effective averaging procedure at the tip-sample interface.

varying magnetic textures which may be present on the surface of the sample.

6.2.1 Polarized Tunneling

Recall that the total tunneling current for unpolarized tunneling between two electrodes A and B takes the form

$$I(V) \propto \overline{\mathcal{M}^2} \rho_B(0) \int_{-eV}^0 dE \rho_A(E)$$

In the spin-polarized case, we have that

$$\begin{aligned} I(V) \propto \overline{\mathcal{M}_{\uparrow\uparrow}^2} \left[\rho_B^\uparrow(0) \int_{-eV}^0 dE \rho_A^\uparrow(E) \right] &+ \overline{\mathcal{M}_{\downarrow\downarrow}^2} \left[\rho_B^\downarrow(0) \int_{-eV}^0 dE \rho_A^\downarrow(E) \right] \\ &+ \overline{\mathcal{M}_{\uparrow\downarrow}^2} \left[\rho_B^\uparrow(0) \int_{-eV}^0 dE \rho_A^\downarrow(E) \right] + \overline{\mathcal{M}_{\downarrow\uparrow}^2} \left[\rho_B^\downarrow(0) \int_{-eV}^0 dE \rho_A^\uparrow(E) \right] \end{aligned} \quad (6.20)$$

where $\uparrow / \downarrow$ refer to the spin along the local magnetizations $\vec{m}_A$ and $\vec{m}_B$ of the electrodes.

Note that in general, $\vec{m}_A \neq \vec{m}_B$. Assuming a non-magnetic barrier, the matrix elements $\mathcal{M}_{ij}$ are proportional to the projection of the j^{th} spinor in electrode A onto the i^{th} spinor in electrode B . Choosing our polar axis to be along $\vec{m}_A$, we can specify the direction of $\vec{m}_B$ with polar coordinates (θ, ϕ) , in which case our spinors become

$$\begin{aligned} |A, \uparrow\rangle &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} & |A, \downarrow\rangle &= \begin{pmatrix} 0 \\ 1 \end{pmatrix} & |B, \uparrow\rangle &= \begin{pmatrix} \cos(\theta/2) \\ \sin(\theta/2)e^{i\phi} \end{pmatrix} & |B, \downarrow\rangle &= \begin{pmatrix} -\sin(\theta/2)e^{-i\phi} \\ \cos(\theta/2) \end{pmatrix} \end{aligned} \quad (6.21)$$

The tunneling matrix then becomes

$$M = \begin{pmatrix} M_{\uparrow\uparrow} & M_{\uparrow\downarrow} \\ M_{\downarrow\uparrow} & M_{\downarrow\downarrow} \end{pmatrix} = M_0 \begin{pmatrix} \cos(\theta/2) & \sin(\theta/2)e^{i\phi} \\ -\sin(\theta/2)e^{-i\phi} & \cos(\theta/2) \end{pmatrix} \quad (6.22)$$

We can quantify the polarization of tunneling electrons by defining

$$N^{\uparrow/\downarrow} = \int_{-eV}^0 dE \rho^{\uparrow/\downarrow}(E) \quad N \equiv N^{\uparrow} + N^{\downarrow} \quad P \equiv (N^{\uparrow} - N^{\downarrow})/N$$

With these definitions and some algebra, the tunneling current can be expressed as

$$I(V) = -4\pi e^2 V \overline{\mathcal{M}_0^2} N_A N_B (1 + P_A P_B \cos(\theta)) \quad (6.23)$$

Therefore, the tunneling conductance $G = I/V \propto (1 + P_A P_B \cos(\theta))$ and, as long as both electrodes are magnetized (i.e. $P_A, P_B \neq 0$), a spatially-varying magnetic texture can be observed in topography.

6.2.2 Differential Conductance

The differential conductance dI/dV exhibits spin sensitivity as well. Differentiating Eq.

[6.20](#) yields

$$\begin{aligned} \frac{dI}{dV} &\propto \overline{\mathcal{M}_{\uparrow\uparrow}^2} \rho_B^{\uparrow}(0) \rho_A^{\uparrow}(-eV) + \overline{\mathcal{M}_{\downarrow\downarrow}^2} \rho_B^{\downarrow}(0) \rho_A^{\downarrow}(-eV) \\ &\quad + \overline{\mathcal{M}_{\uparrow\downarrow}^2} \rho_B^{\uparrow}(0) \rho_A^{\downarrow}(-eV) + \overline{\mathcal{M}_{\downarrow\uparrow}^2} \rho_B^{\downarrow}(0) \rho_A^{\uparrow}(-eV) \end{aligned} \quad (6.24)$$

Defining

$$p = \frac{\rho^\uparrow - \rho^\downarrow}{\rho^\uparrow + \rho^\downarrow}$$

and repeating the previous derivation yields

$$\frac{dI}{dV} = -2\pi e^2 \overline{\mathcal{M}_0^2} \rho_A(0) \rho_B(-eV) (1 + p_A p_B \cos(\theta)) \quad (6.25)$$

Just as in Eq. 6.23, the differential conductance exhibits dependence on the angle θ between the tip and sample magnetization vectors. The major difference is that the P_A and P_B which appear in the former should be interpreted as *average* degree of polarization over the energy interval $[-eV, 0]$, whereas the p_A and p_B which appear in the latter corresponds to the degree of polarization precisely at the Fermi energy and $E = -eV$, respectively.

Because the average of a non-constant quantity is always strictly less than its maximum, a careful choice of the scanning voltage V can often make spin-polarization appear more clearly in dI/dV than in topography. More importantly, assuming a clean and chemically-homogeneous surface, the differential conductance should in principle be constant over regions across which the topography varies (e.g. across steps and terraces), which can make spin-polarization much easier to identify - see Fig. 6.3.

6.2.3 Establishing a Spin-Polarized Tip

There are a number of methods which may be employed to achieve a spin-polarized STM tip, each with pros and cons which depend on the samples being studied and on the details and capabilities of the STM system being used.

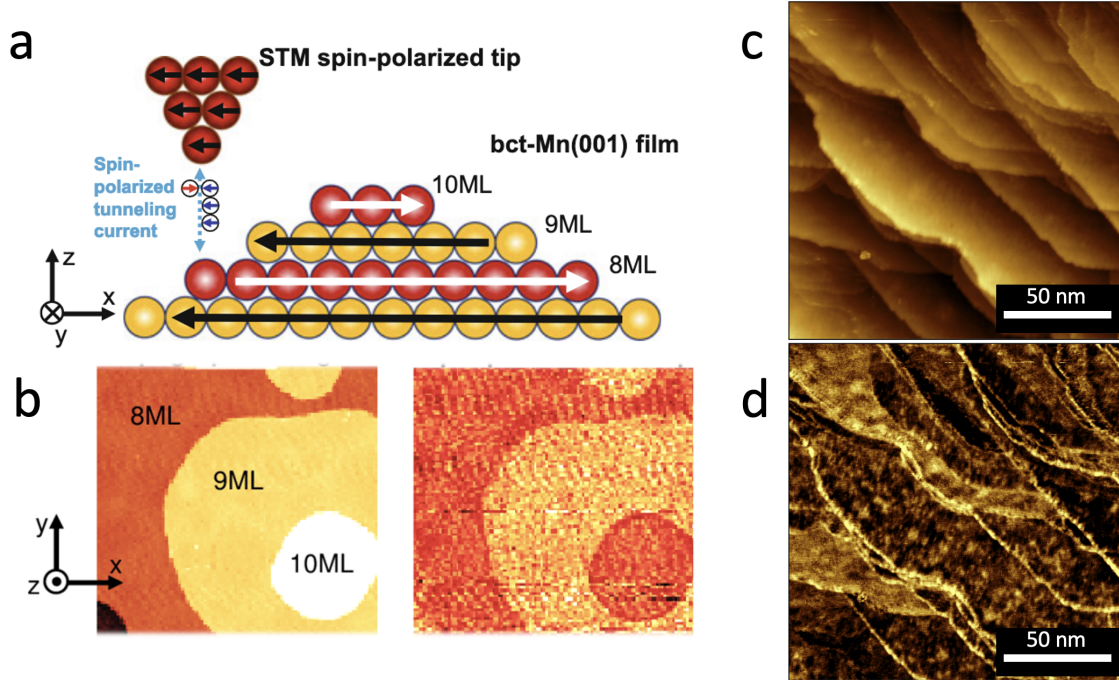


Figure 6.3: Examples of spin-polarized STM. (a) is a cartoon depiction of SPSTM with an in-plane polarized tip on a body centered tetragonal (bct) Mn(001) film. The step-wise antiferromagnetism of bct-Mn(001) results in an enhanced tunneling current when the tip and sample magnetizations are parallel and a suppression when they are antiparallel. (b) shows the topography (left) and differential conductance (right) of a real scan performed with a Fe-coated tungsten tip, taken at +0.2 V. The spin-polarization is not immediately obvious in topography, but is manifested as a very clear pattern of alternating bright and dark steps in dI/dV . (c) and (d) are the topography and differential conductance of a scan I performed with a bulk Cr tip on a Cr(001) single crystal, in which the characteristic, step-wise pattern is visible in dI/dV . (a) and (b) are reused with permission from [1].

6.2.3.1 The Dunking Method

The simplest possibility to implement on any STM system involves performing a controlled crash of an ordinary STM tip into a magnetic material - most commonly iron or nickel - with the goal of picking up a magnetic cluster at the end of the tip.

The main advantage of this approach is that it can be implemented with any STM with no change in instrumentation, and many studies (e.g. [35]) have been performed very effectively with this technique. However, it does suffer from a number of disadvantages. First, it can be difficult

to verify that a magnetic cluster has been picked up in the first place. Ideally one could scan the magnetic sample after the crash and search for e.g. domain walls, but if the characteristic size of magnetic structures in the sample are comparable to or greater than the STM's lateral scan range then this may present a challenge. Once a cluster has been obtained, the magnetic sample needs to be exchanged with the sample being studied and, depending on the nature of the exchange (e.g. the coarse motion of the tip and the subsequent approach on the new sample), the cluster could become dislodged. Finally, by its nature the cluster is unlikely to be atomically sharp, and a large, irregular cluster could introduce imaging artifacts (i.e. multiple tips) and blur features of interest.

6.2.3.2 The Evaporation Method

A more precise approach is to evaporate a thin layer ($\sim 1 - 10$ nm) of magnetic material onto a clean, otherwise non-magnetic tip. A common choice is to evaporate Fe on a tungsten tip [1, 36]. This method avoids the inherent instability and unpredictability of the cluster technique. However, it requires the STM system to feature *in vacuo* tip exchange and the ability to deposit material onto the tip prior to its installation in the STM. Additionally, this approach requires a far larger volume of magnetic material to be present on the tip. If the evaporated material is ferromagnetic, the subsequent stray field is typically extremely difficult to characterize precisely, and could adversely affect the measurement of a magnetically soft sample.

6.2.3.3 The Bulk Tip Method

A third option is to simply etch an STM tip out of a magnetic material. Common choices include ferromagnets iron and nickel [37] as well as antiferromagnets such as chromium and MnNi alloys [38]. This technique has the advantage that any modification to the tip production process occurs *ex vacuo*. However, bulk ferromagnetic tips typically possess an extremely large stray field and are inappropriate for studying any but the magnetically hardest of samples, while antiferromagnetic materials tend to be brittle and difficult to handle or etch.

6.2.4 Extracting a Spin-Polarized Signal

Even with a well-polarized tip and sample, SPSTM is notoriously difficult in practice. It is therefore important to tune the experimental parameters so as to maximize the magnetic contrast.

As mentioned in 6.2.2, the maximal magnetic contrast is typically observed in dI/dV . One approach is to choose a set point voltage V_0 and perform a constant-current scan using a lock-in amplifier (LIA). The LIA adds a sinusoidal modulation $V_{mod} \sin(\omega t)$ to V_0 . The tunneling current then takes the form

$$I(t) \approx I_0 + \frac{dI}{dV} V_{mod} \sin(\omega t - \delta) + I_{noise} \quad (6.26)$$

where the phase shift δ is due to the finite capacitance of the tip-sample junction, and I_{noise} represents random electronic noise. The LIA is able to extract dI/dV by multiplying $I(t)$ by $\sin(\omega t - \delta)$ and averaging over several periods; both I_0 and I_{noise} approximately average to zero.

To optimize this measurement, we need to choose the correct values for V_{mod} , ω , and δ , as well as the set points V_0 and I_0 .

- A larger V_{mod} means a larger modulation, but it also means that the signal is being averaged over a larger voltage range.
- ω should be chosen to be as large as possible (thereby increasing the number of cycles being averaged over) without getting too close to the sample rate of the STM control unit. It is also crucially important that ω be as far as possible from the prominent spikes of the electronic and vibrational noise spectrum - if there is noise with the same frequency as the LIA, then that component will not average out of the signal. A reliable spectrum analyzer should be used to choose an optimal value for ω .
- δ should be chosen to maximize the dI/dV signal when it is out of tunneling. The standard procedure to find the right value of δ is to pull the tip back by a few nm and turn off the z -feedback. From there, the tunneling current should be purely sinusoidal and reflect the capacitive charging and discharging of the tunnel junction. When δ is chosen correctly, the demodulated signal is as large as possible.
- The spin-splitting of the tip and sample densities of states is energy-dependent. V_0 should be chosen to find the largest possible splitting, therefore yielding the largest values of p_A and p_B as defined in 6.25. Larger values of the current set point I_0 result in a larger signal, but also a smaller junction which is more susceptible to disorder, mechanical feedback oscillations, and tip crashes.

All of these factors, along with scan speed and spatial resolution, must be taken into account to obtain a good SPSTM measurement. Of course, as with all STM measurements electronic and vibrational noise should be reduced as much as possible. Once the experimental parameters

have been optimized, differential spectroscopy scans like the ones shown in Fig. 6.3 become accessible.

6.3 STM at Cryogenic Temperatures

Cryogenic conditions are an important tool for many aspects of STM. While the instrument works perfectly well at room temperature and above, the greatly improved vacuum quality and increased stability of the sample surface and tunnel junction which come with a cryogenic environment can provide a drastic improvement in both topographic and spectroscopic measurements. For example, the piezoelectric crystals which control the motion of the tip suffer from hysteresis which is significantly reduced at cryogenic temperatures.

Beyond improved scanning capabilities, many systems of great practical interest - particularly superconductors⁵ and magnetic systems with low Curie temperatures - require cryogenic conditions for their novel properties to emerge. Additionally, many cryostats which are equipped with superconducting magnets require temperatures on the order of 10 K or less in order to perform magnetic studies.

6.3.1 Cooling Mechanisms

A variety of options are available to achieve cryogenic temperatures, ranging from the straightforward submersion of the vacuum can containing the cryostat into a liquid cryogen to the operation of a complex refrigeration system. In this section we discuss those options in increasing order of complexity.

⁵This includes high T_c superconductors, which still require temperatures on the order of 77 K.

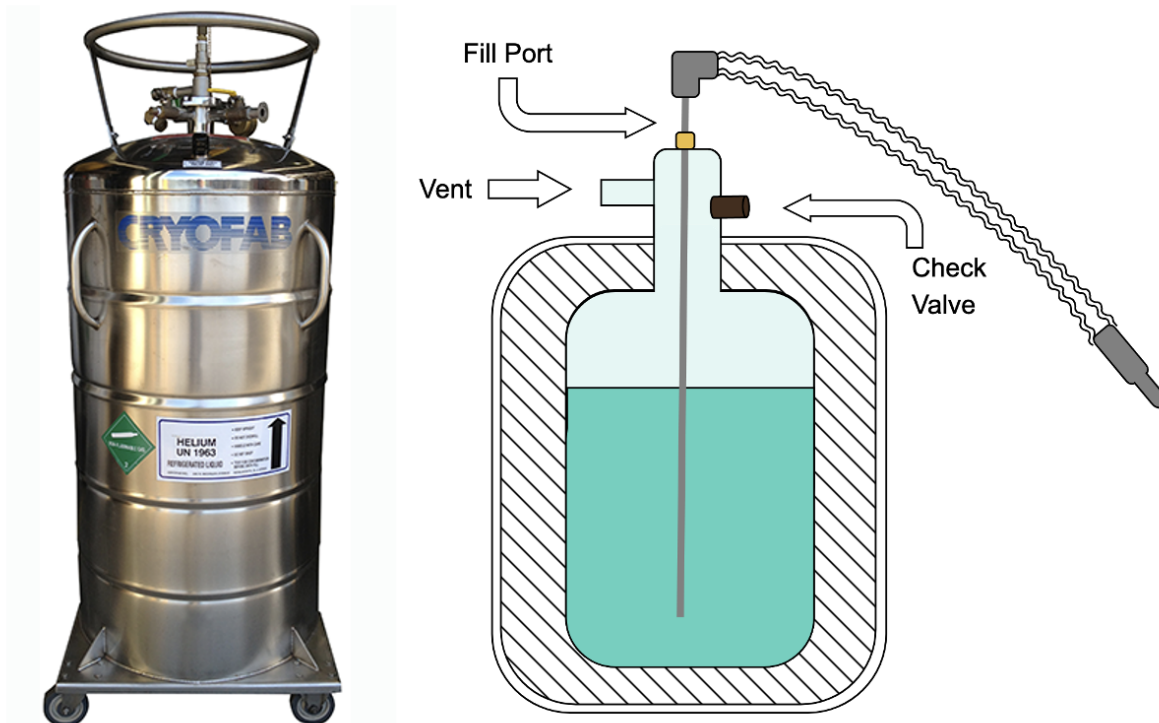


Figure 6.4: A photograph and schematic diagram of a liquid helium transfer dewar. Helium is transferred in and out of the dewar using an insulated hose which is inserted into a fill port at the top of the dewar. A vent port enables the user to control the pressure of the helium atmosphere which exists above the liquid. A safety valve (or more typically, several of them) release excessive pressure if needed. The crosshatched region is typically vacuum space and multilayered radiation shielding.

6.3.1.1 Submersion

The most straightforward method for achieving a cryogenic environment is to simply immerse the evacuated cryostat in an insulated dewar filled with a liquid cryogen such as liquid helium ($T_{LHe} \sim 4$ K) or liquid nitrogen ($T_{LN} \sim 77$ K). The gas produced by the boil-off of the liquid cryogen is vented through a one-way check valve which prevents the back-flow of air into the bath space. When the cryogen is particularly valuable or in short supply⁶, it may be advantageous to capture the boil-off to be re-liquefied.

The dewar itself should be extremely well-insulated to limit the heat flow from the environment. Typically the bath space of the dewar is surrounded a vacuum jacket to eliminate conductive heat flow, and also features super-insulation in the form of Mylar sheets or some other reflective element which substantially reduces radiative heat transfer. An example of a transfer dewar used to transport liquid helium is shown in Fig. 6.4; the dewar which houses a cryostat works on essentially the same principle - however, for further insulation the outer vacuum jacket may itself be cooled with a secondary cryogen.

6.3.1.2 1K Pot

If temperatures below the ambient boiling point of liquid helium are required, then more active cooling measures must be employed. The most common is the use of a 1K pot, which consists of a still within the cryostat vacuum can which draws LHe through a needle valve from the bath space and whose other side is connected to a vacuum pump as shown in Fig. 6.6. When the pump is turned on and the valve is opened, helium is drawn up into the still where it is subject

⁶At the time of writing, liquid helium is the most obvious example of this.

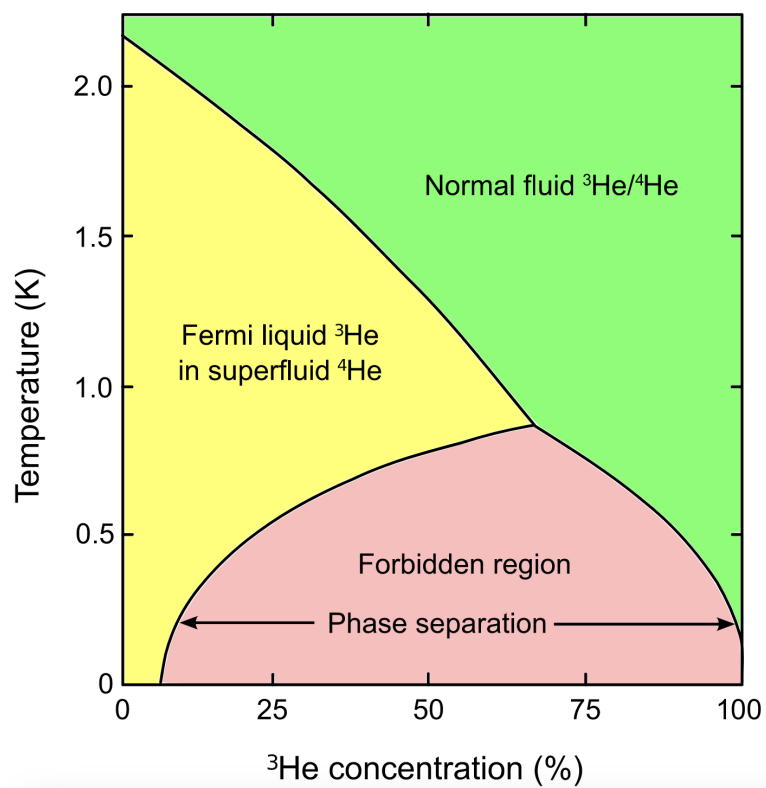


Figure 6.5: A phase diagram for ^3He - ^4He mixtures. Published by Mets501 as Helium phase diagram and licensed under CC BY-SA 3.0.

