Understanding Topological Phases of Matter with Statistical Methods

Dissertation

zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich Physik
der Johann Wolfgang Goethe-Universität
in Frankfurt am Main

von
Thomas Mertz
aus Bad Homburg

Frankfurt am Main, August 2021
(D30)

vom Fachbereich Physik der
Johann Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan:	Prof. Dr. Harald Appelshäuser
Gutachter:	Prof. Dr. Roser Valentí
	Prof. Dr. Falko Pientka
Datum der Disputation:	07.02.2022

"Prediction is very difficult, especially if it's about the future."—Niels Bohr

Zusammenfassung

In dieser Arbeit diskutieren wir die Anwendung statistischer Methoden zur Datenanalyse auf Probleme aus der Festkörpertheorie, für die exakte Lösungen größtenteils nicht verfügbar sind. Wir interessieren uns in diesem Kontext speziell für sogenannte topologische Phasen.

Die Theorie der Festkörper beschreibt in erster Linie Strukturen und daraus hervorgehende Phasen, die periodisch in ihren räumlichen Freiheitsgraden sind. Diese sogenannten festen Phasen der Materie bestehen aus Atomen, die sich in Form periodischer Gitter angeordnet haben und dadurch eine diskrete Translationssymmetrie aufweisen. Dabei sind die Atomkerne aufgrund ihrer höheren Masse und der damit einhergehenden Trägheit vergleichsweise unbeweglich, sodass viele der dynamischen Eigenschaften des Materials vor allem durch das Verhalten der Elektronen beschrieben werden können, siehe z.B. Ladungstransport. Man macht sich dies zunutze, da das Gesamtsystem bestehend aus Atomkernen und Elektronen durch die vielzähligen gegenseitigen Coulomb Wechselwirkungen sehr schwierig zu beschreiben ist, und nutzt die stark unterschiedlichen Zeitskalen, die der Dynamik der beiden Systeme zugrunde liegen, um beide voneinander zu entkoppeln. In theoretischen Modellen nimmt man dann an, dass die Elektronen sich in einem zeitlich konstanten, räumlich periodischen Potential bewegen, das sich aus den mittleren Positionen der Atomkerne ergibt. Das ursprüngliche Problem hat sich damit auf ein Vielteilchensystem von Elektronen reduziert, deren Freiheitsgrade durch ihre kinetische Energie, das periodische Gitterpotential und die Coulomb-Abstoßung zwischen Elektronen beeinflusst werden. In einer ersten Näherung beschreibt man das System unter Vernachlässigung der Wechselwirkung zwischen Elektronen und Berücksichtigung lediglich der kinetischen Energie und des Gitterpotentials, was letztendlich auf entkoppelte Gleichungen für einzelne Elektronen führt. Das Gesamtproblem reduziert sich also auf N Kopien desselben Einteilchenproblems, wobei N die Anzahl der Elektronen im Festkörper ist, das in der Praxis einfach gelöst werden kann. Diese sehr drastische Näherung hat zur Folge, dass viele Effekte, die in der Realität beobachtet werden können, nicht akkurat beschrieben werden, sodass in der Praxis eine wie auch immer geartete Beschreibung elektronischer Korrelationen meist nötig wird. Unglücklicherweise stellt die Komplexität des wechselwirkenden Problems eine große Hürde dar. Während das nichtwechselwirkende Problem sich innerhalb eines Hilbertraums der Dimension d beschreiben lässt, wobei d die Anzahl der Einteilchen-Freiheitsgrade der Elektronen darstellt, bläht sich die Dimension des korrespondierenden Vielteilchensystems exponentiell auf und man erhält stattdessen d^N Dimensionen. Für die theoretische Beschreibung stellt dies schnell eine nicht zu überwindende Hürde dar, wodurch man sich üblicherweise mit approximativen Methoden behelfen muss. Diese approximativen Methoden zur Beschreibung von Vielteilchensystemen bilden gleichzeitig ein großes Feld in der modernen Festkörpertheorie, worin man kontinuierlich versucht, die genäherten Lösungen weiter an experimentelle Beobachtungen anzugleichen und die Menge an lösbaren Problemen zu vergrößern, um schlussendlich bekannte Effekte genauer zu verstehen oder Vorhersagen für interessante Eigenschaften zu machen.

Eine dieser interessanten Eigenschaften bildet die Grundlage eines stark wachsenden Teilgebiets der Festkörperphysik, in dem man sich mit den sogenannten topologischen Phasen

beschäftigt. Ihren Ursprung hatten die dort getätigten Überlegungen in der Entdeckung des Quanten Hall Effekts durch K. von Klitzing im Jahre 1980, durch den man eine Verbindung zwischen der mathematischen Disziplin Topologie und der Physik kondensierter Materie beobachten konnte. Besagte topologische Phasen zeichnen sich in erster Linie durch eine immense Robustheit gegenüber schwachen Änderungen an den Rahmenbedingungen des Systems aus, da ihre topologischen Eigenschaften im Grunde nur von sehr allgemeinen Größen abhängen. Dies lässt sich gut durch makroskopische Beispiele illustrieren. Das Paradebeispiel wäre z.B. der Donut, dessen Anzahl von Löchern meist gleich, nämlich genau eins, ist (abgesehen von Berlinern). Obwohl sich verschiedene Donuts äußerlich oft drastisch unterscheiden, lassen sie sich doch anhand der Anzahl ihrer Löcher kategorisieren. Diese Eigenschaft ist zudem robust, da sie sich nicht ändern kann ohne den Donut selbst zu zerstören. Ein weiteres Beispiel ist eine Schnur, deren beide Enden man miteinander verklebt. Die erhaltene Schlaufe lässt sich in der Ebene beliebig verformen, die Anzahl der Löcher durch die Schlaufe, wieder exakt eins, bleibt dabei jedoch allzeit erhalten. Sollten wir beabsichtigen, diese Anzahl zu ändern, so bleibt als einzige Möglichkeit, die Schnur aufzuschneiden und die Enden auf andere Weise zusammenzufügen, um etwa eine "8" zu formen (Überlappungen sind in der Ebene nicht möglich). Diese Vorgehensweise ist allerdings, genau wie ein Zerreißen oder Zerdrücken des Donuts, destruktiv und repräsentiert drastische Änderungen der grundlegenden Parameter des Systems. Unter normalen Bedingungen, d.h. ohne Einsatz von Scheren oder roher Gewalt, lässt sich die Anzahl der Löcher bei beiden Objekten also nicht verändern.

In der Festkörpertheorie wird die Rolle der Schnur aus unserem Beispiel von der Wellenfunktion der Elektronen eingenommen, während die Schere, die den einzigen Weg, die topologischen Eigenschaften des Systems zu verändern, darstellt, mit der Entartung von Zuständen zusammenhängt. Bei niedrigen Temperaturen (wir gehen hier von Temperaturen nahe des absoluten Nullpunktes aus) werden die elektronischen Zustände anhand ihrer Energie von unten beginnend mit vorhandenen Elektronen aufgefüllt. Die Energie des höchsten besetzten Energieniveaus entspricht dann der sogenannten Fermienergie, die eine besondere Bedeutung für die metallischen Eigenschaften des Systems hat. So sind Materialien, deren Fermienergie inmitten eines kontinuierlichen Energiebandes liegt, Metalle. Dahingegen werden solche Materialien, deren Fermienergie innerhalb einer Energielücke liegt, als Isolatoren bezeichnet, da Ladungstransport nur dann erfolgen kann wenn entsprechende Zustände zur Verfügung stehen, um das System aus seinem Gleichgewichtszustand heraus anzuregen. Ändert man nun behutsam die Eigenschaften eines Materials, beispielsweise die Position der Atome, so ändert sich auch das Anregungsspektrum. Genau wie bei der Schnur lassen sich jedoch Systeme miteinander in Beziehung setzen, die durch solche behutsamen Verformungen ineinander überführt werden können. Der gegenteilige Fall liegt vor, wenn, um eine solche Überführung herzustellen, eine Entartung von mehreren Zuständen bei der Fermienergie zwingend auftreten muss. Diese Zustände, die man nur durch den buchstäblichen Einsatz der Schere ineinander überführen kann, heißen topologisch distinkt oder inäquivalent. Hierbei definiert man auch den trivialen Fall, also Systeme ohne interessante topologische Eigenschaften, durch den atomaren Limes, in dem die Atome weit voneinander entfernt unabhängig daher vegetieren und die Elektronen deshalb an eines der Atome gebunden sind. Nichttriviale Phasen liegen demnach nur vor wenn die Atome Bindungen miteinander eingehen und Elektronen im Kristall delokalisiert sind.

Die interessanten topologischen Eigenschaften, wegen denen man diese Analyse letztendlich überhaupt auf sich nimmt, sind sogenannte topologisch geschützte Zustände, also Zustände, deren Existenz durch die Topologie des Systems garantiert wird, und die sich daher nicht so einfach stören lassen. Betrachten wir eine Probe endlicher Größe, so stellt der Rand eine Schnittstelle zum Vakuum dar, welches die einfachste Realisierung der trivialen Phase repräsentiert. Geht man nun davon aus, dass in der Probe eine nichttriviale topologische Phase vorliegt,

so muss an der Schnittstelle zur trivialen Phase eine Entartung des Energiespektrums auftreten, da sich die topologischen Eigenschaften nur so zu ändern vermögen. Als direkte Konsequenz daraus ergibt sich die Existenz von leitenden Zuständen entlang der Oberfläche der Probe, die nur von der inneren Struktur abhängen und unabhñgig von der Form der Oberfläche sind. Durch diese Garantie für die Existenz dieser Zustände spricht man von geschützten Zuständen, da die leitenden Randzustände von den topologischen Eigenschaften im Inneren der Probe beschützt werden. Eine solche Eigenschaft verspricht offenbar besonders robuste Leitungseigenschaften, welche tatsächlich durch den Quanten Hall Effekt bestätigt wurden, womit man eine sehr präzise Möglichkeit fand, Physikalische Konstanten zu bestimmen. Zudem verspricht diese Robustheit weitreichende Anwendungen im Zusammenhang mit Quantencomputing, wo man auf die Stabilität der mikroskopischen Qubit Zustände angewiesen ist, um die Integrität und damit Zuverlässigkeit der Rechnung sicherzustellen. Hierzu benötigt man insbesondere Zustände, die nur sehr schwach an ihre Umgebung gekoppelt sind, damit sie ihre Quanteneigenschaften wie Verschränkung für möglichst lange Zeit behalten. Ähnlich der Anzahl der Löcher eines Donuts können die topologischen Eigenschaften eines Materials durch eine Zahl ausgedrückt werden, die für alle topologisch äquivalenten Systeme den gleichen Wert hat, weshalb man sie eine topologische Invariante nennt. Im Falle des ganzzahligen Quanten Hall Effekts ist diese Zahl die sogenannte Chern Zahl, die ganzzahlige Werte annimmt, und mit der wir uns im Verlauf dieser Arbeit intensiv beschäftigen werden.