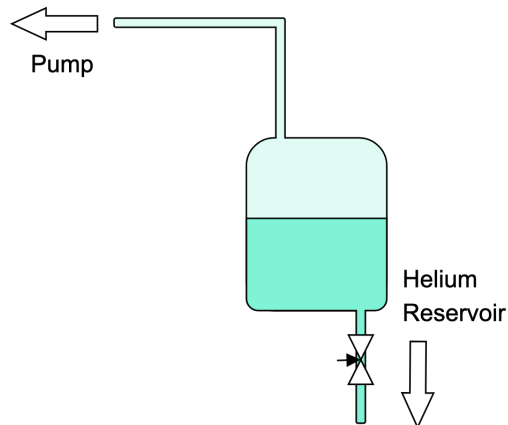


Figure 6.6: A 1K pot, capable of reaching temperatures of ~ 1.5 K. Liquid helium is drawn into the pot by a pump, where it evaporates and provides cooling power. The helium level is controlled by a needle valve, allowing the user to ensure that the helium level is neither above nor below the pot itself.

to evaporative cooling, dropping its temperature to approximately 1.5 K. The still is typically mounted to a plate in the cryostat. The function of the needle valve is to adjust the helium level in the still. For maximum cooling power, the helium level should sit within the body of the still; if the valve is too open or closed, the helium level will sit above or below the still respectively, and the desired base temperature will not be achieved.

6.3.1.3 Dilution Refrigerators

If temperatures below 1K are desired, more sophisticated cooling methods must be employed. The *dilution refrigerator* [2] draws its cooling power from the fact that below approximately 870 mK, liquid helium separates into a ^3He rich phase (consisting of essentially pure ^3He) and a ^3He poor phase (consisting of 6.6% ^3He and 93.4% ^4He) as shown in 6.5. Pumping ^3He across this phase boundary has an analogous cooling effect to evaporative cooling, which can be understood as pumping a liquid across a liquid/gas phase boundary.

A typical design for a dilution refrigerator is shown in Fig. 6.7. The circulation line is filled

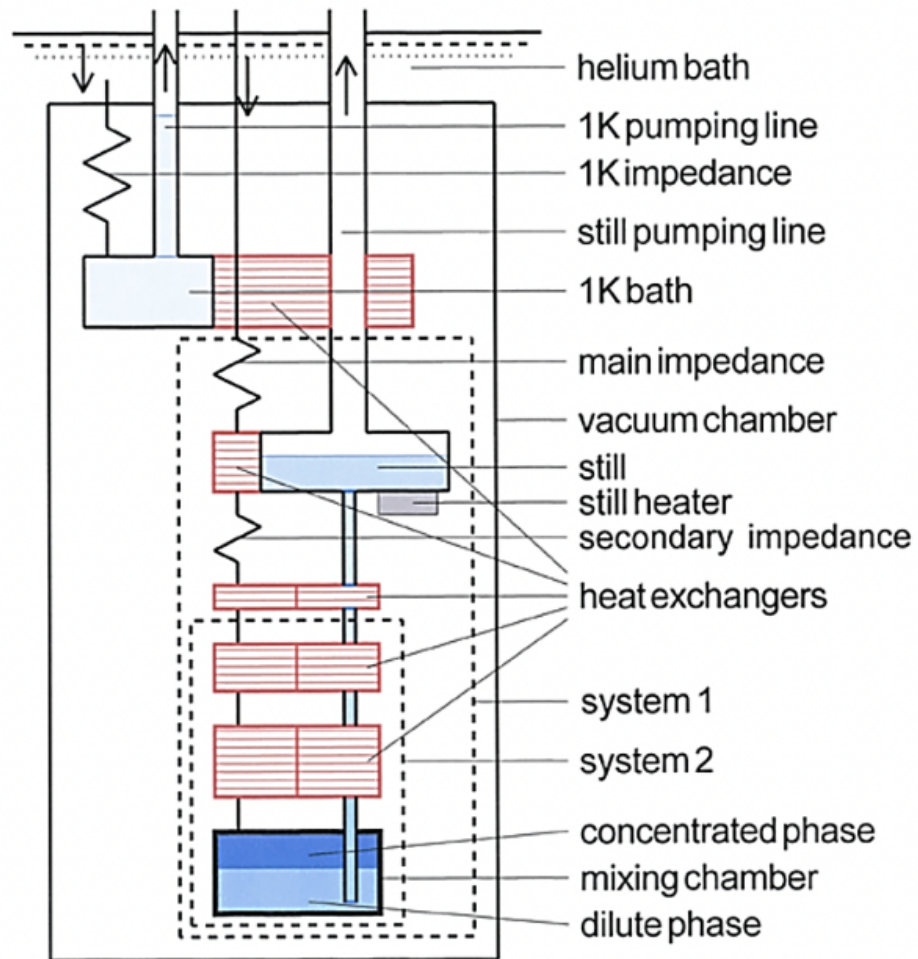


Figure 6.7: A schematic diagram of a dilution refrigerator. Reused with permission from [2].

with a mixture of ^3He and ^4He , which is liquified as it is cooled by the heat exchangers which are in contact with the 1K pot. ^3He has a much higher vapor pressure than ^4He at ~ 1 K, and so when the mixture reaches the still it preferentially evaporates while the ^4He tends to remain liquid. As the ^3He cycles through the system, its evaporation provides additional cooling power until it reaches the critical temperature of 870 mK, at which point the mixture undergoes a phase separation. The mix ratio and circulation rate are chosen such that when this occurs, the phase boundary is in the mixing chamber so that in order to reach the still, ^3He must be pulled across the phase boundary.

This process is endothermic, and is quite analogous to evaporative cooling (just with a different type of phase boundary). As a result, the mixing chamber becomes the coldest part of the dilution unit; our dilution refrigerator is capable of reaching a base temperature of 35 mK.

Chapter 7: Development of Cryogenic Current-Biased STM

Typical STM experiments are performed under equilibrium conditions, in which the characteristics of the surface being studied do not change with time and can be understood through the lens of equilibrium thermodynamics. For example, density of occupied electronic states can generally be well-approximated by multiplying the total density of states¹ by the Fermi-Dirac distribution function

$$f(\epsilon) = [e^{(\epsilon-\mu)/kT} + 1]^{-1} \quad (7.1)$$

where μ is the electron chemical potential, T is the temperature of the sample, and $f(\epsilon)$ is the probability that a single-electron state with energy ϵ is occupied. The $T \rightarrow 0$ approximation in which $f(\epsilon) \rightarrow \theta(\epsilon)$ ² is often sufficient, with notable exceptions such as semiconductors.

However, there are a number of ways in which an STM might be utilized to study non-equilibrium transport phenomena. For example, local sensing of surface phenomena such as the spin Hall effect and the Edelstein (and Rashba-Edelstein) effects - which emerge in the presence of a net current - would be of great interest in the characterization of novel materials. The atomic

¹As computed using the independent electron approximation - see Section 4.1.

² $\theta(x)$ is the unit step function, which we define as

$$f(\epsilon) \rightarrow \theta(\epsilon - \mu) = \begin{cases} 0 & \epsilon - \mu < 0 \\ \frac{1}{2} & \epsilon - \mu = 0 \\ 1 & \epsilon - \mu > 0 \end{cases}$$

precision of an STM makes such techniques sensitive to nanometer-scale variations in transport channels which are averaged out in more conventional, large-scale transport experiments.

7.1 Motivation for Current-biased STM

7.1.1 Scanning Tunneling Potentiometry

In scanning tunneling potentiometry (STP), a lateral current is established in the sample which results in a spatially-varying electric potential. This variation can be extracted from tunneling spectroscopy curves, which in principle allows for the mapping of the local distribution of current over scales of tens or hundreds of nanometers.

7.1.1.1 Mapping the Local Potential

There are a number of ways to extract the local variation in electric potential using spectroscopy, which differ depending on whether or not the sample is gapped. For metallic samples, one might naively imagine taking an IV curve and extracting the value V_0 for which $I = 0$. This value (or rather, $-V_0$) gives the potential of the sample relative to the STM tip, which we take to be (virtually) grounded.

The problem with this is that a realistic STM circuit will generally introduce various current and voltage offsets which must be carefully accounted for before identifying V_0 . A better approach is to take one curve, change the resistance R_T of the tunnel junction, and then take a second curve. Due to the different values of R_T , the curves will intersect at a point which we may identify as (V_0, I_0) , where I_0 is the current offset present in the system.

An even better approach, shown in Fig. 7.1a,c is to take several curves at different values of

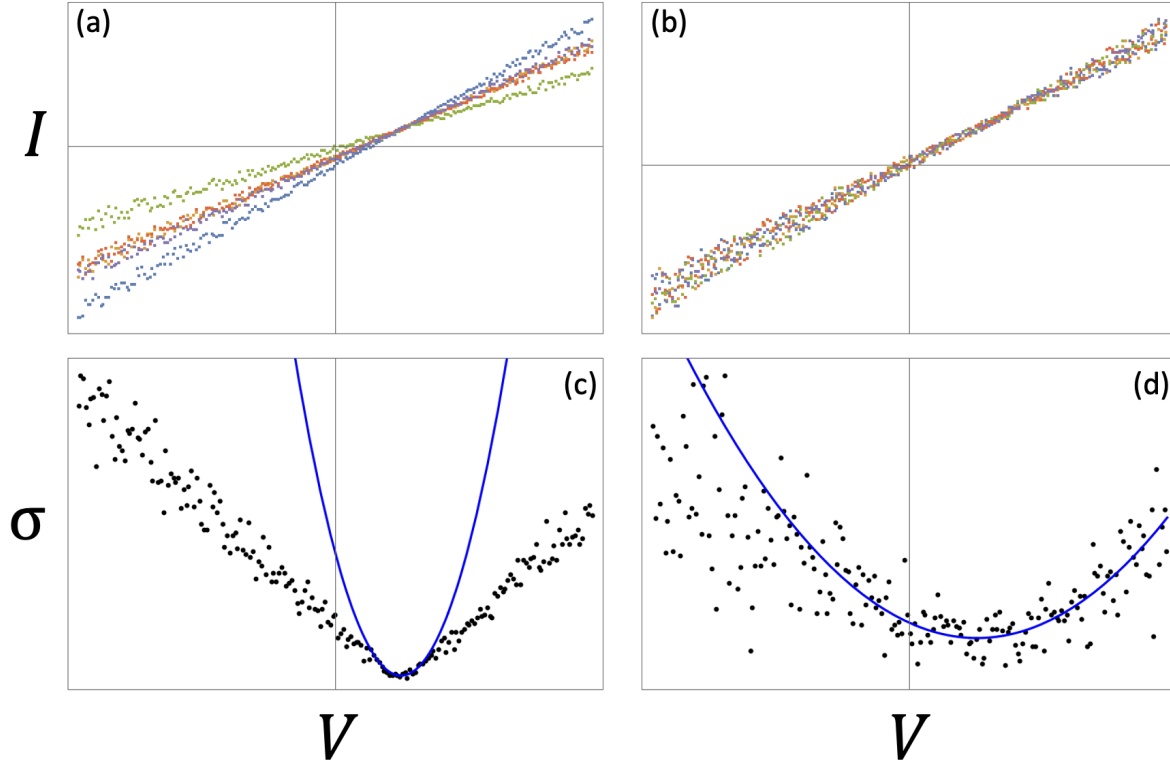


Figure 7.1: Simulated IV curves and their corresponding standard deviations. (a) simulates curves taken at five different values of the tunneling resistance R_T , resulting in a sharper peak in (c). (b) simulates five curves taken at the same value of R_T , and differ due to thermal noise which is minimized at the true zero of the tip-sample voltage. As shown in (d), this method results in a less precise measurement of the current and voltage offsets.

R_T . For each value of V , we compute the standard deviation of the set of curves. The subsequent plot exhibits a well-defined minimum at the point (V_0, I_0) which may be extracted by fitting with a suitable function (e.g. a quadratic).

Alternatively, one can also simply take multiple curves at the *same* value of R_T as shown in 7.1b,d. The so-called *shot noise*³ makes a random contribution to the tunneling given by

$$I_{SN} = \frac{e^2}{\pi\hbar} V \sum_n T_n \quad (7.2)$$

where T_n is the transmission coefficient for the n^{th} tunneling channel. Because this noise is proportional to the voltage across the tunnel junction, it is minimized at the true zero of the tip-sample voltage. This approach is useful if maintaining the tip-sample distance is expected to be important, but it comes at the cost of measurement precision because the standard deviations of curves which vary only due to shot noise will be far less sharply peaked.

If the sample is gapped, then extracting the true zero of the tip-sample bias can be much more difficult. For sufficiently small scales or flat samples, the STM can be operated in constant current/constant height mode, in which the z -position of the tip is held fixed while the feedback modulates the bias voltage in order to maintain a set tunneling current. This technique works well when scanning over small features such as semiconductor junctions [39], but is problematic over large scales or for irregular samples where the lack of z control could lead to a tip crash.

³i.e. Poissonian fluctuations in the current due to the discrete nature of the charge carriers

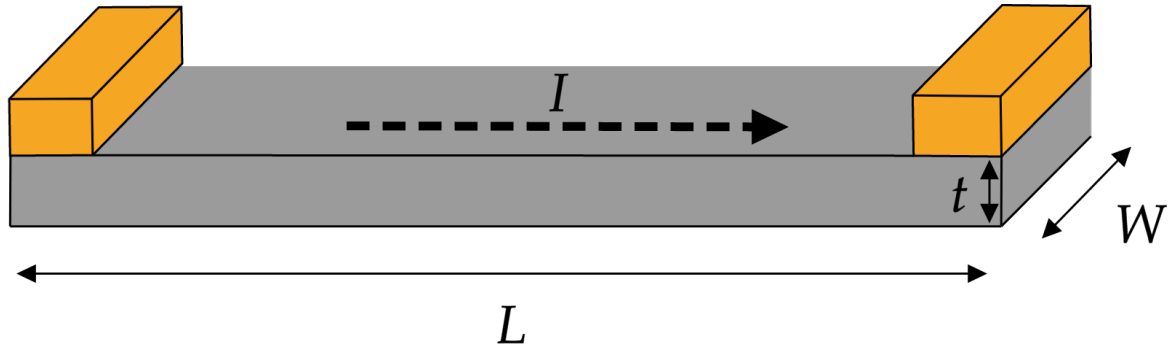


Figure 7.2: A schematic drawing of a sample mounted for scanning tunneling potentiometry. Current is injected and extracted through the gold contacts.

7.1.1.2 Sample Considerations

Typical STP measurements are performed on thin films for a number of reasons. Consider a sample with length L , width W , and thickness t outfitted with electrical contacts through which the lateral current I is injected and extracted, as shown in Fig. 7.2. If the resistivity of the sample is given by ρ , the average rate of change of the electric potential across a uniform sample would be given by the electric field

$$\mathcal{E} = \frac{\Delta\varphi}{\Delta x} = \frac{I\rho}{Wt} = J\rho \quad (7.3)$$

where $J \equiv I/Wt$ is the current density and ρ is the sample resistivity. On the other hand, any spectroscopic technique for determining the local potential on the surface has an associated finite resolution δV . In order to resolve potential variations over a length scale ℓ , we require that

$$\mathcal{E} \gg \frac{\delta V}{\ell} \quad (7.4)$$

Given a fixed sample material and lateral film geometry, $\mathcal{E}$ may be increased either by increasing I or decreasing t . The former option typically leads to unwanted sample heating, thermal expansion, and instability of the tip-sample junction. Note that the power generated is given by

$$P = I^2 R = \mathcal{E}^2 \frac{LWt}{\rho} = \mathcal{E} I L \quad (7.5)$$

Therefore, the power P scales quadratically with $\mathcal{E}$ for fixed t , but only linearly with $\mathcal{E}$ for fixed I . Furthermore, if potential variations are to be correlated with structures (e.g. steps or defects) which are visible in the surface topography, then the structures in question should have a substantial impact on current flow. If the sample is thick, then one would expect surface features to have a negligible impact on current flow, which would occur primarily in the bulk of the film. For both reasons, a thin film is typically - but not always - preferred.

7.1.2 Local Sensing of Spin Transport Phenomena

In addition to potentiometry, the ability to pass a lateral current through a sample during STM experiments allows for the direct study of non-equilibrium transport effects. The localized measurement of an enhanced signal in spin-polarized STM due to a spin-saturated surface has been reported in spin Hall materials [36] and topological insulators [40]. This opens the door to the characterization of nanoscale variations in the spin Hall effect and the local observation of the spin-momentum locked topological surface states. The possibility of using a current-biased STM to see how these effects are manifested in the presence of nanoscale defects and topographical features is a strong motivator for the continued development of these techniques.

7.2 Description of the LPS 4K STM

The LPS 4K STM system (shown in Fig. 7.3) is described in detail in [3]. It consists of a room-temperature ultra-high vacuum (UHV) subsystem and a cryostat which houses the microscope. The UHV subsystem sits on top of an optical table which rests on N₂ gas springs for vibration isolation. The cryostat is connected via a gate valve to the bottom of the transfer chamber, passing through a hole in the optical table and hanging approximately 2 meters below it in a maintenance pit in the floor. A liquid helium (LHe) magnet dewar sits at the bottom of the pit and is raised by a manual winch to enclose the cryostat during normal operation.

7.2.1 The UHV Subsystem

The UHV subsystem is comprised of two chambers separated by a gate valve and mounted on top of an optical table - one for loading and preparing samples and a second which acts as a transfer chamber for moving samples to the cryostat. The preparation and transfer chambers are respectively outfitted with Varian Starcell 300 and 75 ion getter pumps, as well as Vacuum Generators ST22 titanium sublimation pumps. Together, they maintain vacuum levels of approximately 4×10^{-10} mbar and 8×10^{-11} mbar in the prep and transfer chambers, respectively⁴.

Samples are mounted on a sample stud shown in Fig. 7.4 and are loaded into a load-lock which is evacuated by a Varian Mini-Task turbomolecular pump. After approximately 45 minutes, a gate valve is opened and the sample is moved into the preparation chamber. Once inside the prep chamber, samples can be heated to approximately 450° C by a resistive filament,

⁴The higher pressure in the prep chamber is due to frequent exposure to H₂O via the load lock, which is opened to transfer samples in and out as well as during Ar ion sputtering

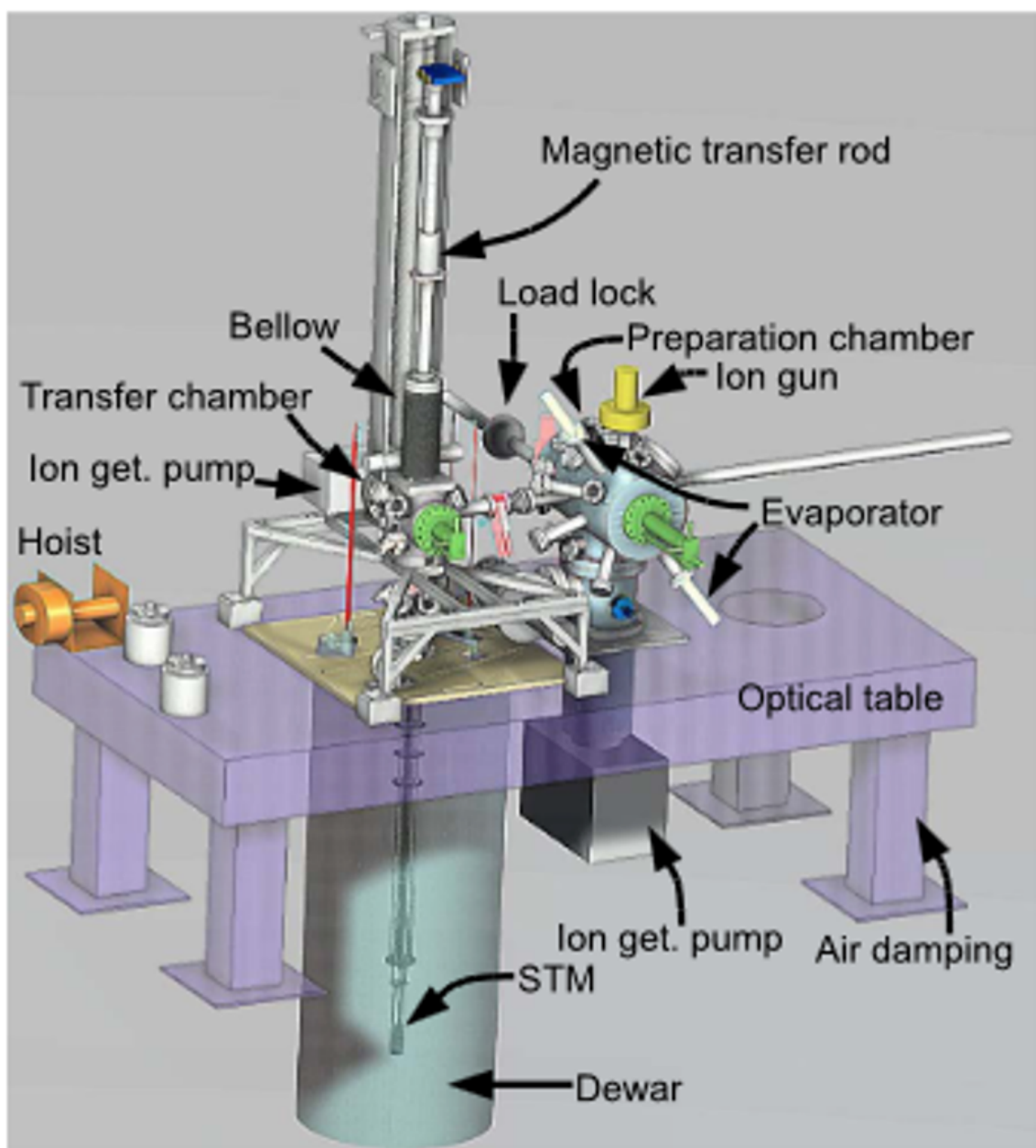


Figure 7.3: A schematic drawing of the 4K STM system, reused with permission from [3]. Samples are loaded into the preparation chamber via a load lock, then into the transfer chamber via a horizontal magnetic transfer rod. A vertical transfer rod picks up the sample and transfers it into the cryostat which is housed within the LHe dewar.

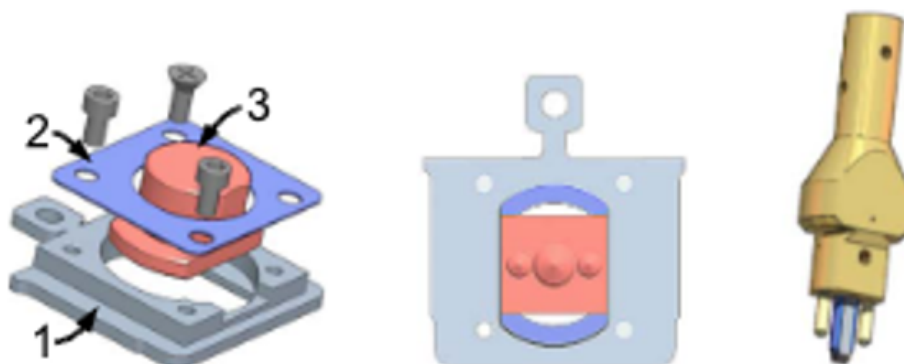


Figure 7.4: A schematic drawing of the sample holder and transfer rod, reused with permission from [3]. The sample holder (1) is outfitted with a spring plate (2) which holds the sample stud (3) in place. Once inserted into the holder through the keyhole in the back, it is rotated to lock it in place. The sample holder is manipulated in the UHV subsystem via the tab at the top. The sample stud is extracted from the sample holder and moved into the cryostat via a transfer rod with a leaf spring friction coupling (right).

or to 750° C by an electron beam. Resistive samples (e.g. silicon) may be heated via a direct current heater which interfaces with a customized sample stud built for that purpose.

In addition to the heating capabilities, the prep chamber is equipped with an Omicron ISE 10/NGISE 10 sputter gun for removing surface contaminants and native oxides. It also features an Extorr XT200 residual gas analyzer which provides insight into the ambient vacuum quality, validation of the gas purity during sputtering, and monitoring the composition of the materials released from samples during thermal outgassing. Finally, the chamber is equipped with an AgI electrochemical cell for iodine deposition, which has been used to study iodine passivation of Ag(111) [41].

Once samples are ready for scanning, they are moved into the transfer chamber and picked up by a transfer rod via a friction coupling which fits into the bottom of the sample stud. The stud is then lowered sample side down into the cryostat and locked into the microscope as shown

in Fig. 7.6.

7.2.2 The Cryostat and Magnet Dewar

The cryostat consists of the two-tiered copper probe shown in 7.5 and is connected via stainless steel conduit tubes to the transfer chamber and to breakout flanges for electrical connections at the top of the table. The higher tier of the probe features a keyhole which centers the transfer rod to ensure its alignment with the microscope, while the lower probe is the stage to which the microscope body is mounted. Low-temperature RuO_x thermometers calibrated at room temperature, 77 K, and 4.2 K are mounted to the lower tier of the cryostat and to the bottom of the microscope. Thin leaves of stainless steel sheet, which are pushed out of the way when the transfer rod is lowered into the cryostat, act as baffles which impede the direct flow of thermal radiation from the UHV subsystem.

The dewar, manufactured by American Magnetics, has a capacity of 80 L and is insulated by an outer vacuum jacket as well as a layer of superinsulation to impede thermal radiation. The hold time of the dewar is approximately seven days when filled with liquid helium, and on the order of six weeks when filled with liquid nitrogen.

A two-axis superconducting magnet sits at the bottom of the dewar and is capable of producing a $9 \text{ T } \hat{z}$ field, a $2 \text{ T } \hat{y}$ field, or a 2D vector field of up to 1 T. When the dewar is raised, the cryostat sits in the center of the magnet coils. The magnet coils are powered independently by an AMI Programmer 420 4 Quadrant Power Supply (Model 4Q05100PS) via leads which may be disconnected when the magnet is not in use or during persistent operation mode, which is activated via a persistent switch. The leads are typically disconnected as often as possible, as the

dewar hold time is reduced by several days when they are left connected.

7.2.3 The Microscope

The STM itself is fundamentally based on the design by Shuheng Pan which was modified first by the Davis group at the University of California [42] and subsequently by us to allow the STM to be mounted to the microscope stage from below. It consists of a cylindrical ceramic body which encloses a mobile sapphire prism on which the tip is mounted. The prism is held in place by a ruby ball supported by a metallic spring plate, which can be loosened or tightened as necessary. Coarse motion of the prism is provided by six piezoelectric stacks which operate simultaneously in the stick-slip configuration. During the "slow phase," the prism is raised or lowered as the piezo stacks deform; during the "fast phase," the piezo stacks reset to their original shapes while the prism is held in place by its inertia.

Within the prism, the STM tip is mounted to the end of a 5-lead cylindrical piezo tube whose deformation provides the fine lateral and vertical scanning motion. The tube is 12 mm long, allowing for a maximum scan area of approximately $2.2 \times 2.2 \mu\text{m}^2$ and a vertical scan range of approximately 400 nm. The tip holder is also surrounded by an electrical ground plate to screen out electromagnetic noise.

Samples are locked in place face-down in a keyhole at the top of the microscope as shown in Fig. 7.6. The sample stud is stabilized by a metallic spring plate which also serves to apply the desired voltage bias to the sample.

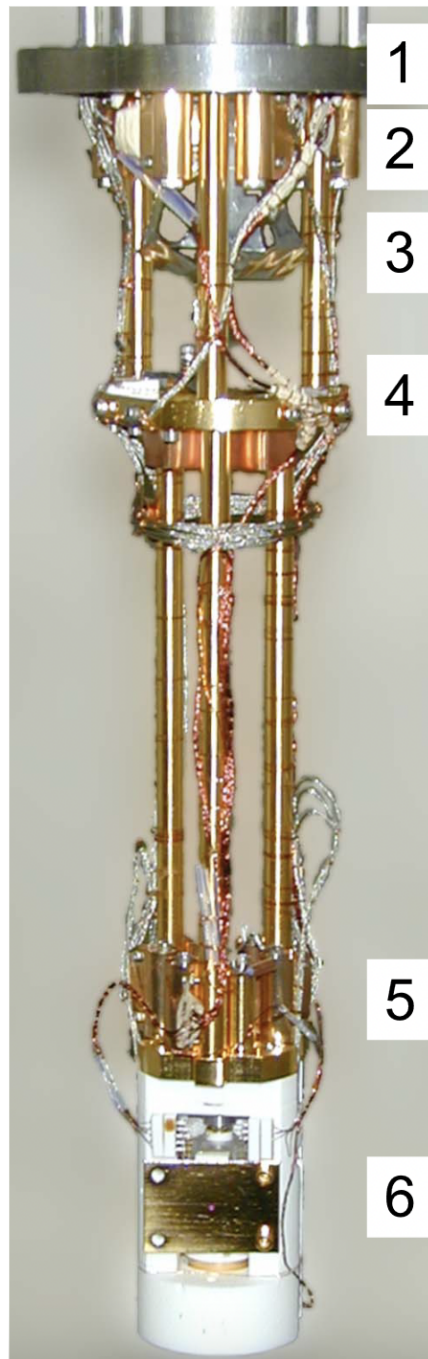


Figure 7.5: The 4K cryostat probe. The vacuum can which houses the cryostat is screwed in to the top flange (1). Copper heat sinks (2) keep the microscope wiring cool, while thin metallic shutters (3) block room-temperature electromagnetic radiation. A keyhole (4) ensures that the transfer rod remains aligned so it can reach the microscope stage (5) and microscope (6). Figure reused with permission from [3].

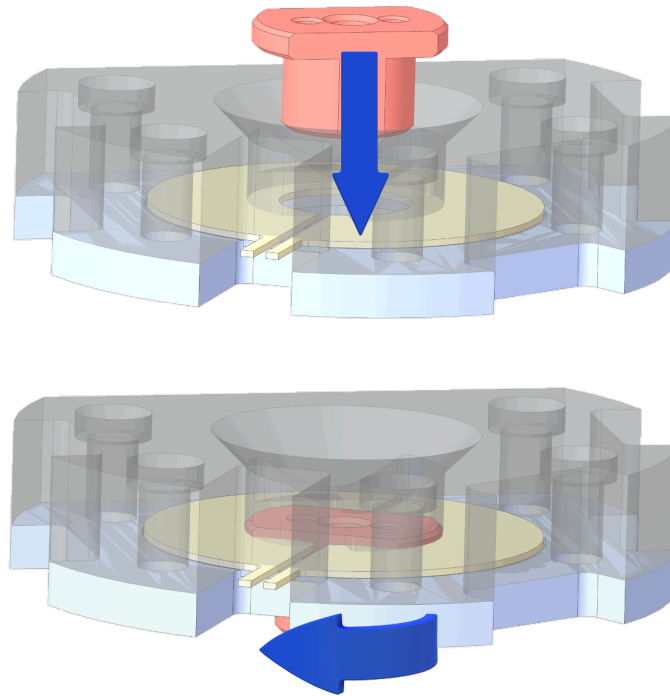


Figure 7.6: The locking mechanism at the top of the 4K microscope. The sample stud is inserted sample-side down into a keyhole and then twisted to lock it in place. The annular bias plate (shown in gold) provides electrical contact to the sample as well as mechanical stability. In this figure, the bias plate has been split in half to allow for current-biasing of the sample.

7.2.4 The Electronics and STM Control

Electronic access to the walker piezos, the scan piezos, the bias plate, the tip, and the ground plane are provided by plugs epoxied into the outer body of the STM. The plugs are connected to the breakout flanges at the top of the table by wires which are wound around the body of the cryostat, which acts as a heat sink. The walker piezos and the thermometers are connected via twisted pairs of manganin wire, while the more sensitive lines (scan piezos, bias, and current) are connected via shielded coax.

On top of the table the current line runs through a 0.5 m BNC cable to a DL Instruments 1211 current preamplifier set to a default sensitivity of 10^{-12} A/V. From there, it runs along with the remaining lines to a Nanonis Mimea control unit.

7.3 Modification Design and Technical Specifications

The modification to allow for lateral current-biasing of the sample consists of three parts. First, the spring plate which applies the bias to the sample stud is split into two pieces which are independently wired to the top of the table. Second, the sample stud is reconfigured to a metal-ceramic composite design which allows a differential voltage to be applied across the wings. Finally, a voltage compensation circuit is inserted into the bias line which ensures that the local tip-sample voltage bias remains fixed when varying the lateral current.

7.3.1 The Sample Stud

The redesign of the sample stud proved to be fairly non-trivial. In the original design, the stud consists of a stainless steel body in which a trench has been cut, as shown in [7.7](#). Two

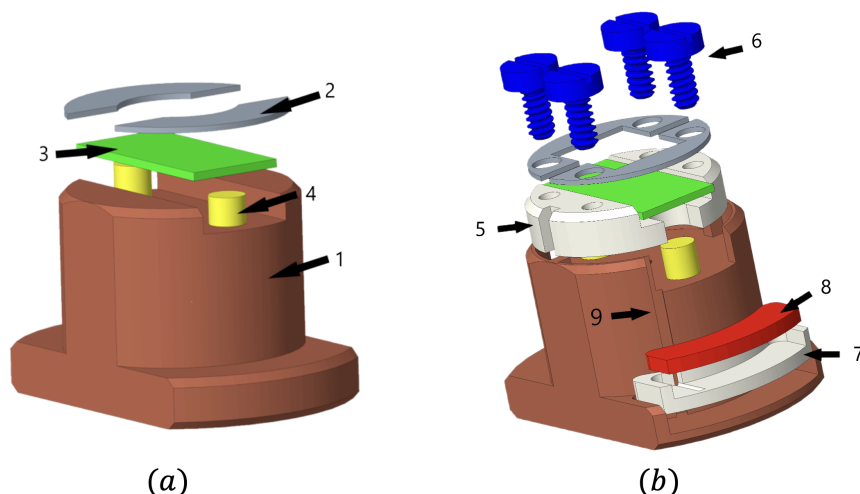


Figure 7.7: The original (a) and redesigned (b) sample studs. (1) The body of the stud is made from SS, to which metallic straps (2) are spot welded. A sample (3) is loaded into the trench before being pressed upward into the straps by set screws (4). In the redesigned stud with copper body, a ceramic insert (5) insulates the sample from the stud body. The straps are fixed to the insert by M1 screws (6). A second ceramic insert (7) isolates a contact on one of the wings (8), which is connected to the adjacent strap via an enameled wire which sits in a groove (9). The other strap is connected to the sample body.

sheet metal straps are spot-welded across the trench to provide electrical contact to the surface of sample, generically assumed to be a film grown on an insulating substrate. Such samples are cut to size, loaded into the trench, and then raised by set screws until they are pressed into the straps; optionally, small pieces of gold foil are inserted between the sample and the straps to improve the electrical contact.

Electrical isolation of the stud wings necessitates the incorporation of insulating materials into one of the two wings and, in some way, into the strap design. This also requires the introduction of adhesives at the metal-ceramic interfaces. Both the thermal stability of the adhesives and differential expansion of the metal and ceramic were important factors in material selection; the final design utilized EPO-TEK H72 (insulating) and H20E (conductive) epoxies along with Macor⁵ ceramic.

⁵The ceramic was chosen both due to its thermal expansion coefficient and the ease with which it can be ma-

Some possibilities which were explored during the design phase included the introduction of ceramic spacers between the stud body and the straps, or between the straps and the sample; in the latter case, the spacers would be covered with metallic foil which would be wired to the insulated wing and to the body of the stud, respectively. Both possibilities were ultimately deemed ineffective due to the difficulty of their implementation and to their relative fragility.

In the final design, a thin cylindrical region is machined out of the top of the sample stud and replaced with a ceramic disk, which is then epoxied in place via non-conductive epoxy. As before, the disk includes a trench into which samples are loaded as well as two threaded holes for the set screws which raise the sample into the straps. The straps themselves are held in place by four M1 "cheese-head" screws. One of the stud wings is insulated via a ceramic insert fixed in place by non-conductive epoxy, and the end of a thin enameled wire is sandwiched between the insert and a copper plate and fixed in place with conductive epoxy. Finally, this wire is wrapped around the nearest screw and mechanically clamped in place, while a corresponding wire connects the opposite strap to the body of the stud.

The use of the ceramic disk greatly reduces the precision required (and therefore the difficulty) in the assembly of the stud. Once the disk is in place it cannot be removed in one piece, but it can be replaced by breaking it into pieces and filing out the residual epoxy. This design has been found to be mechanically robust, with no moving parts save the set screws which raise the sample into the screws. Additionally, the absence of spot-welding in this design means that the body of the stud can be made of copper rather than stainless steel; this is valuable both for heat conduction and for measurements in magnetic environments, since strong magnetic fields would magnetize a stainless steel stud and therefore potentially obfuscate experimental data by

chined.

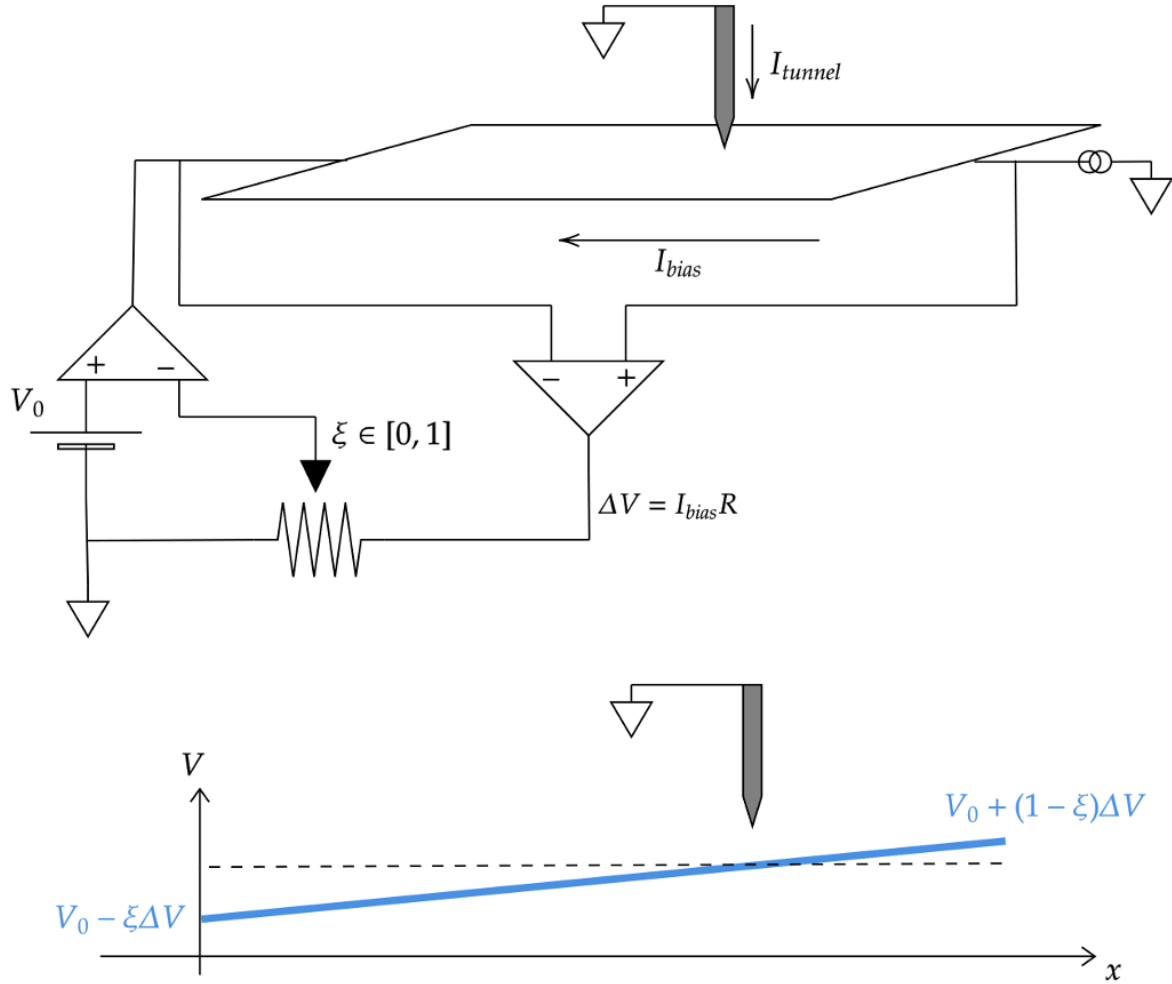


Figure 7.8: The voltage compensation circuit which allows for a dynamic modulation of the bias current. A fraction ξ of the potential difference across the film is subtracted from the nominal bias voltage. When ξ is properly calibrated to the tip location, the local tip-sample voltage remains at V_0 while the bias current and ΔV are allowed to vary.

effectively introducing hysteresis.

7.3.2 The Compensation Circuit

The passage of a lateral current through a resistive sample causes the electric potential to be non-uniform across the sample's surface. As a result, in the absence of a compensation mechanism the local tip-sample voltage generally depends on the lateral current being applied.

To adjust for this effect, we introduced the compensation circuit shown in 7.8. A differential amplifier computes the potential difference $\Delta V = I_{bias}R$ across the sample, and some fraction $\xi \in [0, 1]$ of the result - which is selected by a 10-turn potentiometer - is then subtracted from the nominal voltage bias being applied by the STM control software. Assuming a homogeneous film as a toy model, the tip-sample voltage at a position $x \in [0, L]$ on the film is given by

$$V(x) = V_0 + I_{bias}R \left(\frac{x}{L} - \xi \right) \quad (7.6)$$

As a result, choosing $\xi = x/L$ eliminates the influence of I_{bias} on the tip-sample voltage, allowing us to dynamically modulate I_{bias} without introducing any unwanted voltage offsets. The correct setting for the potentiometer is straightforwardly obtained by passing a small amount of current (e.g. $10 \mu\text{A}$) through the sample and extracting the zero point of the tip-sample bias via one of the mechanisms described in Sec. 7.1.1.1. Once the setting has been adjusted, the bias current can be increased and the calibration can be repeated with higher sensitivity.

7.4 Experimental Characterization

7.4.1 Thermal Characteristics

The first test of the new design involved varying the Ohmic power delivered to a thin film and monitoring the system temperature (as measured by the thermometers connected to the microscope stage of the cryostat and the bottom of the microscope body as described in Sec. 7.2.2) as well as the dynamic lateral and vertical motion of the STM tip.

In this experiment, the tip was made by electrochemically etching 0.2 mm tungsten wire,

which was subsequently cleaned by high-voltage field emission on Au(111). The sample consisted of a Bi_2Te_3 film of thickness 200 nm. An aluminum post⁶ was glued to the surface of the film with conductive epoxy before it was transferred to the load-lock. Once under high vacuum, it was moved into the prep chamber. During the transfer, the post was knocked off which mechanically cleaved the sample and exposed a clean surface which was immediately transferred to the microscope at 4.2 K.

After locating a suitable scan area, the voltage compensation circuit was calibrated as described Sec. 7.3.2. From there, the tip was brought into tunneling range and a bias current $I_{bias} = 3$ mA (corresponding to 40 mW of Ohmic power) was injected into the sample. The tip began to retract due to the combined expansion of the film, substrate, and sample holder, but the expansion was sufficiently slow that the feedback could follow. After approximately 6 minutes, the tip position came to equilibrium ~ 150 nm above its starting point, as shown in 7.9a.

As shown in 7.9b, the thermometers registered temperature increases of 0.25 K and 1.7 K at the microscope stage and the bottom of the microscope, respectively. Of course, neither measurement accurately reflects the temperature change of the film, with the former thermometer being attached to a copper heat sink and the latter to a large thermal insulator approximately 4 cm below the sample location. The long equilibration time of the thermometer readings (with time constants on the order of 3 hours and 1 hour, respectively) can be attributed to the large thermal masses of the components to which they are attached.

⁶Our post consisted of a piece of aluminum welding rod.