Im Jahre 2005 wurde mit dem sogenannten Quanten Spin-Hall Effekt eine weitere topologische Phase vorhergesagt, die sich im Gegensatz zum normalen Quanten Hall Effekt durch eine Erhaltung der Zeitumkehrsymmetrie auszeichnet. Bereits im Jahre 2007 wurde diese Phase experimentell bestätigt. Anders als bei der Chern Zahl kann die mit dieser Phase in Zusammenhang stehende Invariante nur zwei mögliche Werte annehmen (stellvertretend Null und Eins), was automatisch bedeutet, dass alle nichttrivialen Phasen zueinander äquivalent sind. Durch diesen großen Unterschied wird sofort klar, dass Symmetrien eine prominente Rolle im Kontext topologischer Eigenschaften einnehmen. In jüngster Vergangenheit wurde dem Rechnung getragen und man versuchte, auch räumliche Symmetrien in die Beschreibung einzubeziehen, wodurch es gelang, tabellarisch alle 230 Raumgruppen dahingehend zu unterscheiden, ob Entartungen, die die Existenz nichttrivialer Phasen erlauben, für die jeweilige Symmetrie auftreten können. Nichtsdestotrotz sind noch immer aufwändige Rechnungen nötig, um diese Existenz für einen speziellen Fall nachzuweisen, da hierfür stets eine topologische Invariante zu berechnen ist.

Da die Berücksichtigung der gegenseitigen Wechselwirkungen der Elektronen untereinander eine komplett andere Herangehensweise erfordert, um der gesteigerten Anzahl an Freiheitsgrade Rechnung zu trangen, wurden viele Aspekte zunächst nur für nicht-wechselwirkende Systeme behandelt. Eine Beschreibung durch die üblichen Größen, die bei der Behandlung wechselwirkender Systeme zur Anwendung kommen, ist deshalb oft komplizierter. Besonders relevant ist in diesem Zusammenhang die Methode der Greensfunktionen, die die Eigenschaften des Systems in Form einer Menge aus Funktionen von Frequenz bzw. Energie und Impuls ausdrückt. Während die einfachste dieser Funktionen leichter handhabbar ist als die Vielteilchen-Wellenfunktion, so enthält sie doch nur eine reduzierte Menge an Informationen, die sich auf die Beschreibung von Einteilchen-Anregungen beschränken, was allerdings bereits die erwähnten Randzustände einschließt. Zur Lösung des Vielteilchenproblems existiert eine Vielzahl an approximativen Methoden, die von der Methode der Greensfunktionen Gebrauch machen, z.B. die dynamische Molekularfeldtheorie (DMFT), Cluster-Störungstheorie (CPT) oder die Zweiteilchen selbstkonsistente Methode (TPSC). Unter der Voraussetzung, dass die Wechselwirkungen nicht zu stark sind, kann man die Anregungen des Vielteilchensystems effektiv durch ein Einteilchenproblem beschreiben, wobei man dann statt der ursprünglichen Elektronen von Quasiteilchen spricht. In ähnlicher Weise lassen sich auch die topologischen Eigenschaften durch ein solches effektives Ein-

teilchenmodell beschreiben, jedoch mit dem entscheidenden Unterschied, dass die Abbildung in diesem Zusammenhang exakt ist. Für dieses effektive Modell, das nur die Topologie des Systems korrekt wiedergibt, hat sich der Begriff des topologischen Hamiltonians durchgesetzt, aus dem sich die Chern Zahl genau wie im nichtwechselwirkenden Fall berechnen lässt. Neben dieser Beschreibung gibt es auch Formulierungen der Hall Leitfähigkeit, die direkt proportional zur Chern Zahl ist, in Abhängigkeit von der Einteilchen Greensfunktion. Diese ist jedoch aufgrund der Notwendigkeit, über alle Frequenzen zu summieren, numerisch deutlich anspruchsvoller zu berechnen.

Da für die Greensfunktion im wechselwirkenden Fall allgemein keine exakte Lösung bekannt ist, übermannt uns gewissermaßen die Qual der Wahl, welche aus der Vielzahl an verschiedenen näherungsweisen Methoden am besten zu nutzen ist. Eine besonders erfolgreiche Methode ist die bereits erwähnte DMFT, deren Erfolgsgeschichte in erster Linie auf der Beschreibung des wechselwirkungsgetriebenen Metall-Mott-Isolator Phasenübergangs begründet ist. Auch im Kontext der topologischen Phasen erfreut sich DMFT großer Beliebtheit und so sind in der Literatur entsprechende Phasendiagramme für eine Vielzahl an Modellen zu finden. Die hauptsächliche Näherung, die DMFT zugrunde liegt, manifestiert sich darin, dass die sogenannte Selbstenergie, die die frequenz- und impulsabhängigen Korrekturen zum Einteilchenspektrum enthält, ihre Impulsabhängigkeit verliert, streng genommen also lokal ist. Die Chern Zahl ist jedoch tatsächlich ein direktes Maß für eine bestimmte Impulsabhängigkeit der Einteilchenzustände, die im Rahmen des topologischen Hamiltonians durch die Selbstenergie korrigiert werden. Die Vernachlässigung nicht-lokaler Korrekturen wirft daher große Fragen über die Vertrauenswürdigkeit der Näherung und die damit erzeugten Phasendiagramme auf. Wir beschäftigen uns mit diesen Fragen, indem wir ein allgemeines Modell mit DMFT selbst, und zugleich anderen Methoden, die ihrerseits nicht besagter lokalen Näherung unterliegen, behandeln und vergleichende Analysen anstellen. Hierzu erarbeiten wir Maße für die Stärke und Relevanz der Impulsabhängigkeit der Selbstenergie, die wir als die Dispersionsamplitude der Selbstenergie bezeichnen, und mit Hilfe derer wir eine Art Phasendiagramm für das ionische Hubbard Modell auf dem Quadratgitter berechnen. Dieses Modell zeichnet sich insbesondere durch zwei wichtige Eigenschaften aus: Es beinhaltet einen sogenannten trivialen Massenterm, der im Prinzip das ionische Potential beschreibt und zugleich in allen üblichen topologischen Modellen vorkommt, und eine antiferromagnetische Instabilität, die sich durch eine bevorzugt starke Impulsabhängigkeit der Selbstenergie äußert. Letzteres bedeutet, dass das Modell als eine obere Schranke für die Relevanz der Impulsabhängigen Korrekturen dienen kann. In unserem Phasendiagramm unterscheiden wir nun zwei konträre Phasen, die sich jeweils durch starke oder schwache Relevanz der Impulsabhängigkeit auszeichnen und durch die Dispersionsamplitude der Selbstenergie identifiziert werden können. Wir erhalten so das wichtige Ergebnis, dass nichtlokale Effekte größtenteils nur sehr schwach sind und daher in einem großen Teil des Phasendiagramms keiner genaueren Beachtung bedürfen, was automatisch zur Folge hat, dass DMFT paradoxerweise auch die Chern Zahl topologischer Phasen in diesem Bereich sehr akkurat wiedergibt. Wir zeigen außerdem durch Anwendung einer einfachen Analyse der Energieskalen verschiedener Freiheitsgrade des Systems, dass der erwartete Fehler der DMFT Näherung für diese Klasse von Modellen bereits durch das Ergebnis von DMFT selbst abgeschätzt werden kann, im Allgemeinen also keiner Vergleichsrechnung bedarf.

Die Präzision der DMFT Lösung im Bereich kleiner Impulsabhängigkeit fußt grundlegend auf der Annahme, dass kleine Änderungen an den Parametern die topologischen Eigenschaften eines Systems nicht verändern, was wir am speziellen Fall des Haldane Modells explizit demonstrieren. Da jedoch zu erwarten ist, dass nicht nur die Stärke, sondern auch die Art der Impulsabhängigkeit eine Rolle für die topologische Klassifizierung spielt, verlangt die definitive Beantwortung unserer Fragestellung nach einer exakten Lösung, da jede andere Näherungsmethode lediglich Indizien zu liefern vermag, allerdings keine allgemeingültige Schlussfolgerung

zulässt.

An dieser Stelle entscheiden wir uns dafür, die üblichen Methoden hinter uns zu lassen und stattdessen ein komplett anderes Konzept zu erarbeiten. Die exponentielle Komplexität des Vieltelchenproblems bedingt, dass die exakte Lösung in unerreichbarer Ferne verbleibt. Statt auf eine bestimmte Näherungsmethode zu setzen, die die Impulsabhängigkeit auf eine bestimmte Art beinhaltet, schlagen wir eine Methode im Sinne eines stochastischen Algorithmus vor, die im Limes langer Laufzeiten schlussendlich alle möglichen Lösungen evaluiert. Diese Idee weist in ihrer Essenz Ähnlichkeiten zum sogenannten Solovay-Strassen (oder Miller-Rabin) Test auf, der eine Zahl auf ihre Primeigenschaft hin überprüft, und anstatt exponentiell viele Kombinationen von Produkten zu berechnen bis entweder eine Faktorisierung gefunden wurde oder gezeigt ist, dass keine existiert, verschiedene Kombinationen zufällig testet und dadurch eine wahrscheinlichkeitsbehaftete Lösung erhält, die jedoch im Limes langer Laufzeiten gegen die exakte Lösung konvergiert. Im Zusammenhang mit Systemen von Festkörpern wenden wir ein ähnliches Schema an, indem wir zufällige Selbstenergie-Funktionen erzeugen und damit die Robustheit der Chern Zahl bezüglich nicht-lokaler Korrekturen prüfen. Das Ergebnis, das wir im Rahmen dieser Analyse erhalten, lässt sich in Form einer Änderungswahrscheinlichkeit der Chern Zahl darstellen, die wir anschließend genauer untersuchen und dabei eine exponentielle Unterdrückung der Empfindlichkeit bzw. einen exponentiellen Anstieg der Robustheit, als Funktion der Distanz zum lokalen Phasenübergang, also jenem Phasenübergang, der sich ohne jegliche nicht-lokale Korrektur ergibt, feststellen. Zieht man für den lokalen Phasenübergang die DMFT Lösung heran, so ergibt sich sofort, dass die Wahrscheinlichkeit für große Fehler der DMFT Lösung unwahrscheinlich ist und eine Abschätzung des maximal zu erwartenden Fehlers lässt sich als Funktion der Dispersionsamplitude der Selbstenergie ausdrücken.

Der Erfolg der statistischen Herangehensweise dient im Folgenden als Motivation dafür, uns auch anderen Problemen mit derselben grundlegenden Idee anzunehmen. Wir entscheiden uns im Speziellen für die Untersuchung von Methoden zur Vorhersage von topologischen Materialien. In diesem Zusammenhang beschränken wir uns auf nicht-wechselwirkende Modelle. Zum einen aus Komplexitätsgründen, aber auch, weil durch den topologischen Hamiltonian und unsere vorherige Analyse eine Beschreibung wechselwirkender Systeme zumindest im Grundsatz analog vorgenommen werden kann. Ein Algorithmus, der ohne jegliche Information von außen vorgegeben zu bekommen, präzise Vorschläge für Materialien mit topologischen Eigenschaften macht, wird höchstwahrscheinlich Wunschdenken bleiben, allerdings existiert bereits ein großer Wissensschatz als Folge theoretischer und experimenteller Arbeiten. Das Problem, das sich hierbei offenbart ist, dass topologische Eigenschaften nicht in jeder dieser Arbeiten eine Rolle gespielt haben und daher entsprechende Aussagen nicht unbedingt getroffen werden können. Zudem verteilt sich das gesammelte Wissen über viele Teilgebiete der Festkörperphysik, in denen nicht alle Wissenschaftler dieselbe mathematische Sprache sprechen. Um an diesem Punkt eine Annäherung verschiedener Experten zu bewirken, entscheiden wir uns dazu, hier anzusetzen, indem wir uns vornehmen, das Verständnis der Zusammenhänge zwischen topologischen Zuständen auf der einen und der Kristallstruktur auf der anderen Seite zu erweitern. Die Kristallstruktur manifestiert sich in unserem Falle in den Matrixelementen des Hamiltonoperators, die sich als Überlappintegrale der elektronischen Wellenfunktionen ergeben. Diese wohlbekannte Formulierung verspricht eine große Teilhabe verschiedenster Experten und dadurch die erfolgreiche Zusammenführung verschiedener Wissensbereiche.

Wir beschäftigen uns in diesem Kontext auch mit der Frage, inwiefern übliche Verfahren des maschinellen Lernens für unsere Zwecke eingesetzt werden können, indem wir genau analysieren, welche Art Information wir aus Phasendiagrammen ableiten können, wobei wir speziell die Skalierung hin zu vielen Dimensionen im Auge behalten. Allerdings stellt sich schnell heraus, dass viele dieser Methoden sich entsprechende Informationen nur mit großer Mühe entlocken

lassen, was in etwa mit der Komplexität des ursprünglichen Datensatzes in Verbindung gebracht werden kann. Stattdessen zeigen wir, dass ein entsprechender Datensatz auch direkt und ohne Umschweife über künstliche Intelligenz mit Hilfe informationstheoretischer Überlegungen analysiert werden kann, wodurch sich Einblicke in die Relevanz bestimmter räumlicher Freiheitsgrade für die topologische Klassifizierung ergeben. Die Erzeugung der Daten übernimmt ein zufallsbasierter Algorithmus, der im Prinzip einen Teil des hochdimensionalen Phasendiagramms, das ohne Zuhilfenahme entsprechender Methoden nicht verstanden werden kann, abbildet. Damit gelingt es uns schließlich zu zeigen, dass das Haldane Modell und seine topologischen Zustände als Prototyp eines topologischen Modells auf dem Bienenwabengitter aus einem solchen Datensatz hervorgehen, und das ohne jegliche Information abseits der Chern Zahl von außen vorzugeben. Zusätzlich finden wir weitere Zustände, die nicht durch das Haldane Modell abgedeckt werden. Wir wenden die Methode zusätzlich auf das Kagome Gitter an, wo wir, ausgehend von einer gänzlich unvoreingenommenen Position, selbst komplexe Zusammenhänge zwischen der Chern Zahl und den Parametern des Hamiltonoperators, und damit der Kristallstruktur, entdecken können. Diese Analyse endet schließlich mit der Präsentation eines qualitativen Phasendiagramms, das Baupläne für verschiedene topologische Phasen beinhaltet.

Die Struktur dieser Arbeit ist folgendermaßen aufgebaut:

Zunächst geben wir im Anschluss an die Englische Fassung dieser Einleitung eine kurze Einführung in die theoretischen Hintergründe der topologischen Phasen in Kapitel 2, wo wir uns auf die für die später folgende Diskussion relevanten Aspekte beschränken und keinen Anspruch auf Vollständigkeit stellen. Mehr Informationen hierzu finden sich stattdessen in der entsprechenden Fachliteratur, auf die wir bemüht sind hinzuweisen.

In Kapitel 3 beschäftigen wir uns dann mit ausgewählten Methoden der Festkörpertheorie, wobei wir den für diese Arbeit sehr wichtigen nichtwechselwirkenden Fall genauer beleuchten und schließlich einige relevante approximative Methoden zur Lösung des Vielteilchenproblems ansprechen, die für uns relevant sind.

Da wir uns relativ ausgiebig mit Statistik beschäftigen geben wir auch einen kurzen Überblick über die entsprechende Notation und einige fundamentale Konzepte, von denen wir später Gebrauch machen in Kapitel 4. Hier gehen wir auch getrennt auf wichtige Konzepte aus der Informationstheorie und der künstlichen Intelligenz ein.

Der eigenständige Teil folgt auf diese einführenden Kapitel und ist selbst in drei Teile aufgeteilt, die wir anhand von drei in sich geschlossenen Projekten, die aber aufeinander aufbauen, ausrichten. Leser, die mit den Themen der Kapitel 2-4 bereits vertraut sind, seien hiermit ermutigt, gleich zu Kapitel 5 zu springen. Dort präsentieren wir unsere Studie der Impulsabhängigkeit der Selbstenergie am Beispiel des ionischen Hubbard Modells, wobei wir verschiedene numerische Methoden miteinander vergleichen und unter Zuhilfenahme exakter Grenzwerte eine effektive Beschreibung des Phasendiagramms erhalten. Diese Diskussion beschränkt sich größtenteils auf das Quadratgitter, das als obere Schranke für die Impulsabhängigkeit verstanden werden kann, allerdings gehen wir auch kurz auf Ergebnisse für das Dreiecksgitter ein, anhand derer wir tatsächlich eine schwächere Impulsabhängigkeit feststellen.

Unseren statistischen Ansatz für die Analyse der zu erwartenden Fehler der DMFT Methode beschreiben wir in Kapitel 6, wo wir eine Zerlegung der Selbstenergie in lokale und nichtlokale Anteile vollziehen und deren Auswirkungen auf die Chern Zahl separat diskutieren. Der Diskussion des nicht-lokalen Teils schenken wir besondere Aufmerkssamkeit und vertiefen die Ergebnisse aus unserer entsprechenden Publikation durch eine quantitative Analyse noch weiter. Wir kommentieren außerdem die Bedeutung unserer Ergebnisse für das Phasendiagramm

des Haldane-Hubbard Modells, dessen Erscheinung maßgeblich von der gewählten numerischen Methode abhängt.

In Kapitel 7 widmen wir uns schließlich der Entdeckung neuer Materialien mit topologischen Eigenschaften, wobei wir zunächst die Fragestellung hin zu dem Verständnis hochdimensionaler Phasendiagramme konkretisieren und die zu erwartende Form des Ergebnisses genauer beleuchten. Im Zuge dessen untersuchen wir die Eignung einiger Standardmethoden aus dem Bereich des maschinellen Lernens für die Beschreibung topologischer Phasendiagramme. Anschließend stellen wir unsere statistische Herangehensweise vor und zeigen wie informationstheoretische Methoden das Verständnis solcher komplexer Datensätze ermöglichen können, indem wir aus einem zufällig generierten Datensatz die Essenz des Haldane Modells, welches hier als Testumgebung dient, extrahieren. Wir wenden die gleiche Methode danach auf das Kagome Gitter an, wobei wir einige Ansätze weiter verfeinern und dadurch sehr komplexe Zusammenhänge innerhalb der Daten zugänglich machen. Schließlich sprechen wir kurz mögliche Wege hin zur Anwendung auf realistische Materialien an.

Contents

Chapter 1

Introduction

In this thesis, we discuss the merits of statistical methods in condensed matter theory—in particular, in the field of topological phases—where exact solutions are virtually inaccessible and even approximate methods come at high computational cost.

Condensed matter theory describes phases of matter that are predominantly periodic in space. In these solid phases, the atoms that form the material are arranged on a periodic lattice, which gives rise to a discrete translational symmetry. Since the atomic nuclei are relatively immobile compared to the electrons, the dynamical properties of the material are dominated by the behavior of electrons. A description of the whole system is extremely difficult, however, due to the much larger mass of the atomic nuclei compared to that of the electrons, the time scales are vastly different, and therefore, the two systems can be decoupled from each other. One then arrives at a model where electrons are moving through a static lattice that can be thought of as the equilibrium configuration. What remains are only electronic degrees of freedom that consist of a kinetic energy, the periodic lattice potential and the electron-electron Coulomb repulsion between all N electrons. Taking into account only the former two contributions, one arrives at essentially N copies of the same single electron model, which can readily be solved. On the other hand, this rather drastic approximation is unable to capture many effects that are observed in reality, which means that the interactions between individual electrons often have to be accounted for in one way or another. Unfortunately, while the single particle model is defined in a Hilbert space of dimension d, where d is the number of single particle degrees of freedom, the dimension of the corresponding many-body Hilbert space scales exponentially in the number of electrons, i.e., d^N. Clearly, this becomes intractable very quickly and suitable approximation methods are required. A large portion of modern condensed matter theory is devoted to the conception and improvement of such approximation methods, hoping for accurate descriptions of experimental results, an understanding of the underlying physical mechanisms, and finally, the power to predict systems with desirable properties.