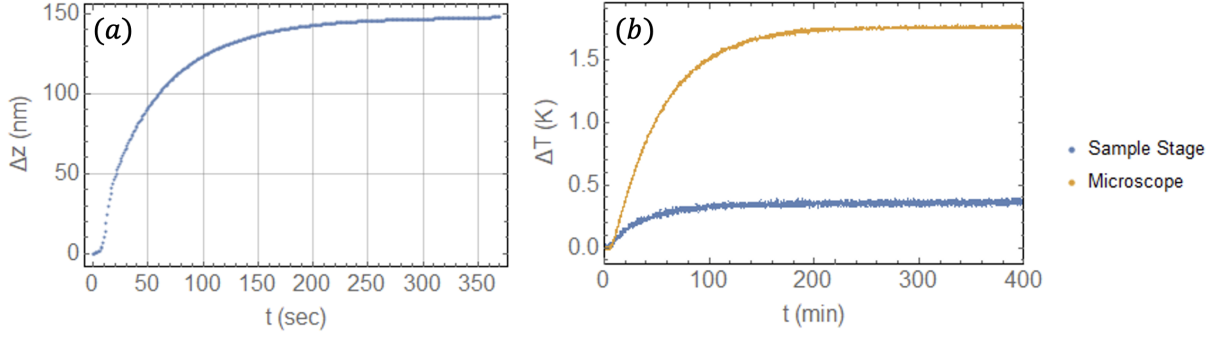


Figure 7.9: Equilibration times for the z position of the tip (a) and the thermometers on the microscope body and sample stage (b). The change in vertical tip position is well within the dynamic range of the z piezo, and can easily be accommodated by the feedback controls. The far longer equilibration time of the thermometers is due to the large thermal masses of the microscope body and sample stage.

7.4.2 Validation via Potentiometry

A second test of the design was to perform an elementary potentiometry experiment on a resistive film. A film of NbTiN with thickness 7 – 10 nm was grown on an MgO substrate using an AJA International A300 series magnetron sputterer. The sputtering target was 99.95% pure $\text{Nb}_x\text{Ti}_{1-x}$, $x = 0.75$. The sample was mounted on the sample stud and gold foil was inserted beneath the sample contacts to minimize contact resistance; at this time, the sample resistance was 3.65 k Ω .

Once loaded into the prep chamber, the film was subjected to heating ($T \approx 300$ C) and Ar^+ sputtering for approximately 20 minutes before being transferred to the microscope at 77 K. As before, this experiment utilized a tungsten tip which was electrochemically etched from 0.2 mm tungsten wire.

Once the sample reached thermal equilibrium, the sample resistance increased to 4.126 k Ω . After the voltage compensation circuit was calibrated as described in Sec. 7.3.2, a bias current

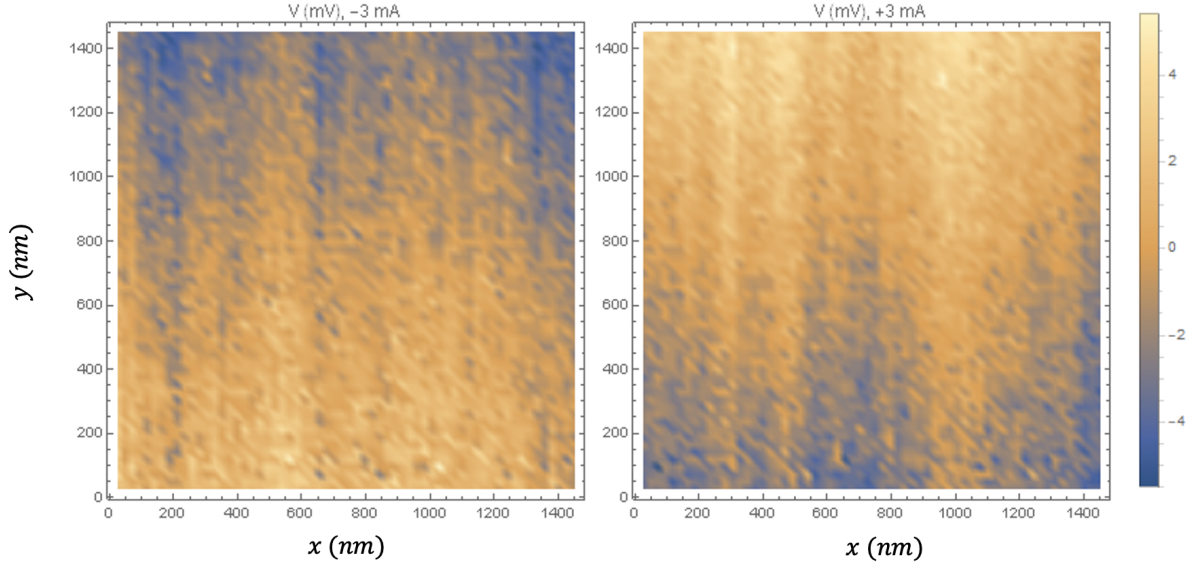


Figure 7.10: Scanning tunneling potentiometry of a NbTiN thin film, with $I_{bias} = -3$ mA (left) and $+3$ mA (right). The average potential difference across the scan window agrees with the expected value (assuming a homogeneous film) to within 5%.

of 3 mA was established and a spectroscopy map with 256×256 resolution was taken over a $1.5 \times 1.5 \mu\text{m}^2$ scan area. Afterward, the measurement was repeated with a bias current of -3 mA. At each pixel of the map, the zero of the tip-sample bias was extracted from the IV curve as described in Sec. 7.1.1.1, and the results are plotted in 7.10. The corner-to-corner potential difference across the scan area is approximately 7.5 mV, in good agreement with the expected value

$$\Delta V = I_{bias} R \frac{\delta x}{L} = (3 \text{ mA})(4.126 \text{ k}\Omega) \frac{1.5\sqrt{2} \mu\text{m}}{3.81 \text{ mm}} \approx 7.12 \text{ mV}$$

The reversal of the voltage gradient when the bias current switched direction provides additional support for the fact that the system works as anticipated.

7.5 Future Work

The modification of the cryogenic STM to support current biasing opens the door to a large number of interesting experiments. This section will consist of a number of suggestions which could be pursued in the future.

7.5.1 Current Mapping and Potentiometry

7.5.1.1 Spin Accumulation

The initial motivation for the modification was the desire to extend the spin-polarized STM measurements described in Sec. 7.1.2 to cryogenic conditions. As a preliminary step, the measurements in [36] and [40] should be simply duplicated with the added stability and higher-quality vacuum associated with low temperatures. This would validate the previous work while providing additional confirmation that the modification to the low-temperature STM behaves as anticipated with dynamically modulated bias currents as well as steady ones.

The thermal characterization of the current-biasing modification suggests that extended measurements utilizing a persistent lateral bias current can be achieved without compromising the stability of the tip-sample junction. This would permit the extraction of a differential spectroscopy map which, when paired with a spin-polarized tip, should exhibit an asymmetry in regions of significant spin saturation which scales with the magnitude of the applied bias current.

An example experimental outline would go as follows. Once the spin-polarization of the tip has been established with a magnetically-ordered reference sample, a spin Hall material or topological insulator is transferred to the system. After a suitably clean area is identified in

topography and the voltage compensation circuit has been calibrated, a persistent lateral bias current is established in the sample, which is allowed to come to thermal equilibrium over a period of several minutes. The topography is verified again to ensure no major changes to the tip or sample occurred during heating.

Once the sample is in thermal equilibrium, a differential spectroscopy map is extracted. From there, the current is reversed (thereby reversing the spin saturation of the surface), and a second spectroscopy map is taken. The difference in the two maps constitutes a 3D data set with coordinates (x, y, V) . Slices of constant V should exhibit strong, spatially-coherent signals corresponding to areas of high spin accumulation when V lies in a spin-polarized tip band.

For a spin Hall material, mapping this spin accumulation would provide insight into any tendency of spins to accumulate near topographical features or defects. For a topological insulator, it should provide a map of the distribution of the spin-saturated surface current guaranteed by its bulk topology. In both cases, the accumulation of spin should scale with the magnitude of the lateral current, providing an extra sanity check as well as a quantitative measure of how the distribution of spins varies when the amplitude of the bias current is modulated.

7.5.1.2 Granular Superconductors

The superconductivity of granular films is a fascinating subject which still resists a full and satisfactory explanation despite being studied for many decades. An experiment in which I studied the spatial variation in the superconducting gap of superconducting granular aluminum is the subject of Chapter 8.

The intent behind this experiment was to provide insight into how well-coupled adjacent

grains tend to be. An alternative experiment using a lateral bias current might go as follows. After the sample is prepared via Ar^+ sputtering and transferred into the STM, a bias current is established and a potentiometry map is extracted using the techniques of Sec. 7.1.1. Conventional wisdom holds that the sheet resistance of the film is due to the presence of oxide layers at the grain boundaries, suggesting that in the presence of a lateral current the potential should vary sharply in these regions while being relatively constant across the individual grains.

A potentiometry experiment would validate (or invalidate) this assumption. Sufficiently precise data would also provide a quantitative characterization of these oxide layers (e.g. their thicknesses) as well as their uniformity over large areas. The percolation model discussed in Sec. 8.1 would suggest that there should be networks of well-coupled grains (with thin oxide barriers between them) which run throughout the film, with the majority of the grains being well-insulated from one another. Such a feature would show up very clearly in a potentiometry map.

It should also be noted that while low temperatures would provide stability and higher quality images, it would *not* be necessary for the sample to be superconducting in order for the coupling between grains to be characterized via potentiometry.

7.5.2 Manipulation of Solitons

A second application of my modification lies in the manipulation of solitons which couple to externally-applied currents. Examples of this include magnetic vortices in type-II superconductors and magnetic structures such as skyrmions, the latter of which are the subject of the study presented in Chapter 9.

7.5.2.1 Magnetic Vortices

An Abrikosov vortex (see Sec. 3.3) features a core region (whose radius is on the order of the coherence length ξ) in which the superconductivity is locally quenched, and through which a quantum of magnetic flux is threaded. An azimuthal supercurrent surrounds the core, decaying into the bulk over a characteristic distance on the order of the London penetration depth λ .

Vortices interact with one another when the supercurrent surrounding one couples to the magnetic flux threading the other and vice-versa. For type-II superconductors, this coupling is repulsive, and a system of many vortices finds its equilibrium in a triangular lattice configuration. However, vortices also interact with currents which are applied externally. In the presence of such a current, magnetic vortices feel a net force in the direction of the current flow.

Since the magnetic field of a superconducting vortex is visible in spin-polarized STM, my modification in principle allows for the study of the formation and manipulation of magnetic vortices in superconductors whose coherence length is on the nanometer scale - a difficult task for more conventional tools such as scanning SQUID magnetometry, whose spatial resolution is typically on the order of hundreds of nanometers or larger. This could be particularly interesting in studying disordered films such as the granular aluminum discussed in Sec. 7.5.1.2, where the presence of disorder on the nanometer scale drastically reduces the effective coherence length (and therefore the size of magnetic vortices). Such an experiment could involve studying how the formation of vortices correlates with the grain structure of the film, and on how the presence of grain boundaries affects their dynamics in the presence of a lateral bias current.

7.5.2.2 Magnetic Skyrmions

A second type of topological soliton which can be manipulated with lateral currents is the magnetic skyrmion, discussed at length in Sec. 3.2.2. Because skyrmions are hosted in magnetic materials, any lateral current applied to a sample will be (at least partially) spin-polarized. Due to the spin-orbit torque between a spin current and a magnetic structure, magnetic skyrmions also feel a force in the direction of an applied current. As a result, their dynamics could be studied by spin-polarized STM coupled with a lateral current bias.

Beyond being merely pushed back and forth, it has been shown [43] that skyrmions can be nucleated by the presence of a spatially-inhomogeneous spin current. Such a current distribution could be established by etching a constriction into the film using a scribe. From there, an applied current could nucleate a population of magnetic skyrmions, whose dynamics could then be studied in the presence of a spin current and an external magnetic field. This would provide an avenue to observe the *skyrmion Hall effect* [44], in which a "current" of magnetic skyrmions is deflected by a magnetic field.

Chapter 8: Experiment: Superconducting Gap Variation in Granular Aluminum Films

8.1 Overview

In the 1960's [45] it was discovered that aluminum thin films grown in low-level oxygen atmospheres exhibit surprising electrical characteristics at low temperature, such as a factor $2.6 \times$ enhancement of the superconducting critical temperature T_C [46] and increased critical magnetic fields H_c and anisotropies $H_{C\parallel}/H_{C\perp}$ [47]. It was quickly determined that the presence of oxygen during the film deposition process induces a granular structure in the film, with the oxygen migrating to the grain boundaries.

The corresponding inhomogeneity in the film has a profound impact on the nature of its superconductivity. The Ginzburg-Landau parameter $\kappa \equiv \lambda/\xi$ distinguishes Type-I ($\kappa < 1/\sqrt{2}$) and Type-II ($\kappa > 1/\sqrt{2}$) superconductors, with the latter exhibiting substantially higher T_C and H_C . Pure aluminum has a Ginzburg-Landau parameter $\kappa_{Al} \approx 0.0015$ [48], placing it firmly in the category of a Type-I superconductor. However, in the presence of disorder-induced scattering with mean free path ℓ , the coherence length is reduced to

$$\frac{1}{\xi} = \frac{1}{\xi_0} + \frac{1}{\ell} \quad (8.1)$$

If $\ell \ll \xi_0$ (the "dirty" superconductor limit), then $\xi \sim \ell$ and $\kappa \sim \lambda/\ell$ can be much larger than it would be if the disorder were absent. It was therefore argued [49] that the properties of the granular Al films could be attributed to a decrease in ξ for films with grain size less than ξ_0 , inducing a transition to Type-II superconductivity.

Since that time, granular superconductors have been studied extensively with various models being put forth to explain their unconventional properties. The oxide layers between grains are generally understood to act as Josephson links, such that the film as a whole behaves as an array of Josephson junctions. Abeles [50] argued convincingly that the charging energy between grains was too large to permit conventional Josephson coupling if the sheet resistance of the film was uniformly distributed among the oxide layers at the grain boundaries, and suggested instead that the films acted as percolation networks of well-coupled grains through a more resistive bulk. On the other hand, measurements of optical conductivity [51] and heat capacity [52] suggest that granular superconductivity is an effectively bulk phenomenon which is well-described by a BCS transition.

The disconnect between the percolation argument and the experimental evidence in favor of the bulk nature of granular superconductivity has remained something of an open question. Theoretical studies have focused on the role that the Mott insulator transition and resonant Kondo tunneling may play in circumventing the percolation argument. From a more applied perspective, substantial work [53, 54] has recently been directed at the characterization of the high kinetic inductance of granular superconductors for use in metrology and quantum computing.

Though much work has been done, questions persist about the unconventional nature of granular superconductivity. Spectroscopy measurements with nanoscale spatial resolution could provide valuable information to determine the nature of the inter-grain coupling and the small-

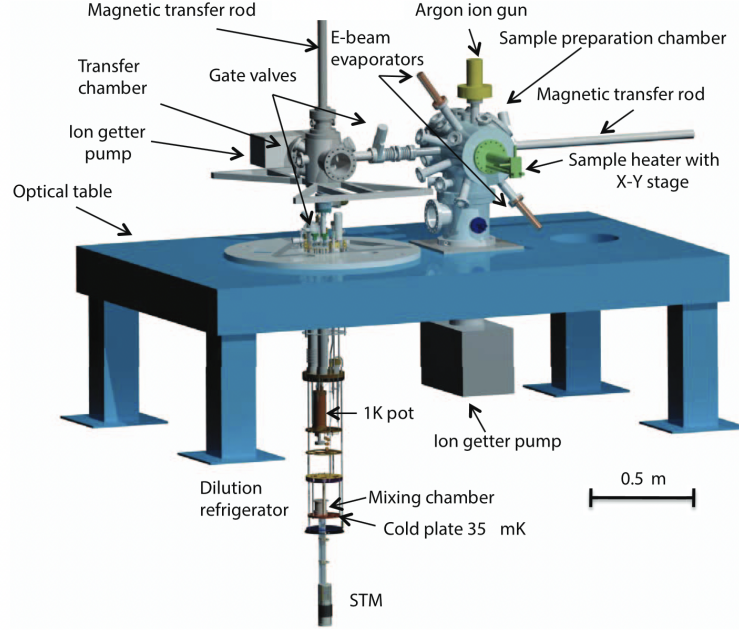


Figure 8.1: The mK-STM system at the Laboratory for Physical Sciences, consisting of a room temperature UHV subsystem for sample preparation and transport and a cryostat outfitted with a 1K pot and dilution refrigerator. Figure published in [4].

scale variation in the superconducting condensate. I therefore decided to utilize our mK-STM to provide insight into the nature of these fascinating systems.

8.2 Experimental Description

The experimental design was fairly straightforward. Aluminum films were sputter deposited onto Al_2O_3 substrates in the presence of low-pressure O_2 . After structural characterization in our ambient atomic force microscope, samples were transferred into UHV and cleaned by Ar^+ sputtering to remove the $\sim 3 - 5$ nm native oxide layer before being transferred into our mK-STM system outfitted with a superconducting Nb tip.

Spectroscopy maps were taken across the surface, with both conductance and differential conductance being measured at each point. The widths of the superconducting gaps were ex-

tracted to correlate them with the film topography and granular structure. The appearance of markedly different gap widths for adjacent grains or the quenching of superconductivity altogether in smaller grains would suggest a highly resistive grain boundary, possibly shedding light on the distribution of the primary current pathways in the film.

8.2.1 The mK-STM

Our mK-STM system is described in [4] and shown in 8.1. A room-temperature UHV subsystem consisting of a preparation chamber and a transfer chamber is mounted on a vibration-isolated optical table. The chambers are maintained at respective pressures of approximately 10^{-10} and 10^{-11} mbar by Varian Starcell 75 and 300 ion getter pumps and Vacuum Generators ST22 titanium sublimation pumps. The cryostat hangs below the table, and is encased in an outer vacuum can during operation. The STM is bolted to the cold plate of the dilution refrigerator, and is enclosed in an inner vacuum can.

8.2.1.1 The UHV Subsystem

The UHV subsystem is essentially identical to that of our 4K STM, described in Chapter 7. One significant difference is that the preparation chamber on the mK system also houses an STM system which can be used to verify the quality of a sample and the effectiveness of any preparation procedures before the sample is moved to the cryostat; this addition was deemed necessary due to the substantial time investment required to cool a sample without overheating the 1K pot or the mixing chamber.

8.2.1.2 The Cryostat and Microscope

The cryostat consists of a 1K pot and dilution refrigerator, as shown in 8.1. The STM is mounted to the bottom of the cold plate, which is the location of the dilution refrigerator's mixing chamber. The microscope is enclosed by an inner vacuum can, and the full cryostat (including the microscope and inner vacuum can) is subsequently enclosed in a secondary, outer vacuum can.

The microscope itself consists of a dual-tip modification to a standard Pan-type [42] STM design as shown in 8.2. The inner and outer tips can be independently and simultaneously controlled, but this feature was not utilized in the experiment. The body of the microscope is primarily made from macor ceramic.

8.2.1.3 The STM Tips

Our STM tips were etched from ~ 0.25 mm diameter Nb wire with a SF_6 plasma controlled via an Oxford Instruments Reactive Ion Etcher. Once fabricated, tips were evaluated for quality using an optical microscope¹ before being installed in the mK-STM. Once pumped down and cooled to base temperature (~ 35 mK), the partially oxidized tips were cleaned via high-voltage field emission on an Au(111) single crystal and sharpened with low-voltage pulsing.

At the time of this study, a superconducting tip had been installed in the system because unlike ordinary metallic tips which participate only in quasiparticle tunneling, superconducting tips constitute a direct probe of the superconducting condensate through the tunneling of Cooper pairs. The tip-sample junction essentially becomes a Josephson junction itself. This gives rise to

¹Tips were periodically sampled and checked via SEM, but were not subsequently installed due to carbon buildup

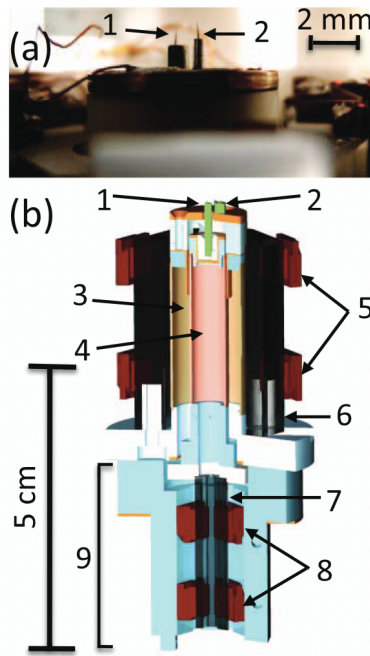


Figure 8.2: A close-up photograph of the outer and inner tip configuration (a) and an overall schematic (b) of the dual-tip mK STM. The outer (1) and inner (2) tips are mounted on separate outer (3) and inner (4) piezo tubes for lateral scanning control. Coarse motion for the outer assembly is provided by external piezo stacks (5) which move an outer sapphire prism (6). The inner tip can be independently controlled via an inner prism (7) and piezo stacks (8), while also riding along with the outer assembly (9). Figure published in [4].

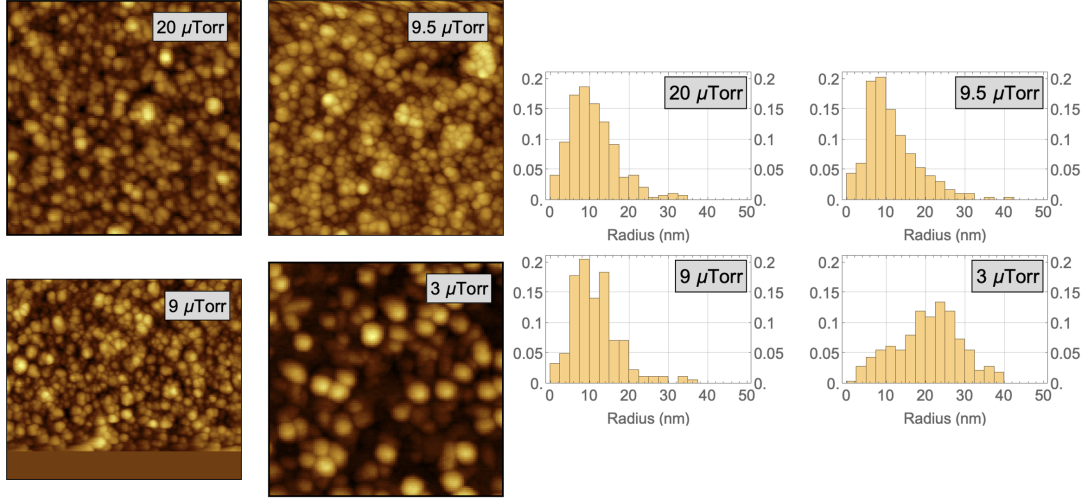


Figure 8.3: AFM topography and histograms of grain sizes for films grown with varying pressures of O_2 in the deposition chamber. The scan windows are all $500 \times 500 \text{ nm}^2$, with the exception of the $9 \mu\text{Torr}$ scan which was truncated (but remains to scale).

additional effects such as Shapiro steps and interesting subgap features which are worthy of their own dedicated study.