One such interesting property is studied in the field of topological phases that has enjoyed much attention in the last few decades. Originally formulated in the aftermath of the groundbreaking discovery of the quantum Hall effect by K. von Klitzing in 1980, topology in physics corresponds to the application of methods from the mathematical field of topology to condensed matter systems. In particular, we are interested in properties that are rather robust and only depend on the presence or absence of certain fundamental features of the system. A common example in the macroscopic world is, e.g., the number of holes through a doughnut. This number does not change when the doughnut is squished together or stretched out, unless it gets torn apart. Consider for example taking a string and attaching the two ends to each other. This creates a ring of some sort, however, if the string is flexible it can assume an abundance of different shapes. What all of these have in common, though, is the existence of exactly one hole,

provided that we stay in a two-dimensional plane. Changing the number of holes would require us to cut the string somewhere and possibly use glue to attach the ends to the string in another way. Both operations are considered drastic since they require us to destroy the string just as in the doughnut example, and therefore, we can conclude that the number of holes is going to be constant as long as we keep scissors and brute force out of the game. Another example are knots, which are basically an extension of the former to three dimensions. Allowing perturbations also in three dimensions, we can introduce knots that can be differentiated in terms of how simple it is to resolve them. A simple loop can be resolved by just pulling on the string. On the other hand, more complicated structures are possible that would require us once again to cut the string, since in order to restore the simple loop we would have to move the string through itself.

In condensed matter physics, the proverbial string are the electronic wave functions of the material, while the scissors, i.e., the action that can change the fundamental properties of the system, turns out to correspond to the introduction of degeneracies between different energy bands. Here, the term "band" refers to one of the single particle quantum numbers. At low temperature (approximately zero), the electronic states are filled from the bottom up, i.e., starting with the smallest energies, until all electrons have an assigned set of quantum numbers. The highest occupied energy is the so-called Fermi energy. If this energy is at the bottom end of a gap in the energy spectrum, there is a general lack of free states that are close in energy, which indicates insulating behavior, since current flow would require the occupation of excited states in order to break out of equilibrium. Changing properties of the material smoothly, such as moving the atomic positions, corresponds to deformations of the energy spectrum. In analogy to the knot, there are sets of configurations that can be related to one another by such smooth variations. However, there are special states that cannot be reached through variations of a normal initial state without closing the energy gap somewhere along the way. Two such states that require the proverbial scissors along any transformation path are called topologically distinct. The trivial case is defined through the atomic limit, where electrons are strongly localized to their individual atoms, i.e., any non-trivial phase must be a consequence of the presence of the lattice.

The study of such topological phases is motivated by the presence of so-called topologically protected states. Considering a sample of finite size, the boundary represents an interface to the vacuum, which can be considered an especially simple realization of the trivial insulator. Assuming that a non-trivial phase is realized in the sample, the band gap must close somewhere between the lattice and the vacuum, that is, at the boundary. As a consequence, despite the insulating behavior in the bulk of the material, there are conducting states available at the interface to the vacuum—irrespective of the shape of the boundary. Due to this guaranteed existence of conducting states, one speaks of topological protection, since the topological properties of the sample protect the conducting boundary states from being annihilated by geometric details of the sample. Such a property promises robustness of the conduction behavior of a material, which has indeed been found in the quantum Hall effect that offered an extremely precise way to determine fundamental physical constants. This robustness is interesting, in particular, also in the context of quantum computing, where the stability of quantum states is essential for the reliability of the computation. Over time, quantum states that are coupled to an environment naturally decohere and lose their unique quantum properties, e.g., entanglement, which leads to errors in the computation. The robustness of topological states, on the other hand, promises a protection of the computational states from such external influences, and could therefore lead to much more robust forms of quantum computers. Similarly to the number of holes through the doughnut, the topological properties of a material can be expressed through a number—the so-called topological invariant. In the case of the integer quantum Hall effect, this number is

the Chern number, which takes integer values and will be studied extensively throughout this thesis.

In 2005, another topological phase (the so-called quantum spin Hall phase) was predicted, where, in contrast to the usual quantum Hall effect, time-reversal symmetry is conserved. The experimental observation of this phase has been achieved only two years later, thus confirming the theoretical prediction. Unlike the Chern number, the invariant obtained in this case can only assume two possible values, indicating that all non-trival phases are related to one another. This profound difference to the quantum Hall states underlines the importance of symmetries on the topology of a material. In recent years, advances have been made to incorporate also spatial symmetries, and a complete classification table containing all 230 space groups has been constructed. This table reveals whether non-trivial band crossings that lead to a change in the topological invariant are possible. It is, however, still a daunting task to find actual realizations of topological phases, since for each configuration one has to either compute a topological invariant or prove that it is smoothly connected to a configuration with a known topological phase.

The effects of electron-electron interactions have largely been neglected for a long time, since a similar concept of topology in terms of the most successful theoretical approximation methods proved to be rather complicated. In condensed matter theory, the method of Green's functions, which encodes the properties of the interacting problem in a set of functions of frequency and momentum, is of particular importance. While the simplest of these functions, the so-called single particle Green's function, is much easier to handle than the many-body wave function, it contains only a reduced amount of information, namely information about the single particle excitations. On the other hand, many convenient approximation methods exist, such as dynamical mean field theory (DMFT), cluster perturbation theory (CPT) and the two-particle self-consistent method (TPSC). Assuming that the interactions are not too strong, i.e., the definition of well-defined quasiparticles that can be related to single particle states is still possible, the topological properties can be defined in terms of an effective single particle Hamiltonian—the topological Hamiltonian. While formulations of, e.g., the Hall conductivity in terms of the single particle Green's function have existed for a long time, these approaches have a number of fundamental problems. In particular, the Green's functions are extremely difficult to compute numerically and the computation of the Hall conductivity is subject to the differentiability of the Green's function—a property that is no longer guaranteed in the strongly interacting regime, where the single particle excitations that the Green's function describes are no longer well-defined.

Lacking an exact solution for the Green's function, we are facing a dilemma rooted in the availability of an entire zoo of approximation methods. One method that has been particularly successful in the past is DMFT, which played a central role in the description of the metal-Mott-insulator transition. Even in the rather young field of topological phases, DMFT enjoys popularity and topological phase diagrams have been computed for all significant topological models. The main approximation in DMFT implies that the correction to the single particle spectrum—the so-called self-energy—is local, i.e., independent of momentum. However, the Chern number is a rather direct measure of geometric properties of the corresponding eigenstates as a function of momentum, which casts considerable doubt on the accuracy of the resulting phase diagrams. We tackle this problem by considering the most generic model within DMFT as well as methods that take the momentum-dependent corrections into account to varying degrees. By introducing measures of the amount and importance of the momentum-dependence in terms of a quantity that we call the *self-energy dispersion amplitude*, we compute a phase diagram for the ionic Hubbard model on the square lattice. This model has two convenient features: first, the trivial mass term found in all common topological models is equivalent to the ionic potential in this case, and second, a strong anti-ferromagnetic instability that favors a strong momentum

dependence is observed. The latter implies that the model can be regarded as an upper limit for the importance of momentum-dependent corrections. In our phase diagram, we distinguish two phases where the non-local corrections due to electron-electron interactions are important and unimportant, respectively, as is encoded in our importance measure. We find that non-local effects are mostly very weak and thus unimportant throughout the largest part of the phase diagram, indicating that despite the apparent paradox, DMFT is rather reliable in predicting the correct topological phase. In addition, we show through a simple argument based on the energy scales involved how the expected error of the local approximation can be judged from within DMFT.

Our preceding argument relies mainly on the assumption that a small momentum dependence does not change the topological properties—a fact that we proceed to demonstrate for one of the prototypical topological models: the Haldane model. Since it is expected that not only the amount, but also the type of the momentum-dependence is somehow important for the topological classification, we need to know an exact result to make a strong statement, as in the absence of precise error bounds any approximate method constitutes merely another guess.

Here, we decide to venture off the beaten track and offer a description following a different paradigm. Due to the high numerical complexity of the many-body problem, the exact result is out of reach. Instead of relying on one particular approximate scheme, we therefore propose a stochastic algorithm that in the limit of long running times eventually looks at all possible solutions. This idea is somewhat similar to the Solovay-Strassen test that determines if an integer is prime. Instead of testing exponentially many combinations hoping that one either finds a factorization or proves that there is none, different combinations are tested at random, which provides a probabilistic answer to the original question that converges to the exact answer for long running times. In the context of topological phases, we apply a similar scheme to test the robustness of the Chern number with respect to non-local self-energy corrections. The answer we obtain is encoded in a probability of change that we then analyze to find an exponential suppression of sensitivity or exponential increase in robustness as a function of the distance to the local phase transition that is provided, e.g., by DMFT.

Motivated by the success of the methodology, we look for other applications and turn our attention to the discovery of topological materials. Here, we neglect electron-electron interactions, since the problem is already difficult enough in terms of a single particle picture due to the large variety of thinkable models and the accompanying large dimension of the associated Hilbert spaces. Clearly, hoping to produce a scheme that engineers a sensible candidate material given no information at all would be too optimistic. However, a large amount of knowledge has already been amassed in the field, procured through the study of model systems and experimental evidence. Unfortunately, with increasing complexity of the underlying theoretical methods, the field has become less democratic as not all experts speak the same mathematical language. Our hope is to improve upon this by fostering an understanding of the relationship between topological states on one hand and the crystal structure on the other. Here, the crystal structure is represented in terms of overlap matrix elements of electronic wave functions—a language that is accessible to theoretical and experimental experts alike, independent of the respective background in topological phases.

By narrowing down what information can realistically be extracted from a phase diagram, we evaluate the use of common machine learning techniques. However, we find that measured against the complexity of the original problem, the resulting models are not necessarily easier to understand. We show that a data set can instead also be investigated directly with information theoretical tools that offer insights into the relevance of particular spatial degrees of freedom for the topological classification. By examining a randomly generated data set that covers a particular portion of the total phase space, we show that, given no information in addition to

the Chern number, the Haldane model can be predicted as the prototypical topological model on the honeycomb lattice. In addition, we discover another topological phase that is not described by the Haldane model. Applying our framework to a generic model on the kagome lattice, we then find, in addition to an abstract phase diagram, that even complex relationships between the Chern number and the phases of the matrix elements of the Hamiltonian can be uncovered by this rather simple methodology.

The thesis is structured as follows:

First, we give a short introduction to the theoretical background of the field of topological phases in Chapter 2. This discussion is limited to the topics most relevant for the later chapters and is by no means exhaustive. We note that more details can be found in the mentioned literature.

In Chapter 3, we discuss several important ingredients of condensed matter theory, starting from a detailed investigation of the non-interacting case and moving on to several approximate methods for the treatment of the many-body problem. Here, we reduce the discussion to a minimum due to the abundance of very good literature on the matter.

Since statistics play an important role in this thesis we give a short introduction into the notation and some elementary concepts in Chapter 4, where we also quickly introduce key ideas from information theory and machine learning.

Following these introductory chapters, we organized the original content of this work into three parts that are related to three different projects that were pursued during the last years and build upon each other. Readers who are familiar with the topics discussed in Chapters 2-4 are encouraged to skip these entirely.

The investigation of the momentum-dependence of the self-energy in the ionic Hubbard model is presented in Chapter 5, where we compare results from different numerical methods and construct an effective description of the phase diagram by taking into account the exact local limits. We focus mostly on the square lattice as a limiting case and show that the result obtained for the triangular lattice indeed features a weaker self-energy dispersion.

The statistical methodology is introduced in Chapter 6, where we decompose the self-energy into local and non-local contributions and discuss the effects of each of them separately. Especially the investigation of the non-local part is carried out in great detail and we comment on the implications for the phase diagram, the shape of which depends strongly on the numerical method used.

We turn our attention to the problem of finding new realizations of topological phases in Chapter 7, where we first discuss our perspective on the problem and evaluate several standard machine learning techniques regarding their usefulness in this context. We then introduce the statistical viewpoint and show how information theoretical methods can be used to extract valuable information from the data and to predict topological model systems. For the development of this approach we mainly study the known example of the Haldane model, which here serves as our test bed, and later apply the resulting method to investigate the phase diagram of kagome models. Finally, we discuss also the path towards an application to more realistic materials.

We close with concluding remarks containing a summary of what has been achieved and an outlook on possible future work in the field.

Chapter 2

Topology in Condensed Matter

The mathematical field of topology is very broad. Specifically, when we use the term "topology" here we mean a specific type of equivalence relation that connects Hamiltonians with one another. In fact, as a consequence of this equivalence relation, physical systems can be classified in terms of labels that do not change when the system parameters are modified smoothly without closing the band gap.

We will start this discussion with a phenomenological review of the Hall effect and its quantum analog and then dive a little deeper into the relation to the mathematical theory. Finally, we will arrive at the Chern number as a topological invariant for time-reversal broken phases in two dimensions and discuss algorithms used for numerical computations. This chapter is kept intentionally short since there is a lot of good literature, e.g., the books by Bernevig [1] and Vanderbilt [2] or the review by Hasan [3]. A more in-depth review has also already been written in a previous thesis of mine [4].

2.1 Hall Effect

The classical Hall effect, named after its discoverer Edwin Hall (in 1879) [5], describes the phenomenon of a voltage occurring perpendicular to the direction of a charge current in the presence of a magnetic field—also perpendicular to the direction of current flow. The basic setup is illustrated in Fig. 2.1, where the metallic sample is illustrated as a slab oriented along the x-y plane that is penetrated by a magnetic field $\mathbf{B}$ along the z-direction. A charge current $\mathbf{j}$ is applied along the x-direction and the voltage measurement (V) is carried out along the y-direction. Clearly, there would be a finite voltage along the x-direction that is associated with the current via Ohm's law $V = RI$. The surprising fact discovered by Hall is that there is also a finite voltage along the y-direction where the net current vanishes, which seems to contradict Ohm's law.

This can be understood as follows. For $\mathbf{B} = 0$ the voltage is of course 0, since the metallic sample can be assumed to be unpolarized and the current flow in x-direction does not alter the charge distribution in y-direction. The finite measured voltage must therefore be caused by the magnetic field. This is easily understood by considering the Lorentz force acting on individual charge carriers. With $\mathbf{j} = \rho\mathbf{v}$, ρ being the charge density of electrons, and the velocity $\mathbf{v}$ pointing in the direction of the current. Note that the sign of $\mathbf{v}$ depends on the sign of ρ. Conventionally, $\mathbf{j}$ describes the direction of movement of positively charged holes as indicated in Fig. 2.1. We then have for the Lorentz force

$$\mathbf{f}_\mathrm{L} = \rho\mathbf{E} + \mathbf{j} \times \mathbf{B} = \rho(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \tag{2.1}$$

where $\mathbf{E}$ is the total electric field. Assuming that the magnetic field points in the z-direction

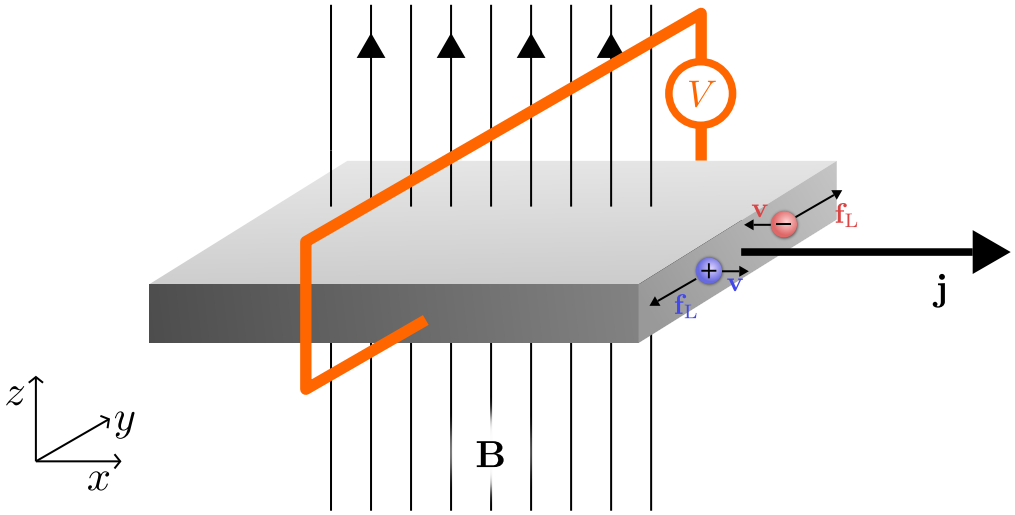

Figure 2.1: Illustration of the Hall effect setup. The metallic sample is probed along three perpendicular directions x, y, z. An electric current $\mathbf{j}$ is flowing along the x-direction and a magnetic field $\mathbf{B}$ is applied along the z-direction. In addition, the voltage V is measured along the y-direction. Assuming an unpolarized sample, the voltage measured would vanish without the magnetic field. Any finite voltage must therefore be a consequence of the applied magnetic field.

and the current flows in the x-direction, i.e., $\mathbf{B} = (0, 0, B)^T$ and $\mathbf{j} = (j, 0, 0)^T$ one obtains simply

$$\mathbf{f}_\mathrm{L} = \rho(\mathbf{E} - vB\mathbf{e}_y). \tag{2.2}$$

There are two finite components, namely f_L^x, which is simply caused by the electric field E_x that drives the applied current, and f_L^y, which is caused entirely by the magnetic field. Since the charges cannot just leave the sample in y-direction, since there are no leads other than those of the voltmeter (which has high resistance) attached, an equilibrium state must be obtained that requires $f_y = 0$, i.e.,

$$E_y = vB = R_H B j_x. \tag{2.3}$$

Here, R_H is the Hall coefficient, which can be defined as

$$R_H = \frac{E_y}{j_x B} = \frac{vB}{j_x B} = \frac{1}{\rho}. \tag{2.4}$$

Given $j_x = {}^I\!/_{dh}$ and $V_H = E_y d$, where d is the depth of the sample (along y) and h the height (along z), the charge density can be related to the experimental parameter I, i.e., the electric current, and the observable Hall voltage V_H

$$V_H = \frac{R_H B I}{h} \tag{2.5}$$

$$\Leftrightarrow \rho = \frac{IB}{V_H h}. \tag{2.6}$$

Since R_H is inversely proportional to the charge density of the charge carriers its measured negative sign reveals that, in fact, moving negative charges produce currents. Apart from the academic interest in measuring the charge carrier density, industrial applications make use of the fact that R_H is a material constant, cf. Eq. 2.4, and can therefore be determined in a controlled environment. The Hall setup can then be applied as a magnetic field sensor that is present in many electronic devices.

2.2 Integer Quantum Hall Effect

The quantum analog of the ordinary Hall effect describes basically the same phenomenon but within a quantum mechanical description. Before diving into any further discussion it makes sense to point out the range of validity of these new observations that were made well after the discovery of the classical Hall effect. Clearly, the physics of the classical Hall effect are not wrong since theory and experimental evidence are in very good agreement. However, the accuracy of the theoretical description depends on the scales on which the effects are investigated. In this case the temperature scale is most relevant. According to statistical mechanics we define expectation values of measurable observables O generally as

$$\langle O \rangle = \frac{\mathrm{tr}\{e^{-\beta H} O\}}{Z}, \tag{2.7}$$

where $Z = \mathrm{tr}\{e^{-\beta H}\}$, H is the Hamiltonian describing the system and $\beta = (kT)^{-1}$ the inverse temperature with Boltzmann's constant k. Note that depending on the ensemble used the single particle spectrum of H can be shifted by the chemical potential, which is not relevant for this discussion. Evaluating the trace over eigenstates of the Hamiltonian we obtain

$$\langle O \rangle = \frac{1}{Z} \sum_n e^{-\beta E_n} \langle E_n | O | E_n \rangle, \tag{2.8}$$

i.e., the expectation value is a weighted average over different quantum expectation values, where the weights are given by the Boltzmann factors $e^{-\beta E_n}/Z$. At high temperatures we have for the energy E_0 of the ground state $kT/E_0 \gg 1$, which implies $\beta E_0 = E_0/(kT) \ll 1$. Therefore, the exponential suppression of contributions from higher-energy states is dampened, i.e., the measured value corresponds to the statistical average over many quantum expectation values, in which the quantum fluctuations inherent to the quantum description eventually average out. Hence the good agreement of measurements with the classical description.

On the other end of the spectrum, i.e., low temperature with $\beta E_0 \gg 1$ the exponential suppression of higher-energy contributions is amplified, which leads to predominantly low energy physics playing a role for the measured observables. In this case quantum fluctuations survive and thus the classical description does not necessarily apply.

The bare fact that there is a distinction between a classical and quantum Hall effect implies that the quantum Hall effect is a low-temperature phenomenon, since all other scales are essentially the same between the two.