With that being said, the use of a superconducting tip does have a drawback in our case, insofar as the observed gap is the convolution of the tip gap and the sample gap. In the absence of a reliable calibration of the tip gap (which could vary substantially as the tip geometry changes throughout its life), it is not possible to precisely extract the sample gap at any given point. It has been observed that variations in the gap of a Nb tip after field emission can vary from several hundred μeV up to the bulk value of $\sim 1.53 \text{ meV}$ [55]. However, if the measurement is focused on gap *variation* over the course of the scan, then in the absence of major structural changes to the tip, any variation in the gap should in principle be attributable to the sample alone.

8.2.2 Sample Preparation

The films utilized in this study were grown to a thicknesses of 70 – 90 nm under a O_2 partial pressures ranging from of 3 μTorr up to 20 μTorr . The initial characterization took place under ambient conditions in our atomic force microscope. The variation of average grain size with the oxygen pressure was noted to behave as expected, with a marginal shift toward smaller grain size for higher pressures for $P \sim 10 - 20 \mu\text{Torr}$ and a substantial jump in grain size at 3 μTorr . $500 \times 500 \text{ nm}$ AFM scans and grain size histograms are shown in 8.3. Between 9 and 20 μTorr , the average grain size was $\sim 10 \text{ nm}$. As anticipated, there was a modest shift in the distribution of grain sizes, with lower pressures corresponding to slightly larger grains. However, this difference became very substantial at 3 μTorr , where the average grain radius jumped to $\sim 25 \text{ nm}$.

After scanning in the AFM, samples were rinsed with IPA and loaded into the mK system prep chamber where they were heated to remove H_2O and cleaned via Ar^+ sputtering to remove the $\approx 3 - 5 \text{ nm}$ native oxide layer from the surface. After sputtering, they were checked for film integrity and residual oxides before being moved through the transfer chamber into the cryostat.

8.2.3 Results

After the integrity of the tip and sample were confirmed via preliminary topography and point spectroscopy measurements, a spectroscopy map was performed over a $500 \times 500 \text{ nm}^2$ region in a 128×128 pixel grid. At each point, an 800 point (I, V) spectroscopy curve was taken. We used a $20\times$ voltage divider to take the spectroscopy curves between $\pm 10 \text{ mV}$, corresponding to a $25\mu\text{V}$ resolution.

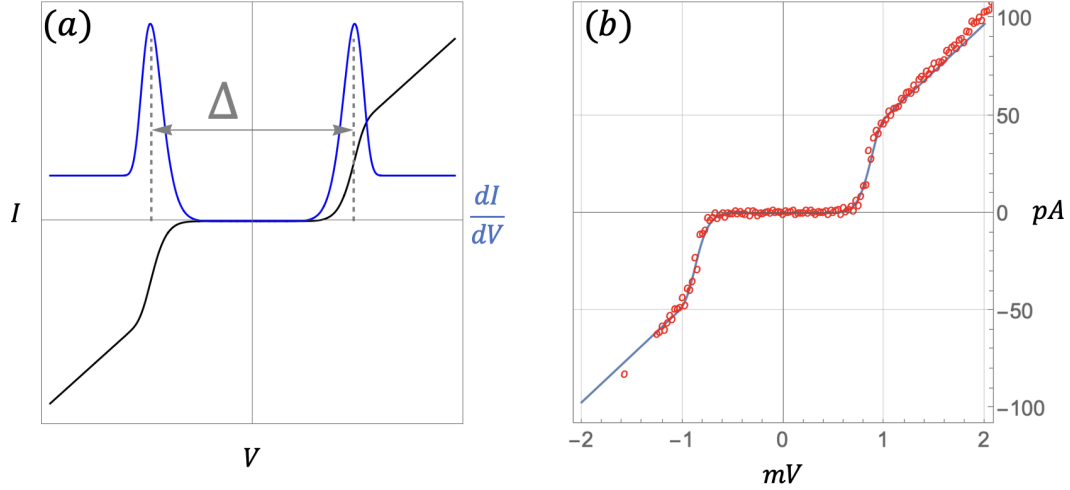


Figure 8.4: (a) The empirical fit function $I(V)$ given in 8.2 and its derivative. (b) An example of the fit being applied to one of the measured spectra. Despite its simplicity, it fits the data well and serves a straightforward way to extract the gap with substantial computational efficiency.

I fit the data using the function

$$I(V) = G_N V \left(1 - \exp \left[- \left(\frac{V - V_0}{\Delta/2} \right)^{10} \right] \right) + I_0 \quad (8.2)$$

with G_N , V_0 , Δ , and I_0 free parameters. This function was chosen because it is an extremely simple and remarkably accurate way to extract the gap data (see Fig. 8.4) and has no particular theoretical underpinnings.

The topography and gap variation data obtained from the spectroscopy map are shown in Fig. 8.5. The contour lines from the topography map are superimposed on top of the gap data for ease of comparison. The red points on the gap map (which are quite hard to find) were areas where the numerical fitting process failed. Its remarkable robustness was likely due in part to the simplicity of the fitting function.

The distribution of gaps over the entire map is shown in Fig. 8.6. The mean and standard deviation of the gaps (once outliers have been excluded) are $\mu_\Delta = 1.75$ mV and $\sigma_\Delta = 257$ μ V.

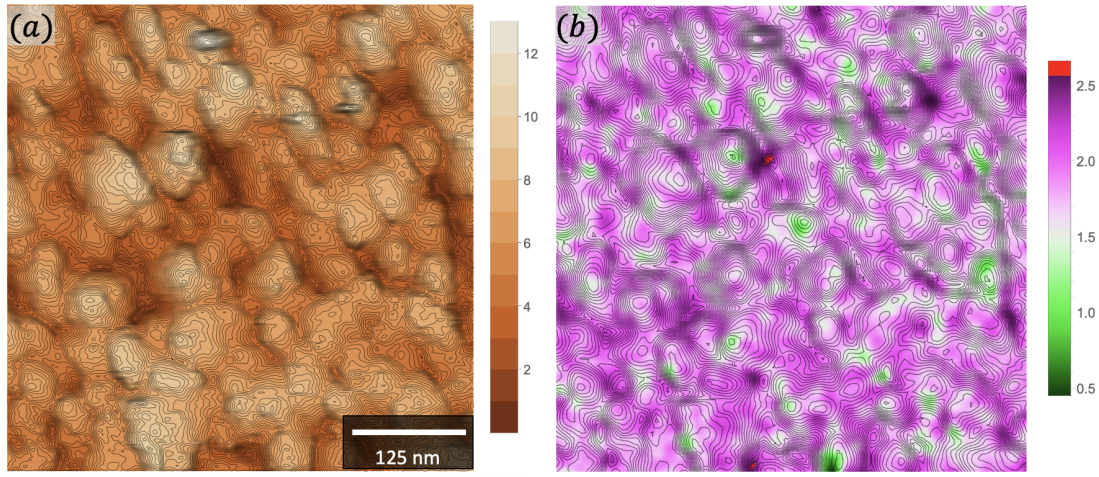


Figure 8.5: The topography (a) and measured gap widths (b) of a granular Al film, taken over a $500 \times 500 \text{ nm}^2$ scan area. The scale on the topography map is in nm, while the scale on the gap map is in mV. Both images have a lateral resolution of 128×128 pixels.

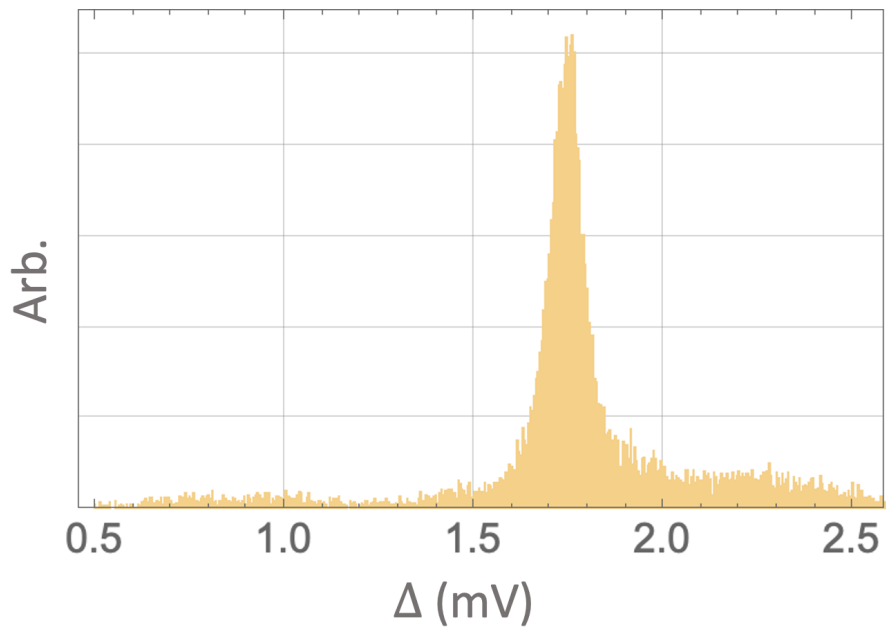


Figure 8.6: Histogram of the measured values for the superconducting map shown in Fig. 8.5. The mean gap is $\mu_{\Delta} = 1.75 \text{ mV}$ and the standard deviation is $\sigma_{\Delta} = 257 \text{ } \mu\text{V}$.

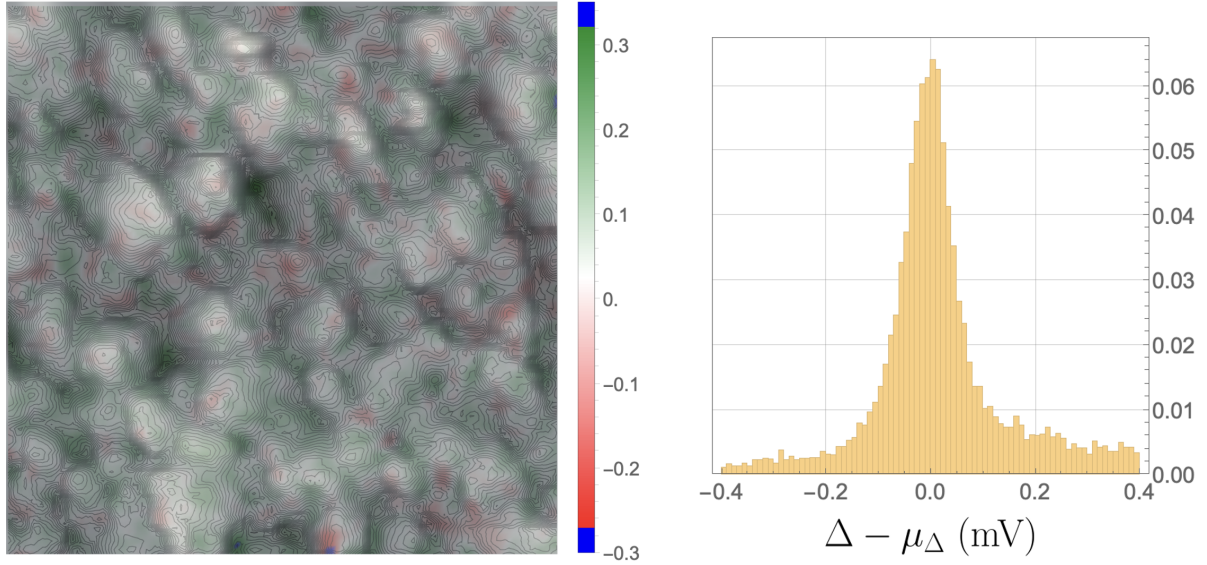


Figure 8.7: The deviation of Δ from its mean value, shown as a spatial plot and as a histogram. The color map has been superimposed on top of a grayscale topography image for visual clarity.

We find these values plausible, given that the gaps of pure (bulk) Nb and Al are 3.05 mV and 240 μ V, respectively [56], and we would not expect the Nb tip to have the same gap as bulk Nb due to its extremely constricted form factor near the apex of the tip.

The deviation of Δ from μ_Δ is shown in Fig. 8.7, along with its histogram. Visually, there seems to be a marked negative correlation between height and gap - the high points on the grains exhibit a noticeable gap suppression, while the gap appears to be enhanced in the valleys in between.

8.3 Discussion

The data obtained in this study suggests that there is a variation in the superconducting gap of the granular Al film on the length scale of the grains, in this case ~ 25 nm. The gap has a tendency to be suppressed near the tops of the higher grains and enhanced in between them.

The variation in Δ is of the correct order of magnitude to be associated with the superconducting condensate, and does not appear to be a tip artifact or a manifestation of experimental noise.

8.3.1 Comparison with Previous Work

Qualitatively similar results in NbN films were obtained as early as 1987 [57]. As in my results, there were surprisingly large variations in Δ (on the order of tens of percents or more) which were anticorrelated with topography and persisted over large distances. A later study which imaged an Abrikosov vortex lattice in a NbN film passivated with an Au layer also observed inhomogeneous superconductivity in zero field [58], but made no attempt to characterize it or to ascertain its origin.

More recently, a superconducting STM study was performed on granular TiN films [5] using both Nb and V tips. The significantly higher quality and volume of data in this study enabled the extraction not only of the superconducting gap Δ , but also the effective temperature T and effective tunneling barrier height Z from the BTK model of Andreev tunneling [59]. The authors concluded that localized heating due to quasiparticle excitations was responsible for the grain-to-grain variations in Δ . If the grains are generically poorly thermally coupled to the rest of the film, then localized heating could be very substantial; for a perfectly isolated Al grain of diameter 40 nm and height 50 nm, an Ohmic power $P = (1 \text{ nA})(1 \text{ V}) = 1 \text{ nW}$ would correspond to a heating rate of $\sim 10^6 \text{ K/s}$.

The most important figure in the TiN study for comparison with my data is shown in Fig. 8.8. The critical take-away is the notable anticorrelation between the topography and Δ map, which mirrors the anticorrelation shown in 8.5. Though my study did not yield data of sufficient

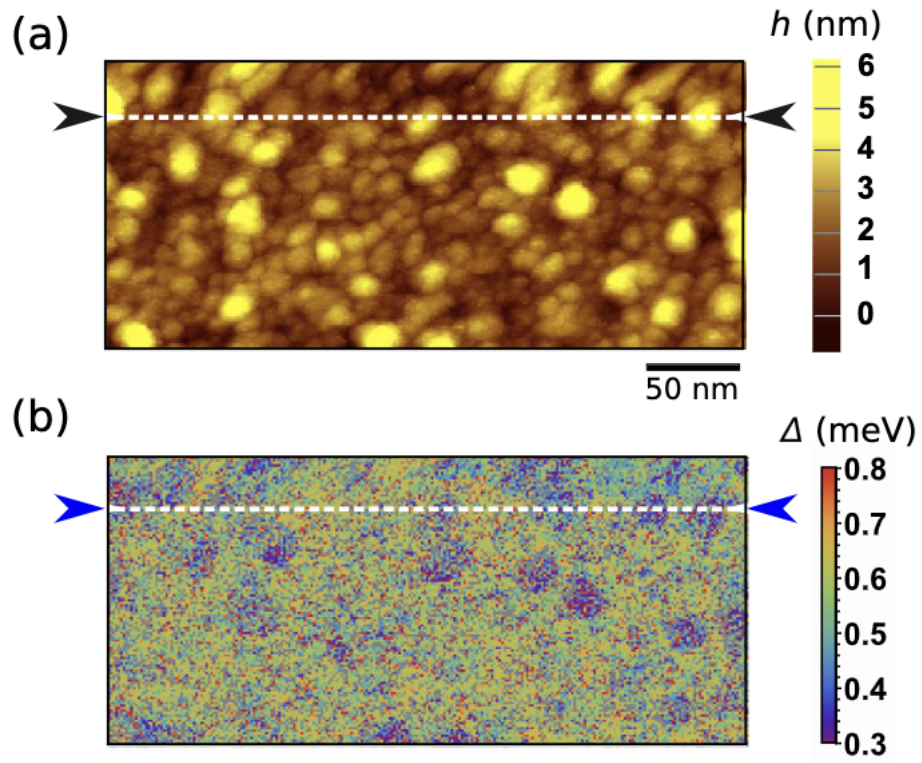


Figure 8.8: The topography (a) and Δ map (b) of a 50 nm thick granular TiN film, taken with permission from [5]. The anticorrelation between the grain topography and Δ is qualitatively similar to that shown in Fig. 8.5.

quality or volume to make a definitive statement about the origin of the Δ variation, it is consistent with the TiN study and its conclusions.

8.3.2 A Future Experiment: Vortex Matter in Granular Superconductors

The dynamics of vortices in granular superconductors has been studied for decades [60]. Recently, anomalously low dissipation by the vortex lattices in granular aluminum films excited by microwaves has motivated a study [61] which explored the idea that the states occupied by the normal electrons present in the cores of superconducting vortices may be gapped.

The dissipation which is observed in vortex matter is generally attributed to the thermal excitation of the normal electrons within the vortex core. A simple model [62] of the electron dynamics in the vortex core of a uniform type-II superconductor predicts a continuous energy spectrum, implying dissipation at all AC frequencies, but if the energy spectrum of the electrons confined to the vortex core were gapped then the quasiparticle excitations - and by extension, the dissipation - would be strongly suppressed below some threshold frequency. Precisely such a gap appears in granular materials when the coherence length ξ becomes comparable to the grain size ℓ ; the non-universal critical point has been found to be $\xi/\ell \simeq 1.4$ for a lattice of hexagonal grains [61].

An example of an experiment based on the one I performed which could shed light on this issue could go as follows. A selection of granular aluminum films with varying grain sizes would be prepared and characterized via macroscopic transport measurements. After being loaded into the UHV prep chamber of the mK-STM, the native oxide layer would be sputtered off before the sample is transferred to the microscope which has been outfitted with a magnetic tip².

²See Sec. 6.2.

After locating a suitable area, the topography and variation in Δ would be extracted in zero field. From there, an out-of-plane magnetic field would be applied to induce a vortex state, and the measurement would be repeated. The spin-polarized character of the tip would allow for the imaging of vortices via effective magnetometry, while the gap variations would be observed in the quasiparticle tunneling spectra. The existence of gaps in the core spectra could be ascertained and correlated with the grain sizes. Finally, the in-situ current biasing modification I made to the 4K-STM (see Chapter 7) could be ported to the mK-system, and the vortices could be subjected to a transverse DC or AC supercurrent.

8.3.3 Conclusion and Outlook

A complete understanding of granular superconductivity is needed in order to realize the promise these materials show for the next generation of quantum technologies. Important figures of merit such as the normal state resistivity, the critical current, and the kinetic inductance can be well-characterized empirically and correlated with the grain topography, but questions remain about the more fundamental behavior of the superconducting order parameter in the presence of irregular polycrystalline disorder.

Further consideration and evidence is needed to interpret what the data from my experiment suggests about the distribution of the resistivity within the film and in the nature of the superconducting condensate. Performing another measurement with a non-superconducting tip could provide more direct insight into the sample gap. Higher quality differential spectroscopy would permit the extraction of a greater volume of data, as in [5]. This would be important in determining whether the observed anticorrelations between the grain topography and supercon-

ducting gap have the same physical origin. It would also be interesting to perform additional studies with more granular films of different grain sizes and compositions to see whether the observed gap variations scale accordingly, or if a transition point exists which marks the onset of completely different behavior. Finally, the introduction of a magnetic field and of spin-polarized STM could open the door to fascinating experiments involving vortex matter, which would have dramatic implications for low-temperature power dissipation in the disordered superconductors which constitute the backbone of a great many next-generation quantum technologies.

Chapter 9: Experiment: Magnetic Skyrmions in Cr₂Te₃ Thin Films

The topological Hall effect [63] is generally considered a strong indicator of the presence of topologically non-trivial spin textures such as magnetic skyrmions in quasi-2D systems such as thin films or 2D electron gases (2DEGs) which arise at heterostructure interfaces. The interaction between the electron spin and the magnetic texture through which it moves can be modeled as a so-called *real-space Berry curvature*, whose action is formally identical to that of a magnetic field where the electron charge is replaced by its spin. The subsequent transverse deflection of the electrons results in an enhancement of the Hall signal in the presence of non-coplanar spin textures.