2.2.1 TKNN Description

In the following we will briefly review the quantum description of the Hall effect proposed by Thouless, Kohmoto, Nightingale and den Nijs in 1982 [6] (the acronym TKNN refers to the authors' initials), which came in the wake of earlier work of Kosterlitz and Thouless [7, 8] regarding phase transitions in two spatial dimensions. Ultimately, David Thouless and Michael Kosterlitz were awarded the 2016 Nobel Prize in Physics together with Duncan Haldane [9] for the major impact their work on topological phase transitions had on the entire field of condensed matter physics [10]. At the time there were already multiple other proposals published [11–15], however, TKNN offered the first explanation for all experimental observations.

In their seminal paper [6] Thouless et al. discuss the Hall effect in a two-dimensional periodic system described by a potential U (we will use their original notation) that satisfies the periodicity requirement

$$U(x, y) = U(x + a, y + b). \tag{2.9}$$

In addition to that a magnetic field B is applied in the z direction, i.e., perpendicular to the x-y plane. This is essentially the same starting point that we used to describe the classical picture with the addition of the periodic potential Eq. 2.9, which implies that electrons are not considered as free particles. According to Bloch's theorem [16–18], which we will look at in Chapter 3, the eigenfunctions of the Hamiltonian without magnetic field satisfy the conditions

$$\psi_{k_1,k_2}(x+a,y) = e^{ik_1 a}\psi_{k_1,k_2}(x,y) \qquad \text{and} \qquad \psi_{k_1,k_2}(x,y+b) = e^{ik_2 b}\psi_{k_1,k_2}(x,y), \tag{2.10}$$

which are fulfilled by the ansatz

$$\psi_{k_1,k_2}(x,y) = e^{i(k_1 x + k_2 y)}u_{k_1,k_2}(x,y), \tag{2.11}$$

with a periodic function $u_{k_1,k_2}(x+a,y) = u_{k_1,k_2}(x,y+b) = u_{k_1,k_2}(x,y)$. The situation with a magnetic field is a bit more complicated as we can see from the Hamiltonian (here in SI units)

$$H(k_1,k_2) = \frac{1}{2m}\left(\mathbf{p} + e\mathbf{A}(x,y)\right)^2 + U(x,y), \tag{2.12}$$

where $\mathbf{A}$ is the vector potential. Using the gauge $\mathbf{A} = (0, Bx)$ such that $\mathbf{B} = B\mathbf{e}_z$ (we omit the z-component in all other cases, since it is only finite for $\mathbf{B}$) and the same ansatz of Eq. 2.11 we find

$$\begin{aligned} H\psi_{k_1,k_2}(x,y) &= \left[\frac{1}{2m}\left(p_x^2 + (p_y + eBx)^2\right) + U(x,y)\right]\psi_{k_1,k_2}(x,y) && (2.13)\\ &= \frac{\hbar^2}{2m}\left[\left((-i\partial_x)^2 + (-i\partial_y + \frac{eBx}{\hbar})^2\right) + U(x,y)\right]\psi_{k_1,k_2}(x,y) && (2.14)\\ &= e^{i(k_1 x + k_2 y)}\frac{\hbar^2}{2m}\left[(k_1 - i\partial_x)^2 + (k_2 + \frac{eBx}{\hbar} - i\partial_y)^2 + U(x,y)\right]u_{k_1,k_2}(x,y), && (2.15) \end{aligned}$$

so u_{k_1,k_2} must be an eigenfunction of the Hamiltonian $H(k_1,k_2)$ corresponding to the expression in brackets in Eq. 2.15. This Hamiltonian is not translation invariant due to the term proportional to x, which means that also u_{k_1,k_2} must break translation invariance in the presence of a magnetic field. However, by defining

$$u_{k_1,k_2}(x+qa,y) = e^{-i\frac{eBqay}{\hbar}}u_{k_1,k_2}(x,y) \tag{2.16}$$

with $q \in \mathbb{Z}$ the Hamiltonian satisfies $H(x+qa,y) = H(x,y)$, i.e., the period in x–direction is increased by a factor q. This constant is chosen such that the total phase factor when moving along $(x,y) \to (x+qa,y) \to (x+qa,y+b) \to (x,y+b) \to (x,y)$ vanishes, i.e. $eBqab/h = p$, with $p/q = eBab/h \in \mathbb{Q}$ the flux per unit cell (in terms of the flux quantum h/e).

The description of the Hall effect now requires an expression for the Hall conductivity (inverse of the Hall resistivity). The expression used by TKNN goes back to Kubo et al. [19], who have derived an expression for the conductivity after perturbing the magnetic Hamiltonian with an electrostatic potential $\phi(x) = eEx$. Without giving the derivation explicitly, see, e.g., Refs. [1,4], one obtains at zero temperature the linear response result

$$\langle j_l\rangle = \frac{ie^2\hbar E}{A}\sum_{\alpha,\beta}\left(\frac{v^l_{\alpha\beta}v^x_{\beta\alpha}}{(E_\alpha - E_\beta)^2} - \frac{v^x_{\alpha\beta}v^l_{\beta\alpha}}{(E_\alpha - E_\beta)^2}\right), \tag{2.17}$$

where $v^l_{\alpha\beta}$ are the matrix elements of the velocity operator w.r.t. occupied states $|\alpha\rangle$ and unoccupied states $|\beta\rangle$ and $l \in \{x,y\}$. A is the cross sectional area of the sample. With $v^l = \frac{1}{\hbar}\frac{\partial H(k_1,k_2)}{\partial k_l}$ we then obtain

$$\langle j_l\rangle = \frac{ie^2 E}{A\hbar}\sum_{\alpha,\beta}\left(\frac{(\frac{\partial H}{\partial k_l})_{\alpha\beta}(\frac{\partial H}{\partial k_1})_{\beta\alpha} - (\frac{\partial H}{\partial k_1})_{\alpha\beta}(\frac{\partial H}{\partial k_l})_{\beta\alpha}}{(E_\alpha - E_\beta)^2}\right). \tag{2.18}$$

For the conductivity we make use of Ohm's law in the form $\mathbf{j} = \sigma\mathbf{E}$, where σ is the conductivity tensor, and therefore with $\mathbf{E} = (E, 0)$, $j_y = \sigma_{yx}E$ we arrive at the equation given in the TKNN paper (up to a global minus sign that is a result of our choice of coordinates):

$$\sigma_{yx} = \frac{ie^2}{A\hbar}\sum_{\alpha\beta}\left(\frac{(\frac{\partial H}{\partial k_2})_{\alpha\beta}(\frac{\partial H}{\partial k_1})_{\beta\alpha} - (\frac{\partial H}{\partial k_1})_{\alpha\beta}(\frac{\partial H}{\partial k_2})_{\beta\alpha}}{(E_\alpha - E_\beta)^2}\right). \tag{2.19}$$

In order to obtain the next equation in the paper that establishes the relation to the eigenstates of H a little algebra is required. Since we couldn't find a reference that provides all steps along the way this is carried out in more detail here. We express the Hamiltonian in terms of $|u\rangle$ as $H = \sum_u |u\rangle E_u \langle u|$. Applying a derivative to this spectral expansion we obtain the relation

$$\frac{\partial H}{\partial k} = \sum_u \left[|\partial_k u\rangle\langle u|E_u + |u\rangle\langle u|\partial_k E_u + |u\rangle\langle\partial_k u|E_u\right], \tag{2.20}$$

where we use the notation $|\partial_k u\rangle := \partial_k|u\rangle$. For the matrix elements this implies

$$\left(\frac{\partial H}{\partial k}\right)_{\alpha\beta} = \langle\alpha|\frac{\partial H}{\partial k}|\beta\rangle = \langle\alpha|\partial_k\beta\rangle E_\beta + \langle\partial_k\alpha|\beta\rangle E_\alpha, \tag{2.21}$$

where we took into account that $\langle\alpha|\beta\rangle = 0$ since they correspond to different quantum numbers. Inserting this into Eq. 2.19 we obtain a lengthy expression. We focus here on the first term in the numerator that yields

$$\langle\alpha|\partial_{k_2}\beta\rangle\langle\beta|\partial_{k_1}\alpha\rangle E_\alpha E_\beta + \langle\alpha|\partial_{k_2}\beta\rangle\langle\partial_{k_1}\beta|\alpha\rangle E_\beta^2 + \langle\partial_{k_2}\alpha|\beta\rangle\langle\beta|\partial_{k_1}\alpha\rangle E_\alpha^2 + \langle\partial_{k_2}\alpha|\beta\rangle\langle\partial_{k_1}\beta|\alpha\rangle E_\alpha E_\beta. \tag{2.22}$$

It is possible to simplify this using a little trick. We note that $\langle\alpha|\beta\rangle = 0$ and therefore $\partial_k\langle\alpha|\beta\rangle = 0$. Expansion of the l.h.s. yields the useful relation

$$\langle\partial_k\alpha|\beta\rangle = -\langle\alpha|\partial_k\beta\rangle. \tag{2.23}$$

Application of Eq. 2.23 then reveals that all terms in Eq. 2.22 are proportional to each other so that we have

$$\text{(Eq. 2.22)} = \langle\partial_{k_2}\alpha|\beta\rangle\langle\beta|\partial_{k_1}\alpha\rangle\left(E_\alpha - E_\beta\right)^2. \tag{2.24}$$

The same relation can be obtained for the second term in the numerator of Eq. 2.19 with reversed order of the derivatives so we that finally arrive at

$$\sigma_{yx} = \frac{ie^2}{A\hbar}\sum_{\alpha\beta}\left[\langle\partial_{k_2}\alpha|\beta\rangle\langle\beta|\partial_{k_1}\alpha\rangle - \langle\partial_{k_1}\alpha|\beta\rangle\langle\beta|\partial_{k_2}\alpha\rangle\right]. \tag{2.25}$$

The sum over unoccupied states $|\beta\rangle$ can be disposed of by using the relation

$$\sum_{\beta\ \text{unocc.}} |\beta\rangle\langle\beta| = \text{Id} - \sum_{\alpha\ \text{occ.}} |\alpha\rangle\langle\alpha|. \tag{2.26}$$

Inserting this one finds that the terms containing double sums over occupied states vanish due to the relation $\langle\partial_k\alpha|\alpha'\rangle = -\langle\alpha|\partial_k\alpha'\rangle$. What remains is

$$\sigma_{yx} = \frac{ie^2}{A\hbar}\sum_{\alpha}\left[\langle\partial_{k_2}\alpha|\partial_{k_1}\alpha\rangle - \langle\partial_{k_1}\alpha|\partial_{k_2}\alpha\rangle\right], \tag{2.27}$$

which is identical to the one from the TKNN paper after replacing vectors $|\alpha\rangle$ with the corresponding wave functions $u_\alpha(x)$ and turning the k sum (here implicit in the sum over all quantum numbers α) into an integral via $\sum_k = \frac{A}{(2\pi)^2}\int \mathrm{d}^2k$

$$\sigma_{yx} = \frac{ie^2}{2\pi h}\sum_\alpha \int \mathrm{d}^2k \int \mathrm{d}^2x \ \left[\frac{\partial u_\alpha^*}{\partial k_2}\frac{\partial u_\alpha}{\partial k_1} - \frac{\partial u_\alpha^*}{\partial k_1}\frac{\partial u_\alpha}{\partial k_2}\right]. \tag{2.28}$$

We can use the identity

$$\begin{aligned}\partial_{k_i}(f\partial_{k_j}g) - \partial_{k_j}(f\partial_{k_i}g) &= (\partial_{k_i}f)\partial_{k_j}g + f\partial_{k_i}\partial_{k_j}g - (\partial_{k_j}f)\partial_{k_i}g - f\partial_{k_j}\partial_{k_i}g \\ &= (\partial_{k_i}f)\partial_{k_j}g - (\partial_{k_j}f)\partial_{k_i}g\end{aligned} \tag{2.29}$$

to arrive at

$$\begin{aligned}\sigma_{yx} &= \frac{ie^2}{2\pi h}\sum_\alpha \int \mathrm{d}^2k \int \mathrm{d}^2x \ \left[\frac{\partial}{\partial k_2}\left(u_\alpha^*\frac{\partial u_\alpha}{\partial k_1}\right) - \frac{\partial}{\partial k_1}\left(u_\alpha^*\frac{\partial u_\alpha}{\partial k_2}\right)\right] \\ &= -\frac{ie^2}{2\pi h}\sum_\alpha \int \mathrm{d}^2k \int \mathrm{d}^2x \ \left[\nabla_k \times (u_\alpha^*\nabla_k u_\alpha)\right]_3,\end{aligned} \tag{2.30}$$

where, again, the minus sign is simply a consequence of the choice of coordinates. Together with $\sigma_{xy} = -\sigma_{yx}$ we arrive at the same sign as TKNN. Applying Stokes' theorem allows us to transform the integral over the Brillouin zone into an integral along the path around the unit cell

$$\sigma_{xy} = \frac{ie^2}{2\pi h}\sum_\alpha \oint_{\partial \mathrm{BZ}} \mathrm{d}\mathbf{k} \int \mathrm{d}^2x \ u_\alpha^*\nabla_k u_\alpha. \tag{2.31}$$

In the paper it is then argued that if the bands do not overlap the wavefunction must be uniquely defined at every point in the Brillouin zone assuming that one chooses a global gauge and the only degree of freedom left is a phase factor $e^{i\theta(k)}$ that appears when skipping from one edge of the Brillouin zone to the other. Due to the normalization condition of Bloch states we then have only one degree of freedom, namely the phase $|u_\alpha(k)\rangle = e^{i\theta(k)}|u_\alpha(k')\rangle$ so that the integral can be evaluated to the total phase change acquired along the path. At $k = 0$ the wave function must be well-defined and therefore only integer multiples of 2π are allowed, which implies that

$$\sigma_{xy} = \frac{e^2}{h}N, \tag{2.32}$$

where $N \in \mathbb{Z}$, i.e., the Hall conductivity is expected to be quantized to integer multiples of e^2/h.

Eq. 2.32 is already one of the main results of the TKNN paper. In order to understand this, however, an explicit calculation has been carried out. The authors assume a simple cosine potential, which yields a model studied already by Harper [20] and Hofstadter [21] and show that the Hall conductivity changes only upon closing the energy gap, i.e., it is constant within each gapped region regardless of the exact parameters of the model. This is an indication of the robustness of the conduction behavior with respect to geometric details in the sample or small variations in the magnetic field. It was later shown that the TKNN integer is the only topological index for these systems and is also related to topological indices in higher dimensions [22].

2.2.2 Berry Phase and Chern Number

We now quickly review the geometric phase introduced by Michael Berry in 1984 [23] that turns out to be related to the expression Eq. 2.31 for the Hall conductivity found by TKNN.

Berry considered a Hamiltonian whose parameters $\mathbf{R}$ are varied along a path C adiabatically. Due to the adiabaticity the n-th eigenstate will remain the n-th eigenstate and only acquire a phase factor $|A(t)| = 1$ with time

$$|\psi(t)\rangle = A(t)|n, \mathbf{R}(t)\rangle. \tag{2.33}$$

Without going into the details of the derivation one arrives at an expression containing two contributions [23, 24]

$$A(t) = \exp\left[-\frac{i}{\hbar}\int_0^t E_n(\mathbf{R}(s))\ \mathrm{d}s\right] \exp\left[\int_C \langle n, \mathbf{R}|\nabla_{\mathbf{R}}|n, \mathbf{R}\rangle\ \mathrm{d}\mathbf{R}\right], \tag{2.34}$$

where the first term is simply the dynamical phase obtained from the unitary time-evolution of the state. The second term is Berry's geometric phase that depends only on the geometric properties of the eigenstates $|n, \mathbf{R}\rangle$ as a function of the parameters $\mathbf{R}$. The second integrand can be shown to be purely imaginary and therefore we define the geometric phase as

$$\gamma(C) = i\int_C \langle n, \mathbf{R}|\nabla_{\mathbf{R}}|n, \mathbf{R}\rangle\ \mathrm{d}\mathbf{R}. \tag{2.35}$$

If the path C considered is a closed loop it was shown that the Berry phase is gauge invariant and assuming $\mathbf{R}$ two be two-dimensional it can be expressed via Stokes' theorem as

$$\gamma(C) = i\int \nabla_{\mathbf{R}} \times \langle n, \mathbf{R}|\nabla_{\mathbf{R}}|n, \mathbf{R}\rangle\ \mathrm{d}^2R, \tag{2.36}$$

where the integrand is called the Berry curvature. Apparently, the integral in the TKNN formula Eq. 2.29 is identical to the Berry curvature, when taking k_1, k_2 as the parameters of the Hamiltonian that are varied throughout the Brillouin zone.

The Berry curvature integrated over the entire Brillouin zone can be seen to correspond to the so-called Chern number [1, 2], which is a topological invariant that encodes the homotopy invariance between different sets of fibre bundles. The Chern number for a fibre bundle that is given by the Bloch Hamiltonian $H(\mathbf{k})$ is defined as

$$C = \frac{i}{2\pi}\sum_\alpha \int_{\mathrm{BZ}} \left[\partial_{k_1}\langle k, \alpha|\partial_{k_2}|k, \alpha\rangle - \partial_{k_2}\langle k, \alpha|\partial_{k_1}|k, \alpha\rangle\right] \mathrm{d}^2k, \tag{2.37}$$

where the sum goes over all occupied bands α. Clearly, $C \in \mathbb{Z}$ follows from comparison with Eq. 2.36 since for the closed path $\gamma(C)$ is an integer multiple of 2π. The Hall conductivity is therefore proportional to the Chern number and therefore serves as a direct probe of a topological invariant. Technically, the Chern number serves as a check for the possibility of defining a smooth global gauge [25]. Considering that it is possible to choose such a gauge throughout the entire Brillouin zone and also across the edges we can map the Brillouin zone to a torus due to the periodicity of the wave functions in k. This, however, means that the closed path of integration which corresponds to the boundary $\partial\mathrm{BZ} = 0$ vanishes and as a consequence also the Chern number vanishes. $C = 0$ is therefore an indication that Stokes' theorem can be applied. This case where a global gauge exists is called the trivial case, whereas all other cases with $C \neq 0$ are called non-trivial, since such a global gauge cannot be chosen and therefore Stokes' theorem cannot be applied to the toroidal Brillouin zone.

Changes of the Chern number can only happen if the adiabaticity requirement is violated, which is the case if bands cross at the Fermi level. In this case the adiabatic transport of one state through the Brillouin zone is hindered by missing (unoccupied) states. Technically, the Chern

number is not defined for such metallic systems due to this reason. Considering two systems with different Chern numbers it is guaranteed that these cannot be connected adiabatically without closing the band gap somewhere along the path. This fact is also known as the *bulk-boundary correspondence* [3, 26, 27], which states that protected metallic states appear at the boundary of a topologically non-trivial sample—protected by the topology of the system, which guarantees their existence even under smooth transformations as long as the bulk band gap remains finite.

The Berry phase and curvature were defined using Eq. 2.33, which implies multiplication with a simple phase factor. This situation is extremely simplified and does not apply to a generic case where multiple bands are present and instead the phase factor would have to be replaced by a $U(N)$ transformation. This so-called non-abelian case has been studied by Wilczek and Zee as a generalization of the Berry phase, see Ref. [28].

2.2.3 Experimental Observations

The idealized experimental setup is generally the same as for the measurement of the classical Hall effect, however, in order to observe the quantum effects a very low temperature is required.

Historically, the quantum Hall effect was discovered experimentally before a theoretical understanding of the phenomenon was known. In 1980 Klaus von Klitzing performed a Hall measurement on GaAs-$Al_{0.3}Ga_{0.7}As$, a semiconducting heterostructure, at low temperatures and discovered an exactly quantized value of the Hall resistance as a function of small variations of the magnetic field [29, 30]. In addition, also changes of the sample size led to the measurement of the same value indicating a strong robustness of the Hall resistance w.r.t. experimental parameters. The measured value was quantized to $R_{\mathrm{H}} = {}^{h}/_{4e^2}$, which corresponds to a Hall conductivity $\sigma_{\mathrm{H}} = 4\frac{e^2}{h}$, i.e. a Chern number $C = 4$.