9.1 The Topological Hall Effect

The conductivity tensor of a 2D material generally takes the form of a 2×2 matrix. Such an object can always be decomposed as the sum of a scalar part, an antisymmetric part, and a traceless, symmetric part:

$$\sigma \equiv \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \sigma_0 \mathbf{1} + \sigma_H \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} + \sigma_\Delta \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{pmatrix} \quad (9.1)$$
$$\sigma_0 = \frac{a+d}{2} \quad \sigma_H = \frac{b-c}{2} \quad \sigma_\Delta = \frac{\sqrt{(a-d)^2 + (b+c)^2}}{2} \quad \theta = \tan^{-1} \left(\frac{b+c}{a-d} \right)$$

By definition, $\mathbf{J} = \sigma \mathbf{E}$, which means that σ_0 contributes a current parallel to E while σ_H contributes a current orthogonal to $\mathbf{E}$. The third contribution is due to the presence of rotational asymmetry, which results in an "easy" axis and a "hard" axis; aligning our coordinates with these axes yields the result

$$\begin{pmatrix} J_x \\ J_y \end{pmatrix} = \begin{pmatrix} (\sigma_0 + \sigma_\Delta)E_x + \sigma_H E_y \\ (\sigma_0 - \sigma_\Delta)E_y - \sigma_H E_x \end{pmatrix} \quad (9.2)$$

In many cases, it is simpler to work with the resistivity tensor $\rho = \sigma^{-1}$ instead, which takes the form

$$\rho = \rho_0 \mathbf{1} + \rho_H \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} - \rho_\Delta \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{pmatrix} \quad (9.3)$$

$$\rho_0 = \frac{\sigma_0}{\sigma_0^2 + \sigma_H^2 - \sigma_\Delta^2} \quad \rho_H = \frac{\sigma_H}{\sigma_0^2 + \sigma_H^2 - \sigma_\Delta^2} \quad \rho_\Delta = \frac{\sigma_\Delta}{\sigma_0^2 + \sigma_H^2 - \sigma_\Delta^2}$$

When a material is placed in a magnetic field $\mathbf{B} = B\hat{z}$, there is a contribution to ρ_H which is linear in B ; this is the *ordinary Hall effect*. If the film possesses an out-of-plane magnetization M , there is an additional contribution $\propto M$ which gives rise to the *anomalous Hall effect*. In addition to these two effects, it is observed that some magnetic materials exhibit a localized enhancement of ρ_H in the neighborhood of their coercive fields, resulting in a characteristic bump on their hysteresis loops; this is referred to as the *topological Hall effect*.

The name is derived from the fact that a topologically non-trivial magnetization - in the form of a lattice of magnetic skyrmions - would give rise to such a feature. Whether the effect is indeed skyrmionic in origin is a subtle question which is material-dependent - see Sec. [9.1.2](#).

9.1.1 2DEG Model

To get a good idea of how the topological Hall effect works, we can study a toy model consisting of a 2D free electron gas coupled to a spatially-varying background magnetization of fixed magnitude. The relevant Hilbert space is $L^2(\mathbb{R}^2) \otimes \mathbb{C}^2$, and our model Hamiltonian takes the form

$$H := -\frac{1}{2}\nabla^2 \otimes \mathbf{1} - \alpha \sum_{i=1}^3 \vec{n}_i \otimes \sigma_i \quad (9.4)$$

where α defines the strength of the coupling to the magnetic field, $\vec{n}_i$ is the unit vector which specifies its direction, and σ_i is the i^{th} Pauli matrix. Typically we suppress the explicit tensor product notation and simply write

$$H := -\frac{1}{2}\nabla^2 - \alpha \vec{n} \cdot \vec{\sigma}$$

The latter term has eigenvalues $\pm\alpha$ corresponding to spinors which are oriented parallel or antiparallel to $\vec{n}$. This suggests that we may simplify the problem by making a position-dependent change of spin basis such that spin up and down mean parallel and antiparallel to $\vec{n}$, respectively. Letting $\vec{n}$ be expressed in terms of spatially-varying polar and azimuthal angles θ and ϕ as

$$\vec{n} = \begin{pmatrix} \sin(\theta) \cos(\phi) \\ \sin(\theta) \sin(\phi) \\ \cos(\theta) \end{pmatrix}$$

this change of basis is implemented by the unitary operator

$$U = \begin{pmatrix} \cos\left(\frac{\theta}{2}\right) & -e^{-i\phi} \sin\left(\frac{\theta}{2}\right) \\ e^{i\phi} \sin\left(\frac{\theta}{2}\right) & \cos\left(\frac{\theta}{2}\right) \end{pmatrix} \quad (9.5)$$

In the transformed basis, we have

$$\vec{n} \cdot \sigma \mapsto U^\dagger (\vec{n} \cdot \sigma) U = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

as expected. However, because θ and ϕ are position-dependent, the momentum operator $-i\nabla$ transforms as

$$-i\nabla_j \mapsto U^\dagger (-i\nabla_j) U = -i\nabla_j + \sum_k a_{jk} \sigma_k, \quad \begin{pmatrix} a_{j1} \\ a_{j2} \\ a_{j3} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2}((\partial_j \theta) \sin(\phi) + (\partial_j \phi) \sin(\theta) \cos(\phi)) \\ \frac{1}{2}((\partial_j \theta) \cos(\phi) - (\partial_j \phi) \sin(\theta) \sin(\phi)) \\ \frac{1}{2}(\partial_j \phi)(1 - \cos(\theta)) \end{pmatrix}$$

Therefore, the Hamiltonian becomes

$$H \mapsto U^\dagger H U = \frac{1}{2}(-i\nabla + A)^2 - \alpha \sigma_3 \quad (9.6)$$

where $A_j = \sum_k a_{jk} \sigma_k = \begin{pmatrix} a_{j3} & a_{j1} - ia_{j2} \\ a_{j1} + ia_{j2} & -a_{j3} \end{pmatrix}$ constitutes a matrix-valued vector potential.

9.1.1.1 The Adiabatic Limit

Because A is proportional to the spatial derivatives of θ and ϕ , we may treat this problem perturbatively under the assumption that the spatial variation in the background magnetization is sufficiently slow. More precisely, we may study the motion of wavepackets in the $k\ell \ll 1$ regime, where k is the average wavevector of the packet and ℓ is the characteristic length scale of the variations in $\vec{n}$. This is the so-called *adiabatic* regime of the topological Hall effect¹.

In this picture, the 0^{th} order energy eigenstates are given by

$$\phi_{\uparrow\mathbf{k}}^0(\mathbf{x}) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \otimes e^{i\mathbf{k}\cdot\mathbf{x}} \quad \phi_{\downarrow\mathbf{k}}^0(\mathbf{x}) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \otimes e^{i\mathbf{k}\cdot\mathbf{x}} \quad (9.7)$$

which have energies $E_{\uparrow/\downarrow}^0 = \frac{k^2}{2} \mp \alpha$ and can be interpreted as quasi-free electrons whose spin is respectively parallel to and antiparallel to the background magnetization at each point. To get a physical sense of the effect the vector potential has on the system, we consider the motion of a wavepacket of the form

$$\psi_{\uparrow/\downarrow}(\mathbf{x}, t) = \int d^2k f(\mathbf{k}) \phi_{\uparrow/\downarrow\mathbf{k}}(\mathbf{x}) e^{-iE_{\uparrow/\downarrow}t} \quad (9.8)$$

The force on this wavepacket is given to lowest nonzero order by

$$\frac{d}{dt}\langle \mathbf{p} \rangle_{\psi_{\uparrow/\downarrow}} = -i\langle [\mathbf{p}, H] \rangle_{\psi_{\uparrow/\downarrow}} = -i\langle \nabla(A \cdot \nabla) \rangle_{\psi_{\uparrow/\downarrow}} = \pm \langle \mathbf{p} \rangle_{\psi_{\uparrow/\downarrow}} \times \mathbf{b} \quad (9.9)$$

where $\mathbf{b} = \vec{n} \cdot (\partial_x \vec{n} \times \partial_y \vec{n}) \hat{z}$ acts as an effective magnetic field. Defining $\mathbf{a} \equiv (a_{13}, a_{23}, 0)$, we

¹The topological Hall effect survives in the non-adiabatic regime, but the treatment is more subtle, c.f. [63].

may write $\mathbf{b} = \nabla \times \mathbf{a}$. The quantity $\mathbf{a}$ is often referred to as a *real-space Berry connection*, and $\mathbf{b}$ as the corresponding *real-space Berry curvature*. Note that if the off-diagonal terms in 9.6 are neglected, the Hamiltonian decouples into separate spin-up and spin-down Hamiltonians of the form

$$H_{\uparrow/\downarrow} = \frac{1}{2}(-i\nabla \pm \mathbf{a})^2 \mp \alpha \quad (9.10)$$

In summary, in the adiabatic limit in which the spatial variation in the background magnetization is small, the system can be understood as a quasi-free 2DEG whose spins are respectively parallel and antiparallel to the background magnetization, and which are subject to respective effective magnetic fields $\pm \mathbf{b}$. Averaging over the entire film of area A , we find

$$\bar{\mathbf{b}} = \frac{\hat{z}}{A} \int dx \int dy \vec{n} \cdot (\partial_x \vec{n} \times \partial_y \vec{n}) \equiv \frac{2\pi\Sigma}{A} \hat{z} \quad (9.11)$$

where we recognize the skyrmion number Σ from 3.2.2. Since an ordinary magnetic field produces a Hall resistivity ρ_{xy} , $\rho = \sigma^{-1}$ given by

$$\rho_{xy} = \frac{B}{n_c q} \quad (9.12)$$

where n_c is the density of charge carriers² and q is their charge, it follows that $\bar{b}$ makes an analogous contribution which we call the topological Hall effect.

²This assumes a single species of charge carriers for simplicity, but this assumption is inessential.

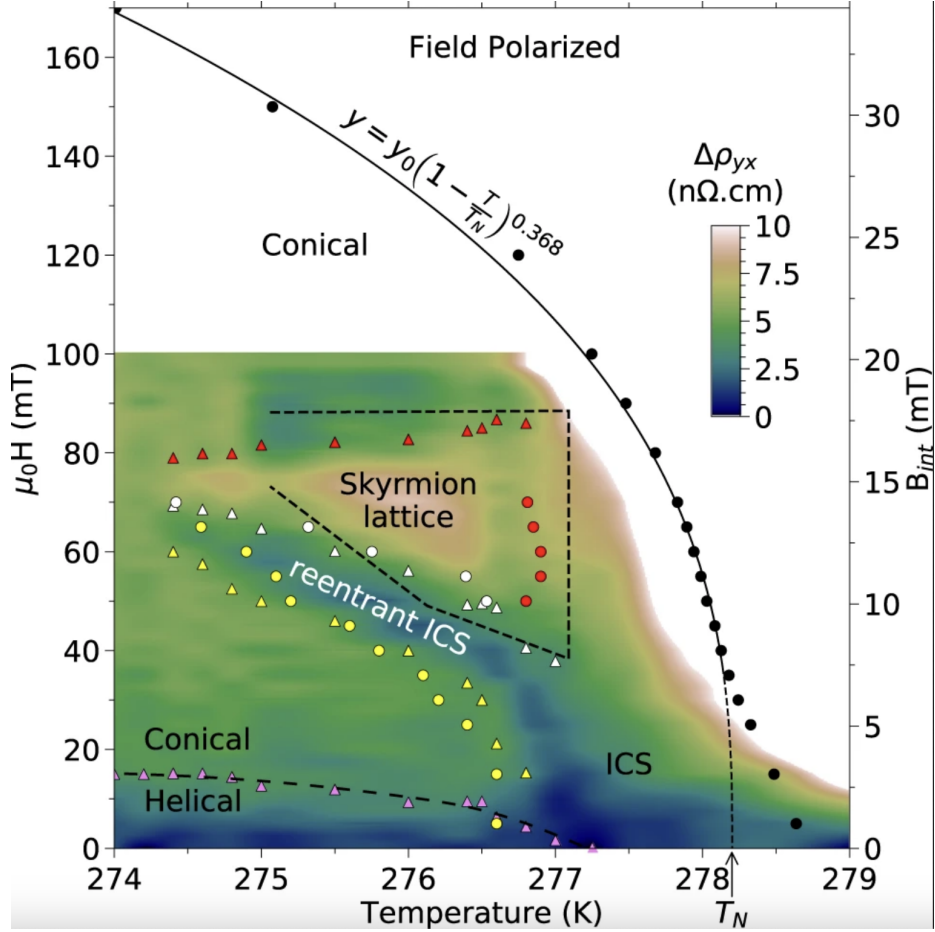


Figure 9.1: A magnetic phase diagram for MnSi, featuring a Skyrmion lattice phase identified via the Hall resistivity ρ_{xy} (color scale). Figure published in [6] and reused under a Creative Commons license.

9.1.1.2 Skyrmion Emergence at the Coercive Field

As noted in Sec. 3.2.3, the phase space for a 2D magnetic system can be quite complex, with the equilibrium configuration being a balance between a number of contributions. For example, Fig. 9.1 (taken from [6]) the existence of multiple magnetic phases as a function of temperature and out-of-plane magnetic field. As the magnetic field is swept during the course of an experiment, the system will generically occupy different phases which may or may not have the topological character discussed in the previous section.

In particular, the standard presentation of the topological Hall effect takes the form of localized enhancement of the Hall resistivity in the neighborhood of the coercive field during a hysteresis measurement, as shown in Fig. 9.2. Referring back to Sec. 3.2.3, the annihilation barrier for a stable magnetic skyrmion can be expressed in the form

$$E \sim -\alpha + \frac{\beta}{(\gamma - H_z)} \quad \alpha, \beta > 0 \quad \gamma - H_z > 0 \quad (9.13)$$

for $\alpha, \beta, \gamma, \delta$ independent of H_z (but dependent on material parameters and skyrmion dimensions). Requiring this to be positive yields

$$H_z > \gamma - \beta/\alpha \quad (9.14)$$

Therefore, a stable skyrmion can exist only in the range

$$\gamma - \frac{\beta}{\alpha} < H < \gamma \quad (9.15)$$

Qualitatively, H favors larger skyrmions over smaller ones. If $H < \gamma - \beta/\alpha$ then a skyrmion is unstable against collapse, whereas if $H > \gamma$ then a skyrmion will grow without bound and reverse the magnetization of the entire film. If $\beta/\alpha < \gamma$, then a stable skyrmionic phase can exist in a small neighborhood of $H = \gamma$ before the magnetization of the film begins to flip.

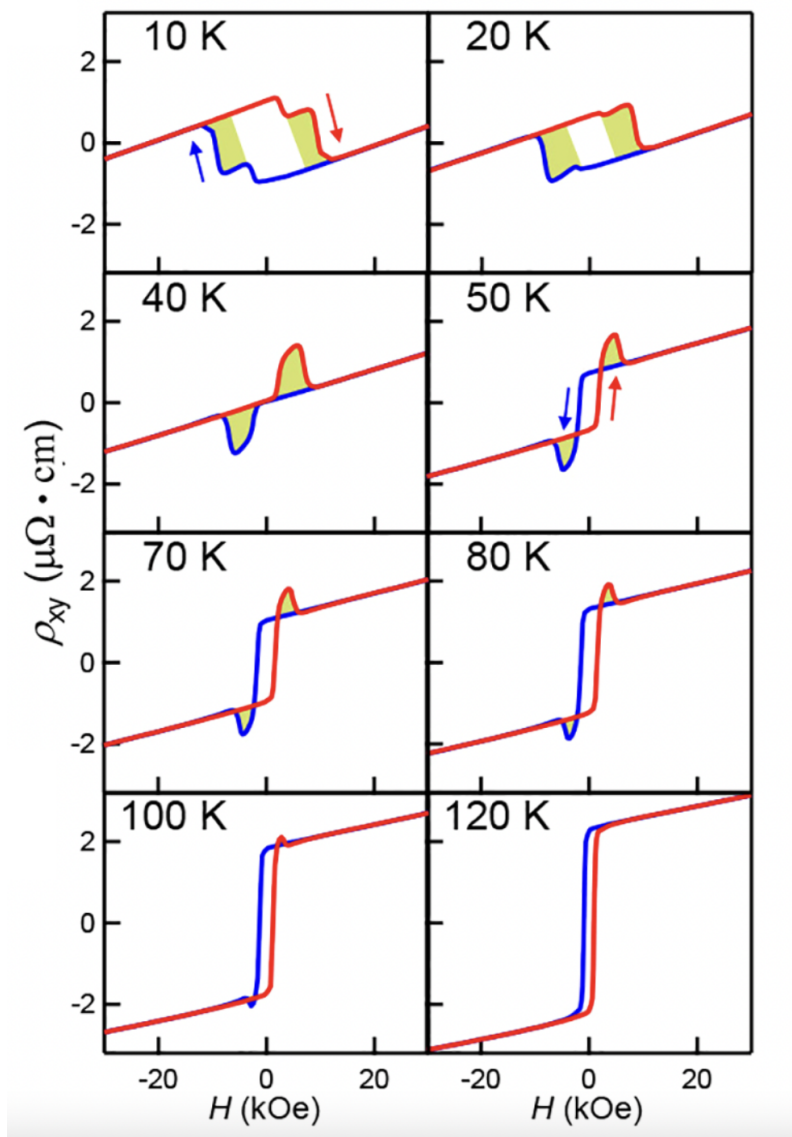


Figure 9.2: A typical presentation of the topological Hall effect, observed in a $\text{Bi}_2\text{Te}_3/\text{CrTe}_2$ heterostructure. The the sign of the anomalous Hall resistivity changes at $T = 40$ K, but the topological hall effect is not dramatically changed. Figure published with permission from [7], Copyright 2021 American Chemical Society.

9.1.2 An Alternative Model

Though a macroscopic³, topologically non-trivial magnetic texture is the underlying cause of the topological Hall effect, it is not the only potential cause of a localized feature on a hysteresis loop. One of the many⁴ subtleties of the anomalous Hall effect emerges when one considers the possibility of a sample consisting of several coexisting domains which exhibit different coercive fields.

A review of the two-component anomalous Hall effect, sometimes called the multiple-domain model, can be found in [65]. In practice, it can be difficult to identify whether a localized feature on a Hall signal is due to a skyrmionic phase or to the superposition of multiple AHEs, as illustrated in Fig. 9.3. If the AHEs have opposite signs and coercive fields, their superposition can exhibit features which are nearly identical to those observed in the topological hall effect.

Recent work (see [65]) has suggested methods for distinguishing the THE from the two-component AHE via transport measurements alone. Careful consideration of the signal in minor (non-saturating) hysteresis loops as well as the dependence of the effect on temperature and applied gate voltage can provide evidence for or against the presence of a topological magnetic phase. However, these methods may fail to be conclusive, especially in films with more than two coexisting magnetic domains.

The most convincing evidence for a skyrmionic THE comes from direct imaging of the magnetic structures themselves. The primary ways in which this can be done are via Lorentz transmission electron microscopy or spin-polarized scanning tunneling microscopy, shown in Fig. 9.4. With these tools, it is possible to image magnetic textures on nanometer scales, providing

³i.e. $\Sigma/A \neq 0$.

⁴For a thorough review of the anomalous Hall effect and its long and controversial history, see [64].

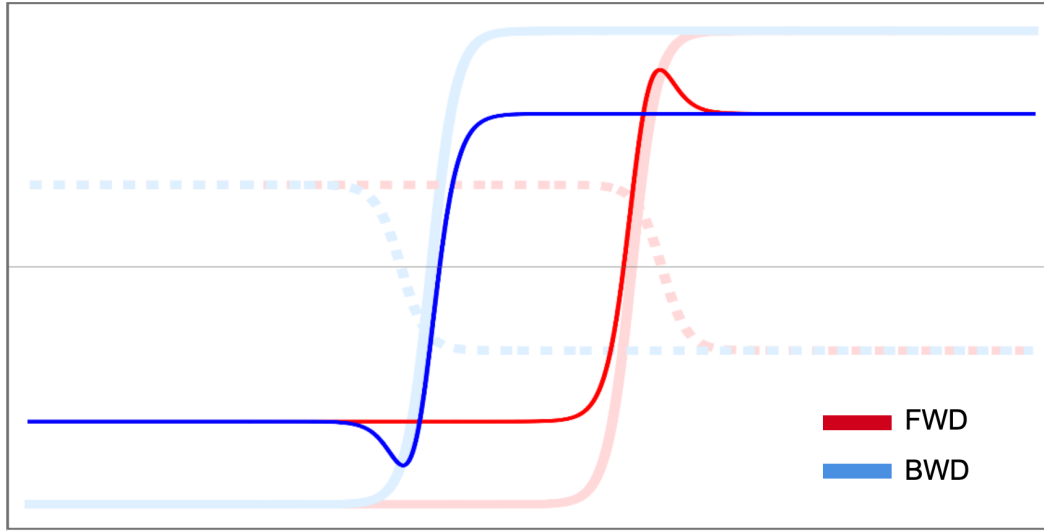


Figure 9.3: A cartoon illustration of a two-component AHE, in which the dark curve is the sum of the two lighter curves. The superposition of two ordinary AHE signals with opposite signs and different coercive fields gives rise to localized features which mimic the characteristic sign of a true (skyrmionic) topological Hall effect.

direct confirmation of the presence or absence of a topological contribution to the Hall effect.

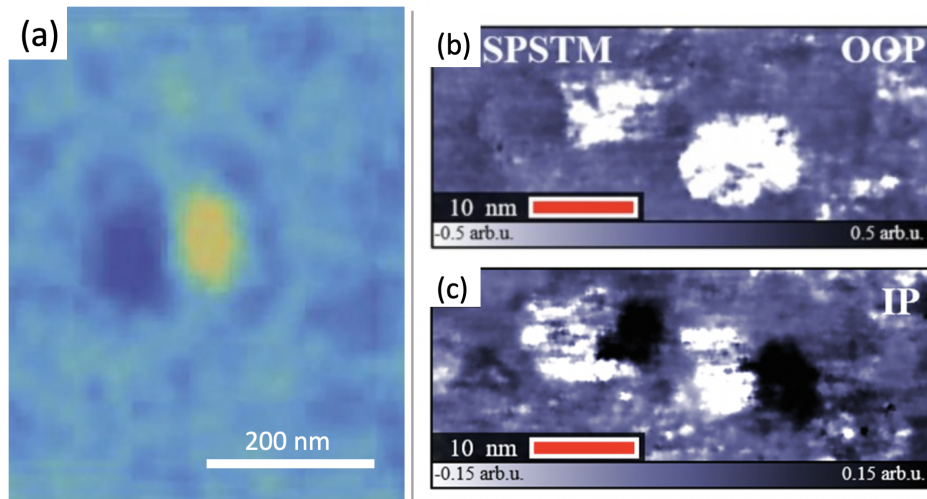


Figure 9.4: Imaging of magnetic skyrmions by Lorentz transmission electron microscopy (a) and out-of-plane and in-plane spin-polarized STM (b,c). Lorentz TEM has the advantage of not relying on an atomically sharp tunnel junction, but SPSTM offers the ability to resolve both in-plane and out-of-plane magnetization. The LTEM image is reproduced with permission from [8], while the SPSTM image is reproduced with permission from [9].

9.2 Possible Topological Hall Effect in Cr_2Te_3

In a recent work [10], thin films of ferromagnetic Cr_2Te_3 were observed to exhibit the localized features characteristic of a topological Hall effect, as shown in Fig. 9.5. This observation is surprising because Cr_2Te_3 is a centrosymmetric crystal; the bulk inversion symmetry implies that the bulk DMI coefficient vanishes. On the other hand, there are no interfaces to which one would expect to associate a strong interfacial DMI. Since skyrmions require a DMI in order to be stabilized, a skyrmionic phase of Cr_2Te_3 is not expected to exist.

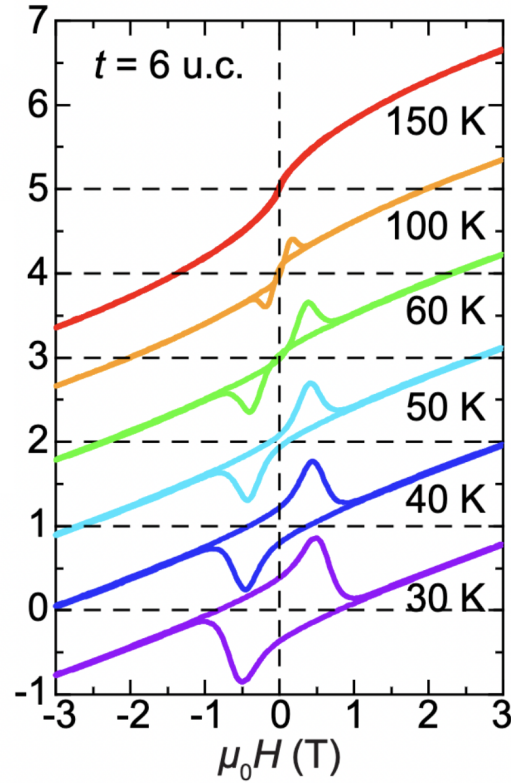


Figure 9.5: Hall measurements performed on a quasi-2D Cr_2Te_3 film with a thickness of 6 unit cells. The emergence of pronounced, localized humps in the vicinity of the coercive field suggests either the presence of a topological Hall effect or of a multi-domain anomalous Hall effect. Figure published with permission from [10].

This suggests that the observed features are more likely due to a multi-component AHE

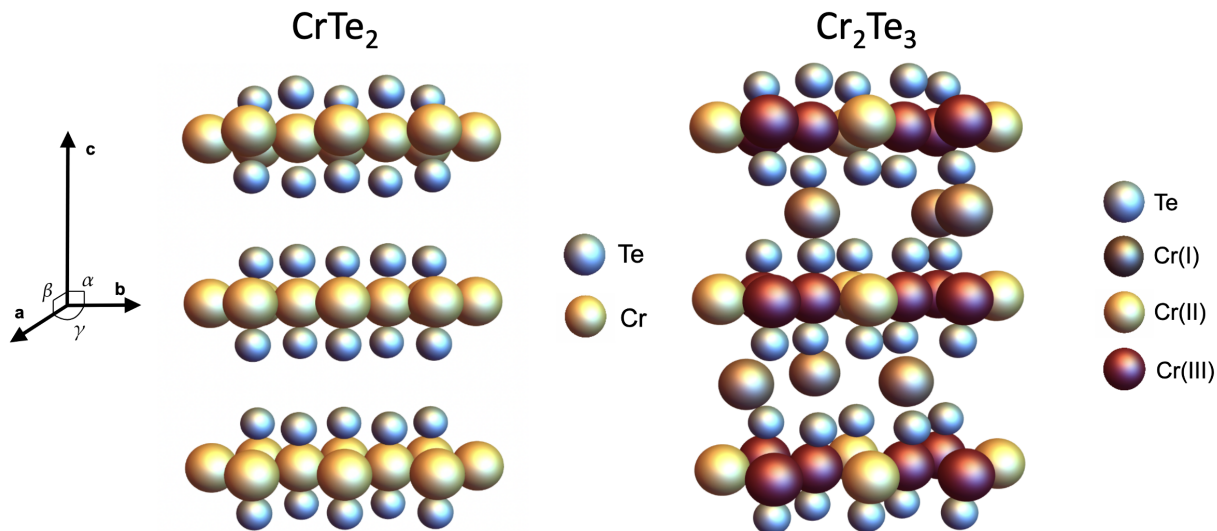


Figure 9.6: The crystal structures of CrTe_2 (left) and Cr_2Te_3 (right). CrTe_2 has a trigonal unit cell ($\alpha = \beta = 90^\circ, \gamma = 120^\circ$) with lattice constants $a = b = 3.86 \text{ \AA}$, $c = 6.2 \text{ \AA}$. Cr_2Te_3 also has a trigonal unit cell, but the lattice constants are enlarged to $a = b = 6.9 \text{ \AA}$, $c = 12.5 \text{ \AA}$, as the larger unit cell is required by the intercalation layers.

than to a THE. However, to provide more direct and convincing evidence, we studied the material using spin-polarized STM, as described in Sec. 6.2

9.2.1 Structure and Properties of Cr_2Te_3

9.2.1.1 Crystal Structure

The crystal structure of Cr_2Te_3 can be understood by starting from CrTe_2 as a backbone structure and adding partially occupied intercalation layers of Cr atoms. The two structures are shown in Fig. 9.6. CrTe_2 is a van der Waals (vdW) material consisting of Te-Cr-Te trilayers which are weakly coupled to one another by vdW interactions. Each trilayer consists of a triangular lattice of Cr atoms sandwiched between two triangular lattices of Te atoms. Each Cr atom is covalently bonded to six equivalent Te atoms (three above and three below) with bonds of length $2.71 \text{ \AA}$.

Cr_2Te_3 is obtained from CrTe by adding intercalation layers of Cr atoms in the vdW gaps between trilayers. The intercalation layers also form triangular lattices, rotated by $\pm 30^\circ$ with respect to the trilayer lattices and scaled up in linear dimension by a factor of $\sqrt{3}$. This produces three inequivalent Cr sites - (I) in the intercalation layers, (II) in the trilayers with no intercalated Cr directly above or below, and (III) in the trilayers with an intercalated Cr either directly above or directly below. Each Cr(I) atom is bonded to six equivalent Te atoms with bond length 2.78 Å, and each Cr(II) atom is bonded to six equivalent Te atoms with bond length 2.75 Å. The Cr(III) sites are bonded to two groups of three equivalent Te atoms. The bonds facing the nearest intercalated Cr(I) site are of length 2.78 Å, while the bonds facing the opposite group are of length 2.73 Å.

9.2.1.2 Magnetic Properties

Quasi-2D films of Cr_2Te_3 have attracted recent attention [66], [67] both for their potential use as 2D ferromagnets for applications in spintronics and data storage and for their potential integration with magnetic topological insulators. They exhibit very large ($K_u \sim 7 \times 10^5 \text{ J/m}^3$) perpendicular magnetic anisotropy, a high $T_c \sim 160 - 180 \text{ K}$, and a substantial coercivity on the order of 1 T at 4K. Its saturation magnetization is $\sim 4 \mu_B/\text{Cr} \approx 4 \times 10^5 \text{ A/m}$.

The centrosymmetry of the crystal structure⁵ implies that the bulk DMI coefficient must vanish, so in the absence of an interfacial effect magnetic skyrmions should not exist. Nevertheless, the magnetic structure of Cr_2Te_3 is subtle and not yet fully understood, due in part to the complex interactions between inequivalent Cr sites. The observed saturation magnetization is notably less than theoretically expected, and the existence of non-collinear spin textures has been

⁵ Cr_2Te_3 is inversion-symmetric about the Cr atom at the center of the unit cell, c.f. Fig. 9.6.

suggested as a possible explanation [68].

9.2.2 Experimental Description

Epitaxial films of Cr_2Te_3 were grown on Al_2O_3 and STO substrates via MBE by our collaborators in the Moodera group at MIT. The films varied in thickness from $\sim 5 - 20$ unit cells. Capping layers consisting of either Se or Te were deposited to protect the samples during transport. We came to favor Se capping because of the ease of its removal; fewer oxides survived to contaminate the Cr_2Te_3 surface, while Te oxides needed to be removed by light Ar^+ sputtering to obtain a clean surface.

Upon arrival, the samples were cut to approximately $4 \times 2 \text{ mm}^2$ to fit our sample holders. Debris was cleared from the sample surface with IPA and dry He gas before being loaded into our UHV prep chamber ($P \approx 10^{-10}$ mbar). After any native oxides were removed by Ar^+ sputtering ($E \approx 250$ eV), the films were heated above the sublimation point of the capping layer and monitored with a residual gas analyzer to ensure that the entire layer had been removed. From there, the samples were transferred to the 4K cryostat and allowed to cool for several hours before being studied with our low-temperature STM.

9.2.2.1 Procedure

Samples were initially imaged in zero magnetic field to determine their topographical characteristics. The large scale morphology of the films is shown in 9.7 and did not vary appreciably with film thickness. On large ($\sim 1 \text{ }\mu\text{m}$) scales the films were irregular, with island-like nanostructures of characteristic lateral size $\sim 50 \text{ nm}$ and heights $\sim 2 - 5 \text{ nm}$.

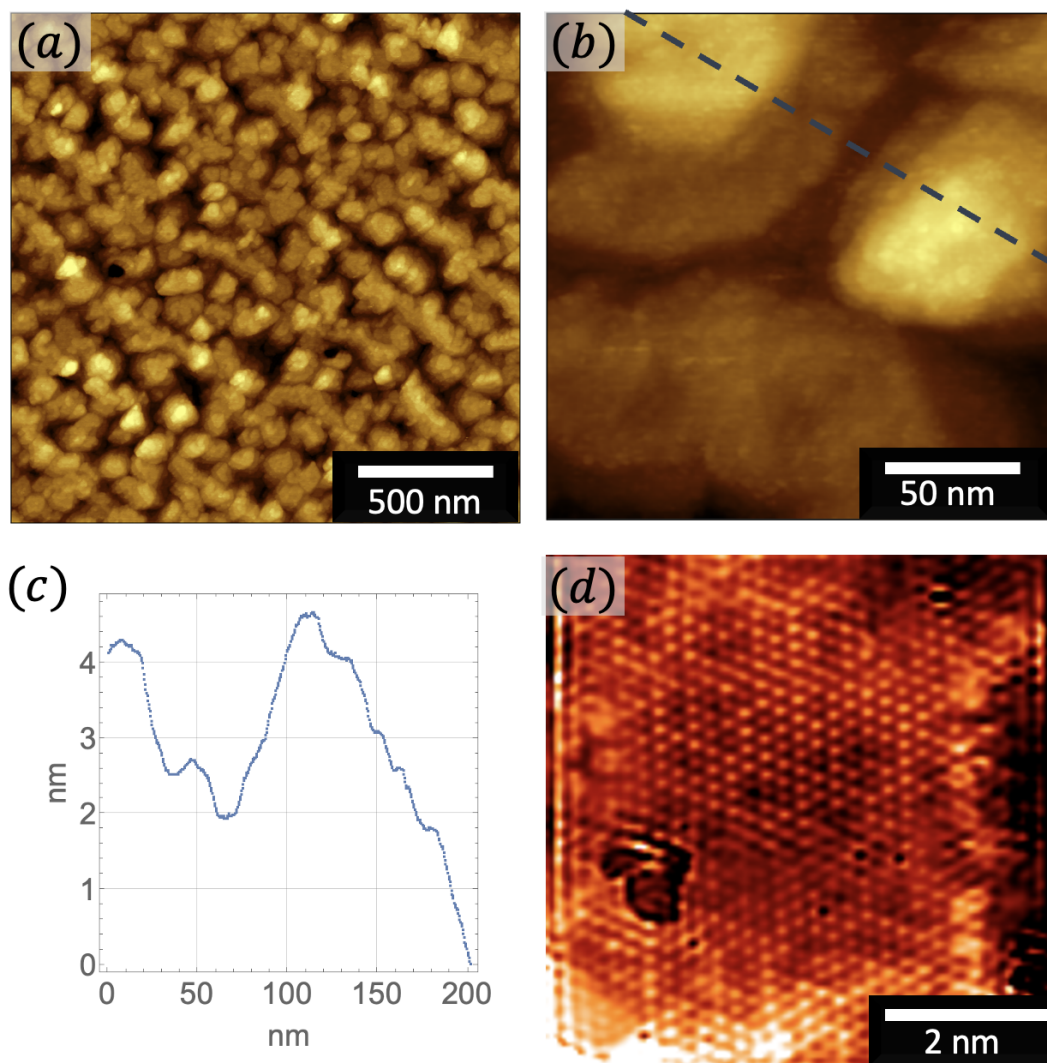


Figure 9.7: Surface morphology of a Cr_2Te_3 film at (a) $2\ \mu\text{m}$ and (b) $250\ \text{nm}$ scales with an electrochemically etched bulk Cr tip. (c) shows the height profile of the line cut shown in (b). (d) is a filtered atomic resolution image of one of the island features and was taken with an electrochemically etched tungsten tip. All images taken at $I_t = 100\ \text{pA}$ and $V = 100\ \text{mV}$.

To search for the emergence of a skyrmionic phase, we ran the sample through a hysteresis loop based on the data in Fig. 9.5. An out-of-plane magnetic field $\mu_0 H_z = +2$ T was applied to saturate the magnetization of the film before being slowly returned to zero and incrementally increased in the opposite direction. At each value of $\mu_0 H_z$, a simultaneous map of surface topography and differential conductance was taken using a lock-in amplifier, after which the tip was retracted by several hundred nanometers, the magnetic field was incremented, and the scan was repeated at the next field value.

9.2.2.2 Results

The differential conductance maps from the large $2 \times 2 \mu\text{m}^2$ scan are shown in Fig. 9.8. The color scale reflects the spin-projected density of states in the sample, with bright regions oriented parallel to the tip and the dark regions oriented in the opposite direction. The topography was identical throughout the measurement, indicating no major structural changes to the tip or surface. The spectroscopic images were decorrelated from topography by constructing a dI/dV vs. z scatter plot, applying a least-squares linear fit, and subtracting the result from the original data. No skyrmion lattice phase is visible on this scale; however, indications of stripe domains can be seen in the -0.9 T image in Fig. 9.8b.

The results of the small scale $250 \times 250 \text{ nm}^2$ scans are shown in Fig. 9.9. Processing for the images was the same as for the larger scan. We observe the emergence of two double-lobed structures of the same chirality at $\mu_0 H_z = +0.9$ T and their subsequent disappearance at $+1$ T. No corresponding features appeared in topography.

As shown in Fig. 9.10, the emergence of these double-lobed structures matches the ex-

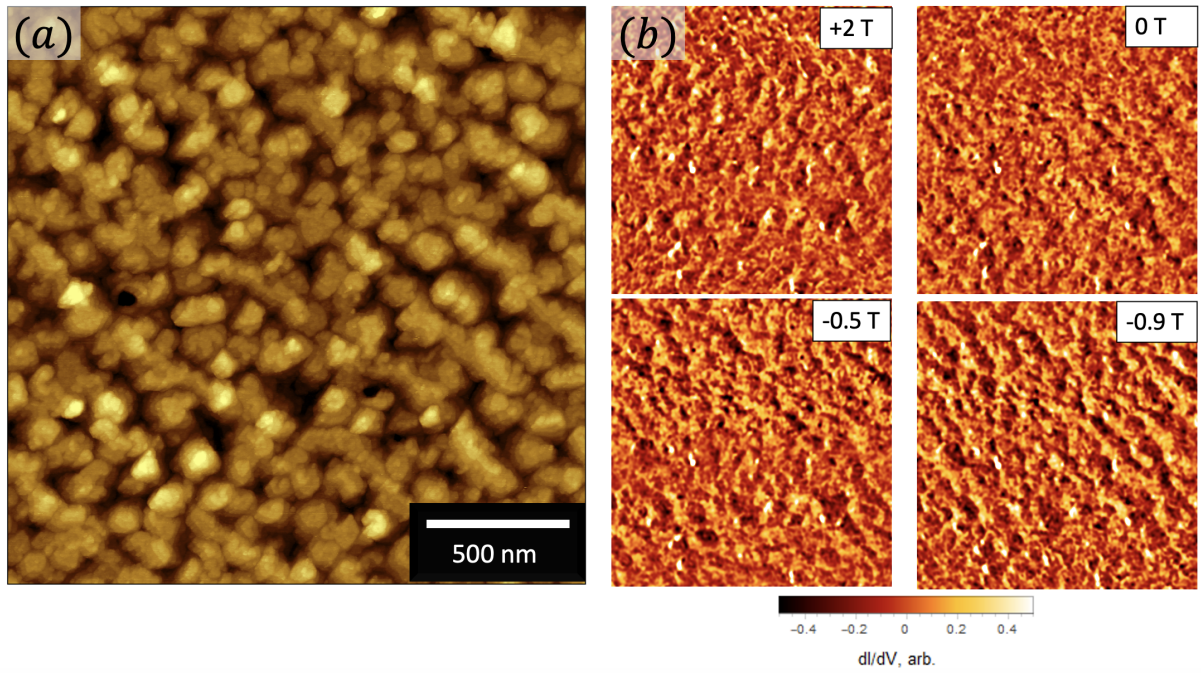


Figure 9.8: (a) Topography and (b) differential conductance maps of a Cr_2Te_3 film for $\mu H_z = +2$ T, 0 T, -0.5 T, and -0.9 T, taken with an in-plane polarized STM tip. dI/dV images were processed via Fourier filtering and decorrelation from topography. Possible stripe domains emerge in the -0.5 T and -0.9 T scans, but no signs of a skyrmionic phase are visible.

pected coercive field of the film. A line cut through the differential conductance map shows approximately symmetric positive and negative distortion. We note also the qualitative similarity to the in-plane SPSTM images of magnetic skyrmions in Fig. 9.4c.

9.2.2.3 Conclusions

The results of my experiment did not suggest the presence of a large-scale skyrmionic phase which could account for the Hall effect observed in transport measurements. Together with the lack of theoretical support for the existence of an interaction which could stabilize a skyrmion lattice, we interpret these results as being consistent with the conclusion in [10] that the observed Hall effect is due to a multi-component AHE.

However, as seen in Fig. 9.10, several isolated magnetic structures with the qualitative signature of magnetic skyrmions did emerge in the vicinity of the sample's coercive field. Because the exact polarization vector of our tip was unknown, the precise nature of the in-plane polarization of these structures could not be determined and they could not be definitively identified as topological. However, their presence raises the question as to whether signs of a larger-scale skyrmionic phase may have simply been overwhelmed by the topographical irregularity in the film. Additional corroboration via Lorentz TEM at a range of scales could provide a more conclusive demonstration in either case.

Additional studies with spin-polarized STM might focus on the collection of more data to corroborate the conclusions I have drawn in this study. In particular, repeating the measurement procedure using finer resolution in the external magnetic field could shed light on the formation, growth, and disappearance of the skyrmion-like features I observed. Optimization of the param-

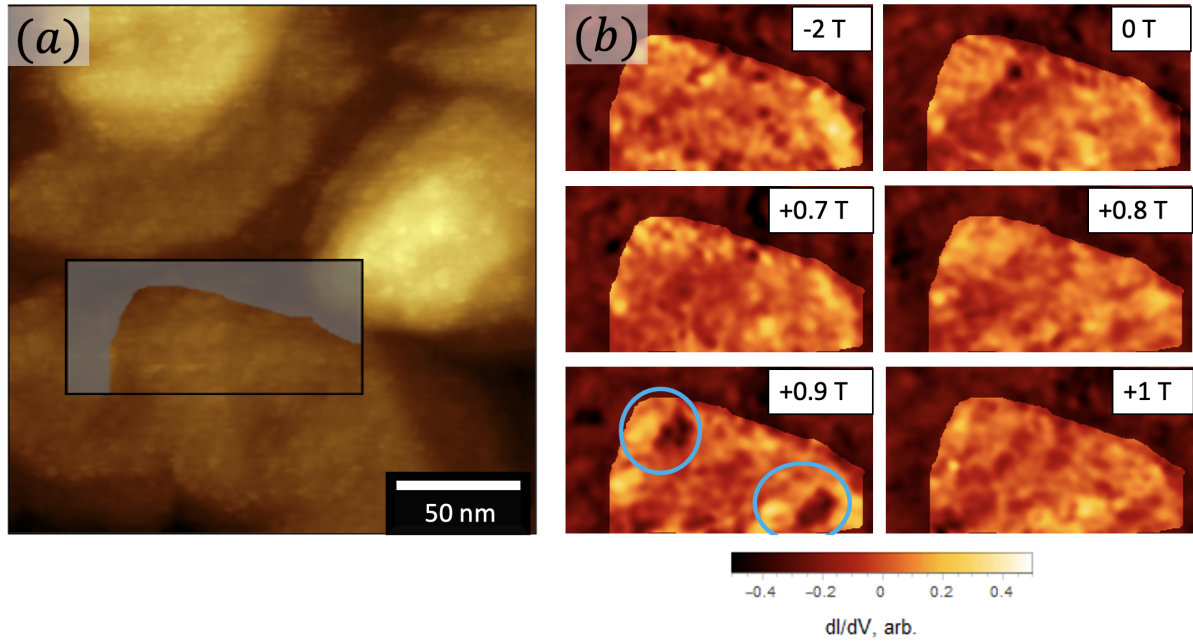


Figure 9.9: (a) Topography and (b) differential conductance maps of a Cr_2Te_3 film for $\mu H_z = -2$ T, 0 T, 0.5 T, and 0.9 T, taken with an in-plane polarized STM tip. A mask was applied to emphasize the terrace under examination. dI/dV images were processed via Fourier filtering and decorrelation from topography. Possible magnetic skyrmions emerge at $\mu_0 H_z = 0.9$ T before disappearing at higher fields.

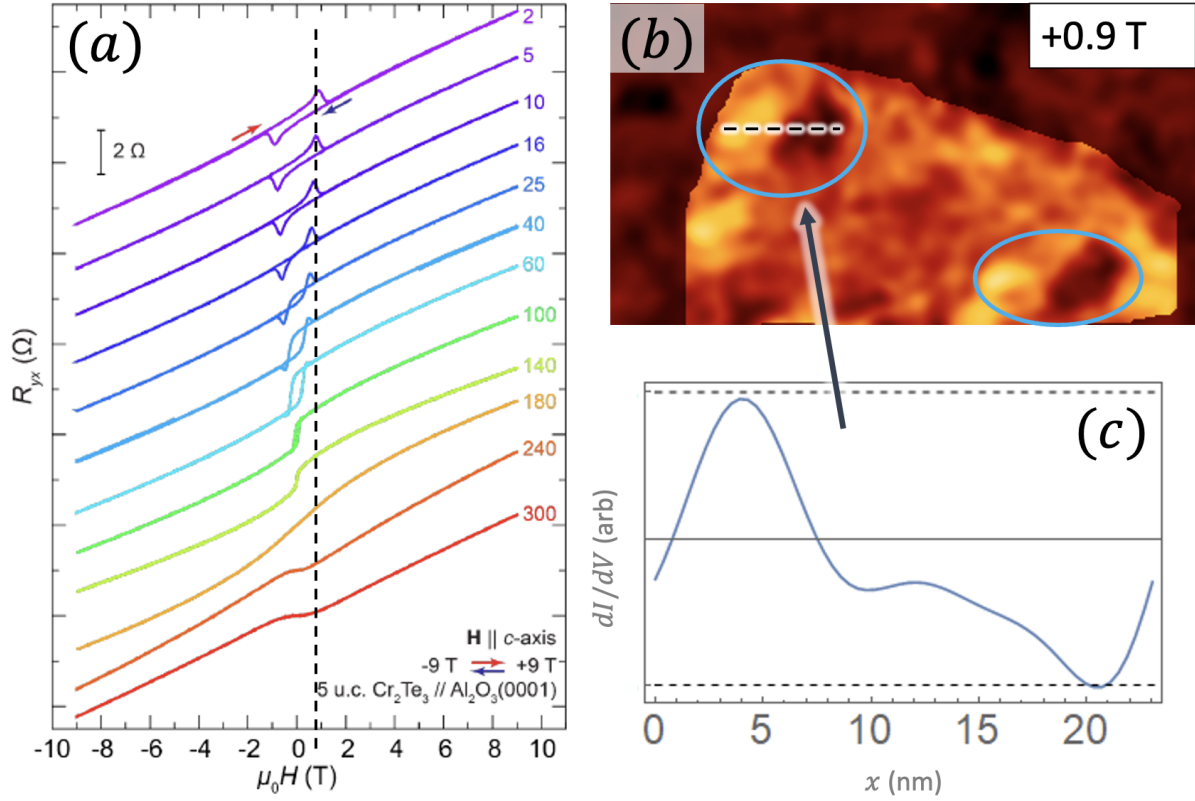


Figure 9.10: A comparison of transport measurements (a) vs. the emergence of possibly skyrmionic structures at $\mu_0 H_z = 0.9 \text{ T}$. Based on the hysteresis curves, this is in the immediate vicinity of the coercive field of the film. (c) shows a line cut through the differential conductance map which exhibits symmetric positive and negative variation across the structure, as one would expect from a magnetic skyrmion. Figure (a) published with permission from [10].

eters of the lock-in amplifier which was used to collect the differential spectroscopy data would sharpen any magnetic features and help disentangle them from the topographical noise. More complete characterization of the magnetic tip would also be extremely helpful in understanding the precise magnetization of these structures, and could more definitively confirm their topological nature (or lack thereof). Of course, greater spatial uniformity in the Cr_2Te_3 topography would provide a drastic improvement in the data quality, and so any advances in sample fabrication would have significant impact on the results.

Beyond simple repetition of the broad measurement procedure utilized in this study, it would also be interesting to make use of the current-biasing modification which was presented in Chapter 7. Because a lateral current through Cr_2Te_3 would be partially spin-polarized, magnetic skyrmions would be subject to a force in the presence of short lateral current pulses. The dynamics of the observed features in the presence of such a spin current would provide further clues as to their nature and origin, and an artificial constriction could be etched into the film to allow for the artificial nucleation of additional skyrmions as described in [43].

Finally, it would be interesting to repeat the localized hunt for magnetic skyrmions in more complex structures. For example, a Cr_2Te_3 film could be interfaced with topological insulators, which is known to produce an interface which exhibits an extremely strong topological Hall effect [69]. This interfacial THE could be studied locally using the methods employed in this study, provided that the interface is sufficiently close to the top surface of the heterostructure. This interface also hosts spin-momentum locked interface states, and so once again the current-biasing capability described in Chapter 7 could be used to push a non-equilibrium current and produce a very strong skyrmion Hall effect [44] whose dynamics could be studied locally.

Bibliography

- [1] Toyo Kazu Yamada. *Spin-Polarized Scanning Tunneling Microscopy*, pages 637–641. Springer Singapore, Singapore, 2018.
- [2] H. Zu, W. Dai, and A.T.A.M. de Waele. Development of dilution refrigerators—a review. *Cryogenics*, 121:103390, 2022.
- [3] Michael Dreyer, Jonghee Lee, Hui Wang, and Barry Barker. A low temperature scanning tunneling microscopy system for measuring si at 4.2 k. *Review of Scientific Instruments*, 81(5):053703, 2010.
- [4] Anita Roychowdhury, M. A. Gubrud, R. Dana, J. R. Anderson, C. J. Lobb, F. C. Wellstood, and M. Dreyer. A 30 mK, 13.5 T scanning tunneling microscope with two independent tips. *Review of Scientific Instruments*, 85(4), 04 2014. 043706.
- [5] Wan-Ting Liao, T. P. Kohler, K. D. Osborn, R. E. Butera, C. J. Lobb, F. C. Wellstood, and M. Dreyer. Scanning tunneling andreev microscopy of titanium nitride thin films. *Phys. Rev. B*, 100:214505, Dec 2019.
- [6] Maxime Leroux, Matthew Stolt, Song Jin, Douglas Pete, Charles Reichhardt, and Boris Maiorov. Skyrmion lattice topological hall effect near room temperature. *Scientific Reports*, 8, 10 2018.
- [7] Xiaoqian Zhang, Siddhesh C. Ambhire, Qiangsheng Lu, Wei Niu, Jacob Cook, Jidong Samuel Jiang, Deshun Hong, Laith Alahmed, Liang He, Rong Zhang, Yongbing Xu, Steven S.-L. Zhang, Peng Li, and Guang Bian. Giant topological hall effect in van der waals heterostructures of crte₂/bi₂te₃. *ACS Nano*, 15(10):15710–15719, 2021. PMID: 34460216.
- [8] Shawn D. Pollard, Joseph A. Garlow, Jiawei Yu, Zhen Wang, Yimei Zhu, and Hyunsoo Yang. Observation of stable néel skyrmions in cobalt/palladium multilayers with lorentz transmission electron microscopy. *Nature Communications*, 8(1):14761, 2017.
- [9] Joseph P. Corbett, Keng-Yuan Meng, Jacob J. Repicky, Reyes Garcia-Diaz, James.R. Rowland, Adam.S. Ahmed, Noburo Takeuchi, Jonathan Guerrero-Sanchez, Fengyuan.Y. Yang, and Jay.A. Gupta. Spin polarized stm imaging of nanoscale néel skyrmions in an srir₃/srro₃ perovskite bilayer. *Applied Surface Science*, 599:153766, 2022.
- [10] Hang Chi, Yunbo Ou, Tim B. Eldred, Wenpei Gao, Sohee Kwon, Joseph Murray, Michael Dreyer, Robert E. Butera, Alexandre C. Foucher, Haile Ambaye, Jong Keum, Alice T.

- Greenberg, Yuhang Liu, Mahesh R. Neupane, George J. de Coster, Owen A. Vail, Patrick J. Taylor, Patrick A. Folkes, Charles Rong, Gen Yin, Roger K. Lake, Frances M. Ross, Valeria Lauter, Don Heiman, and Jagadeesh S. Moodera. Strain-tunable berry curvature in quasi-two-dimensional chromium telluride. *Nature Communications*, 14(1):3222, Jun 2023.
- [11] G. Binnig and H. Rohrer. Scanning tunneling microscopy. *Surface Science*, 126(1):236–244, 1983.
- [12] John Dagata, J. Schneir, Howard Harary, C Evans, Michael Postek, and J. Bennett. Modification of hydrogen-passivated silicon by a scanning tunneling microscope operating in air. (56), 1990-12-31 00:12:00 1990.
- [13] K. J. Dwyer, Michael Dreyer, and R. E. Butera. Stm-induced desorption and lithographic patterning of cl–si(100)-(2 × 1). *The Journal of Physical Chemistry A*, 123(50):10793–10803, 2019. PMID: 31725292.
- [14] Klaus von Klitzing. The quantized hall effect. *Rev. Mod. Phys.*, 58:519–531, Jul 1986.
- [15] Mikio Nakahara. *Geometry, Topology, and Physics*. CRC Press, 2003.
- [16] Theodore Frankel. *The Geometry of Physics: An Introduction*. Cambridge University Press, 2011.
- [17] F.P. Schuller. The geometrical anatomy of theoretical physics. University Lecture, 2013.
- [18] N. D. Mermin. The topological theory of defects in ordered media. *Rev. Mod. Phys.*, 51:591–648, Jul 1979.
- [19] Hans-Benjamin Braun. Fluctuations and instabilities of ferromagnetic domain-wall pairs in an external magnetic field. *Phys. Rev. B*, 50:16485–16500, Dec 1994.
- [20] Niklas Romming, André Kubetzka, Christian Hanneken, Kirsten von Bergmann, and Roland Wiesendanger. Field-dependent size and shape of single magnetic skyrmions. *Phys. Rev. Lett.*, 114:177203, May 2015.
- [21] Felix Büttner, Ivan Lemesh, and Geoffrey S. D. Beach. Theory of isolated magnetic skyrmions: From fundamentals to room temperature applications. *Scientific Reports*, 8(1):4464, Mar 2018.
- [22] Ashcroft and Mermin. *Solid State Physics*. Brooks/Cole, 1976.
- [23] Efthimios Kaxiras and John D. Joannopoulos. *Quantum Theory of Materials*. Cambridge University Press, 2019.
- [24] Nicola Marzari, Arash A. Mostofi, Jonathan R. Yates, Ivo Souza, and David Vanderbilt. Maximally localized wannier functions: Theory and applications. *Rev. Mod. Phys.*, 84:1419–1475, Oct 2012.

- [25] Christian Brouder, Gianluca Panati, Matteo Calandra, Christophe Mourougane, and Nicola Marzari. Exponential localization of wannier functions in insulators. *Phys. Rev. Lett.*, 98:046402, Jan 2007.
- [26] Jacques Des Cloizeaux. Energy bands and projection operators in a crystal: Analytic and asymptotic properties. *Phys. Rev.*, 135:A685–A697, Aug 1964.
- [27] Jiaheng Li, Zetao Zhang, Chong Wang, Huaqing Huang, Bing-Lin Gu, and Wenhui Duan. Topological semimetals from the perspective of first-principles calculations. *Journal of Applied Physics*, 128(19), 11 2020.
- [28] Giandomenico Palumbo. Non-abelian tensor berry connections in multiband topological systems. *Phys. Rev. Lett.*, 126:246801, Jun 2021.
- [29] Lukasz Fidkowski, T. S. Jackson, and Israel Klich. Model characterization of gapless edge modes of topological insulators using intermediate brillouin-zone functions. *Phys. Rev. Lett.*, 107:036601, Jul 2011.
- [30] Shinsei Ryu, Andreas P Schnyder, Akira Furusaki, and Andreas W W Ludwig. Topological insulators and superconductors: tenfold way and dimensional hierarchy. *New Journal of Physics*, 12(6):065010, jun 2010.
- [31] Liang Fu, C. L. Kane, and E. J. Mele. Topological insulators in three dimensions. *Phys. Rev. Lett.*, 98:106803, Mar 2007.
- [32] J. Bardeen. Tunnelling from a many-particle point of view. *Phys. Rev. Lett.*, 6:57–59, Jan 1961.
- [33] J. Tersoff and D. R. Hamann. Theory of the scanning tunneling microscope. *Phys. Rev. B*, 31:805–813, Jan 1985.
- [34] Alex D Gottlieb and Lisa Wesoloski. Bardeen’s tunnelling theory as applied to scanning tunnelling microscopy: a technical guide to the traditional interpretation. *Nanotechnology*, 17(8):R57, mar 2006.
- [35] Christian E. Matt, Yu Liu, Harris Pirie, Nathan C. Drucker, Na Hyun Jo, Brinda Kuthanazhi, Zhao Huang, Christopher Lane, Jian-Xin Zhu, Paul C. Canfield, and Jennifer E. Hoffman. Spin-polarized imaging of strongly interacting fermions in the ferrimagnetic state of the weyl candidate cebi. *Phys. Rev. B*, 105:085134, Feb 2022.
- [36] Ting Xie, Michael Dreyer, David Bowen, Dan Hinkel, R. E. Butera, Charles Krafft, and Isaak Mayergoyz. Scanning tunneling microscopy measurements of the spin Hall effect in tungsten films by using iron-coated tungsten tips. *AIP Advances*, 8(5), 12 2017. 055914.
- [37] Masahiro Haze, Hung-Hsiang Yang, Kanta Asakawa, Nobuyuki Watanabe, Ryosuke Yamamoto, Yasuo Yoshida, and Yukio Hasegawa. Bulk ferromagnetic tips for spin-polarized scanning tunneling microscopy. *Review of Scientific Instruments*, 90(1):013704, 01 2019.

- [38] S Murphy, J Osing, and I.V Shvets. Fabrication of submicron-scale manganese–nickel tips for spin-polarized stm studies. *Applied Surface Science*, 144-145:497–500, 1999.
- [39] Yang Dong, R. M. Feenstra, R. Hey, and K. H. Ploog. Scanning tunneling potentiometry of semiconductor junctions. *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena*, 20(4):1677–1681, 08 2002.
- [40] Siddharth Tyagi, Michael Dreyer, David Bowen, Dan Hinkel, Patrick J. Taylor, Adam L. Friedman, Robert E. Butera, Charles Krafft, and Isaak Mayergoyz. Study of surface spin-polarized electron accumulation in topological insulators using scanning tunneling microscopy. *IEEE Magnetics Letters*, 11:1–4, 2020.
- [41] Michael Dreyer, Joseph Murray, Esther Frederick, and Robert E. Butera. Imaging three phases of iodine on ag (111) using low-temperature scanning tunneling microscopy. *Surface Science*, 721:122081, 2022.
- [42] S. H. Pan, E. W. Hudson, and J. C. Davis. The refrigerator based very low temperature scanning tunneling microscope. *Review of Scientific Instruments*, 70(2):1459–1463, 1999.
- [43] Wanjun Jiang, Pramey Upadhyaya, Wei Zhang, Guoqiang Yu, M. Benjamin Jungfleisch, Frank Y. Fradin, John E. Pearson, Yaroslav Tserkovnyak, Kang L. Wang, Olle Heinonen, Suzanne G. E. te Velthuis, and Axel Hoffmann. Blowing magnetic skyrmion bubbles. *Science*, 349(6245):283–286, 2015.
- [44] Wanjun et. al. Jiang. Direct observation of the skyrmion hall effect. *Nature Physics*, 13(2):162–169, Feb 2017.
- [45] B. Abeles, Roger W. Cohen, and W. R. Stowell. Critical magnetic fields of granular superconductors. *Phys. Rev. Lett.*, 18:902–905, May 1967.
- [46] B. Abeles, Roger W. Cohen, and G. W. Cullen. Enhancement of superconductivity in metal films. *Phys. Rev. Lett.*, 17:632–634, Sep 1966.
- [47] G. Deutscher and S. A. Dodds. Critical-field anisotropy and fluctuation conductivity in granular aluminum films. *Phys. Rev. B*, 16:3936–3942, Nov 1977.
- [48] M. David Maloney, Francisco de la Cruz, and Manuel Cardona. Superconducting parameters and size effects of aluminum films and foils. *Phys. Rev. B*, 5:3558–3572, May 1972.
- [49] Roger W. Cohen and B. Abeles. Superconductivity in granular aluminum films. *Phys. Rev.*, 168:444–450, Apr 1968.
- [50] B. Abeles. Effect of charging energy on superconductivity in granular metal films. *Phys. Rev. B*, 15:2828–2829, Mar 1977.
- [51] Aviv Glezer Moshe, Eli Farber, and Guy Deutscher. Optical conductivity of granular aluminum films near the mott metal-to-insulator transition. *Phys. Rev. B*, 99:224503, Jun 2019.

- [52] T. Worthington, P. Lindenfeld, and G. Deutscher. Heat-capacity measurements of the critical coupling between aluminum grains. *Phys. Rev. Lett.*, 41:316–319, Jul 1978.
- [53] Nicholas A. Masluk, Ioan M. Pop, Archana Kamal, Zlatko K. Minev, and Michel H. Devoret. Microwave characterization of josephson junction arrays: Implementing a low loss superinductance. *Phys. Rev. Lett.*, 109:137002, Sep 2012.
- [54] Lukas Grünhaupt, Martin Spiecker, Daria Gusenkova, Nataliya Maleeva, Sebastian T. Skacel, Ivan Takmakov, Francesco Valenti, Patrick Winkel, Hannes Rotzinger, Wolfgang Wernsdorfer, Alexey V. Ustinov, and Ioan M. Pop. Granular aluminium as a superconducting material for high-impedance quantum circuits. *Nature Materials*, 18(8):816–819, Aug 2019.
- [55] S. H. Pan, E. W. Hudson, and J. C. Davis. Vacuum tunneling of superconducting quasiparticles from atomically sharp scanning tunneling microscope tips. *Applied Physics Letters*, 73(20):2992–2994, 11 1998.
- [56] C. Kittel and U.C. Kittel. *Introduction to Solid State Physics*. Play; 9. Wiley, 1986.
- [57] J. R. Kirtley, S. I. Raider, R. M. Feenstra, and A. P. Fein. Spatial variation of the observed energy gap in granular superconducting NbN films. *Applied Physics Letters*, 50(22):1607–1609, 06 1987.
- [58] T. Nishizaki, A.M. Troyanovski, G.J.C. van Baarle, P.H. Kes, and J. Aarts. Stm imaging of vortex structures in nbn thin films. *Physica C: Superconductivity*, 388-389:777–778, 2003. Proceedings of the 23rd International Conference on Low Temperature Physics (LT23).
- [59] G. E. Blonder, M. Tinkham, and T. M. Klapwijk. Transition from metallic to tunneling regimes in superconducting microconstrictions: Excess current, charge imbalance, and supercurrent conversion. *Phys. Rev. B*, 25:4515–4532, Apr 1982.
- [60] Ulrich Eckern and Albert Schmid. Quantum vortex dynamics in granular superconducting films. *Phys. Rev. B*, 39:6441–6454, Apr 1989.
- [61] Dmitry E. Kiselov, Mikhail A. Skvortsov, and Mikhail V. Feigel'man. Gapful electrons in a vortex core in granular superconductors. *SciPost Phys.*, 15:008, 2023.
- [62] C. Caroli, P.G. De Gennes, and J. Matricon. Bound fermion states on a vortex line in a type ii superconductor. *Physics Letters*, 9(4):307–309, 1964.
- [63] K. S. Denisov, I. V. Rozhansky, N. S. Averkiev, and E. Lähderanta. General theory of the topological hall effect in systems with chiral spin textures. *Phys. Rev. B*, 98:195439, Nov 2018.
- [64] Naoto Nagaosa, Jairo Sinova, Shigeki Onoda, A. H. MacDonald, and N. P. Ong. Anomalous hall effect. *Rev. Mod. Phys.*, 82:1539–1592, May 2010.

- [65] Lixuan Tai, Bingqian Dai, Jie Li, Hanshen Huang, Su Kong Chong, Kin L. Wong, Huairuo Zhang, Peng Zhang, Peng Deng, Christopher Eckberg, Gang Qiu, Haoran He, Di Wu, Shijie Xu, Albert Davydov, Ruqian Wu, and Kang L. Wang. Distinguishing the two-component anomalous hall effect from the topological hall effect. *ACS Nano*, 16(10):17336–17346, 2022. PMID: 36126321.
- [66] Mengying Bian, Aleksandr N. Kamenskii, Mengjiao Han, Wenjie Li, Sichen Wei, Xuezheng Tian, David B. Eason, Fan Sun, Keke He, Haolei Hui, Fei Yao, Renat Sabirianov, Jonathan P. Bird, Chunlei Yang, Jianwei Miao, Junhao Lin, Scott A. Crooker, Yanglong Hou, and Hao Zeng. Covalent 2d cr₂te₃ ferromagnet. *Materials Research Letters*, 9(5):205–212, 2021.
- [67] D. M. Burn, L. B. Duffy, R. Fujita, S. L. Zhang, A. I. Figueroa, J. Herrero-Martin, G. van der Laan, and T. Hesjedal. Cr₂te₃ thin films for integration in magnetic topological insulator heterostructures. *Scientific Reports*, 9(1):10793, Jul 2019.
- [68] S. J. Youn, S. K. Kwon, and B. I. Min. Correlation effect and magnetic moments in Cr₂Te₃. *Journal of Applied Physics*, 101(9), 05 2007.
- [69] Xiaoqian Zhang, Siddhesh C. Ambhire, Qiangsheng Lu, Wei Niu, Jacob Cook, Jidong Samuel Jiang, Deshun Hong, Laith Alahmed, Liang He, Rong Zhang, Yongbing Xu, Steven S.-L. Zhang, Peng Li, and Guang Bian. Giant topological hall effect in van der waals heterostructures of crte₂/bi₂te₃. *ACS Nano*, 15(10):15710–15719, 2021. PMID: 34460216.