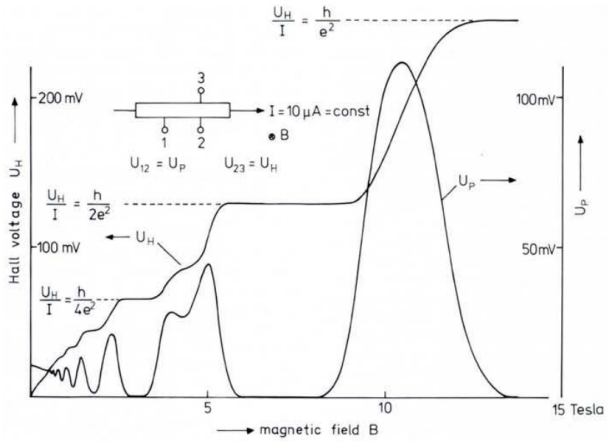

Figure 2.2: Hall voltage measured by von Klitzing et al. on a GaAs-$Al_{0.3}Ga_{0.7}As$ sample as a function of the magnetic field. The plateaus at the quantized values $U_{\mathrm{H}}/I = h/(e^2)$, $U_{\mathrm{H}}/I = h/(2e^2)$, $U_{\mathrm{H}}/I = h/(4e^2)$ corresponding to Chern numbers $1, 2, 4$, respectively, can be seen clearly. The transitions between two plateaus are continuous and occur whenever the longitudinal voltage U_P that is measured parallel to the current (along x in figure Fig. 2.1) is finite. [figure from Ref. [31]]

In Fig. 2.2 we show a measurement on a GaAs-$Al_{0.3}Ga_{0.7}As$ heterostructure from Ref. [31]. The inverse Hall conductivity is given by $R_{\mathrm{H}} = U_{\mathrm{H}}/I$ so that the plateaus observed are in direct correspondence to the plateaus in the Hall conductivity. In fact, it is possible to determine the Chern number from this measurement. Unlike the Chern number, though, the Hall resistance is a continuous function of the magnetic field, which implies that there are non-integer values in an intermediate regime between two plateaus. We observe that these intermediate regimes coincide with positive values of U_{P}, the longitudinal voltage. As described earlier, in the presence of strong

magnetic fields the Lorentz force on the charge carriers induces a non-negligible perpendicular electric field so that the matrix-valued nature of the conductivity becomes important. With $\mathbf{j} = \sigma\mathbf{E}$ we obtain by defining the inverse $\rho = \sigma^{-1}$ (resistivity) $\mathbf{E} = \rho\mathbf{j}$. The condition $\rho\sigma = \mathrm{Id}$ implies that

$$\begin{pmatrix} \rho_{xx} & \rho_{xy} \\ \rho_{yx} & \rho_{yy} \end{pmatrix} \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix} = \begin{pmatrix} \rho_{xx}\sigma_{xx} + \rho_{xy}\sigma_{yx} & \rho_{xx}\sigma_{xy} + \rho_{xy}\sigma_{yy} \\ \rho_{yx}\sigma_{xx} + \rho_{yy}\sigma_{yx} & \rho_{yx}\sigma_{xy} + \rho_{yy}\sigma_{yy} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{2.38}$$

and therefore

$$\begin{aligned} \rho_{xx}\sigma_{xx} + \rho_{xy}\sigma_{yx} &= 1 \\ \rho_{xx}\sigma_{xy} + \rho_{xy}\sigma_{yy} &= 0 \\ \rho_{yx}\sigma_{xx} + \rho_{yy}\sigma_{yx} &= 0 \\ \rho_{yx}\sigma_{xy} + \rho_{yy}\sigma_{yy} &= 1. \end{aligned} \tag{2.39}$$

These equations are satisfied by

$$\rho = \frac{1}{\sigma_{xx}\sigma_{yy} - \sigma_{xy}\sigma_{yx}} \begin{pmatrix} \sigma_{yy} & -\sigma_{xy} \\ -\sigma_{yx} & \sigma_{xx} \end{pmatrix}, \tag{2.40}$$

and with $\sigma_{xx} = \sigma_{yy}$ and $\sigma_{xy} = -\sigma_{yx}$ we have

$$\rho_{xx} = \frac{\sigma_{xx}}{\sigma_{xx}^2 + \sigma_{xy}^2}. \tag{2.41}$$

For small σ_{xy}, i.e., weak magnetic fields, $\rho_{xx} = {}^1\!/_{\sigma_{xx}}$ essentially. However, if σ_{xy} is large we can neglect σ_{xx} in the denominator and therefore $\rho_{xx} \propto \sigma_{xx}$. Since $U_\mathrm{P} = R_{xx}I$ and $I = \mathrm{const}$, vanishing U_P indicates that R_{xx} vanishes and therefore also σ_{xx}, i.e., the material is insulating. On the other hand, in the intermediate regimes U_P is finite indicating that the sample is metallic. This observation is in agreement with the bulk-boundary correspondence, which requires a metallic region in between two distinct non-trivial regions with different Chern numbers.

The Hall conductivity of Eq. 2.29 does not depend on the precise geometry or other properties of the sample, therefore it makes sense that the measured values do not depend on small changes. However, all calculations assumed clean samples in a sense that the crystals are perfectly periodic, which is unrealistic for actual samples that feature some degree of disorder. Regarding the robustness w.r.t. these sources of noise there have been several studies [11, 13–15] that showed that the ideal bands get broadened once disorder is introduced, however, only the centers of the bands are extended states that behave as the states in an ideal system and therefore contribute to the conductivity. The tails on the other hand are localized and therefore cannot blur out the perfect quantization of the Hall conductivity.

2.3 Edge States

While the discussion of TKNN assumed an infinite system, all physical samples are of finite size and therefore the Chern number is technically not well-defined for any realistic material. It has, however, been shown that in finite systems the Hall current is localized to the edge of the system and extended along the edge. Together with the bulk-boundary correspondence the measured Hall conductivity is therefore a probe of the topology of the idealized infinite system, since the topological nature of the phase is encoded in the robustness of these edge states w.r.t. perturbations to the system [26, 27].

It has first been shown by Halperin [15] that the Hall current is indeed carried by extended quasi-one-dimensional states that are localized to the edges of the sample and that these states

are also robust w.r.t. disorder. In a simplified picture Halperin argues that the states are simply "displaced, locally, to go around these regions". Therefore, the non-trivial system features states that avoid scattering with impurities provided that the impurity density is not too large.

2.4 The Tenfold Way Classification

While the discovery and the subsequent discussion of the quantum Hall effect ultimately led to the importance of the field of topological states of matter in physics today it has been found that the integer quantum Hall state is only one of many possible topological phases that can be classified by a variety of topological invariants. In fact, Altland and Zirnbauer were able to relate topological information about physical systems to a classification scheme introduced by Cartan 70 years earlier [32–34]. The resulting set of symmetry classes has been used by Schnyder, Ryu et al. to define a periodic table to topological insulators [35–37] that is shown in Table 2.1 and will be explained in the following.

Symmetry				d							
Cartan label	TRS	PHS	SLS	1	2	3	4	5	6	7	8
A	0	0	0	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$
AIII	0	0	1	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0
AI	1	0	0	0	0	0	$\mathbb{Z}$	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}$
BDI	1	1	1	$\mathbb{Z}$	0	0	0	$\mathbb{Z}$	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$
D	0	1	0	$\mathbb{Z}_2$	$\mathbb{Z}$	0	0	0	$\mathbb{Z}$	0	$\mathbb{Z}_2$
DIII	-1	1	1	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}$	0	0	0	$\mathbb{Z}$	0
AII	-1	0	0	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}$	0	0	0	$\mathbb{Z}$
CII	-1	-1	1	$\mathbb{Z}$	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}$	0	0	0
C	0	-1	0	0	$\mathbb{Z}$	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}$	0	0
CI	1	-1	1	0	0	$\mathbb{Z}$	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}$	0

Table 2.1: Periodic table of topological insulators. The left-most column lists the Cartan labels of the 10 distinct symmetric classes. The anti-unitary symmetries (time-reversal TRS, particle-hole PHS and sublattice symmetry SLS) pertaining to each class are listed, where 0 means that the symmetry is not present and ± 1 indicates that the symmetry operator squares to ± 1. To the right the topological classification is listed for spatial dimensions $d = 1, \ldots, 8$. 0 means no non-trivial class, $\mathbb{Z}$ an integer topological invariant and $\mathbb{Z}_2$ a binary topological invariant. The pattern is periodic in d with period 8. [table from Ref. [3]]

Roughly speaking the general idea behind this classification scheme is the realization that symmetries place constraints on the possible matrices that can represent a Hamiltonian. Assuming that the Hamiltonian has no particular symmetry there is only one constraint, namely it has to be hermitian in order to guarantee real eigenvalues and unitary time evolution. Given the size of the Hilbert space, which depends on the type of spin and the number of sublattices, this constraint reduces the amount of possible matrices. Let e.g. $\dim(\mathcal{H}) = N$, then an arbitrary complex matrix has $2N^2$ degrees of freedom: the real and imaginary parts of the N^2 matrix elements. If, however, the matrix must be hermitian, then we have

$$H = H^\dagger, \tag{2.42}$$

and therefore the real parts of the upper and lower triangular matrices must be equal and the imaginary parts must be the additive inverse, which implies zero imaginary part on the diagonal. The total number of real degrees of freedom is therefore $N(N+1)/2$ and for imaginary parts $N(N-1)/2$. In total these are N^2, i.e., exactly half the number in comparison to the

most general matrix. It is clear that this reduction of the number of independent degrees of freedom has consequences for the properties of the matrix and therefore the physics it represents. Following the same idea originating in random matrix theory that deals with disordered systems and studies the possible properties of such randomized Hamiltonians under certain constraints [32, 38], Schnyder et al. look at two types of symmetries that refer in principle to time-reversal symmetry (TRS) and particle-hole symmetry (PHS) [35]

$$\begin{array}{llll} H = CH^TC^{-1}, & CC^\dagger = 1, & C^T = \pm C & \text{(TRS)} \\ H = -CH^TC^{-1}, & CC^\dagger = 1, & C^T = \pm C & \text{(PHS)}, \end{array} \tag{2.43}$$

where C is the matrix representing the respective symmetry transformation. Eq. 2.43 is a bit difficult to understand. Starting from the left we have, since $H = H^\dagger = (H^T)^*$, $H^T = H^*$ and therefore the symmetry operation C includes complex conjugation which we can express as $S = CK$, where K is complex conjugation, i.e., S is an anti-unitary symmetry. The matrix C is unitary and with the third equation $CC^* = \pm 1$ or equivalently $S^2 = \pm 1$. This means that for both TRS and PHS there exist two distinct forms which square to 1 or -1, respectively. In total this leaves three possibilities (0, 1, -1), where 0 indicates the absence of the symmetry, that can be combined between the two to yield 9 distinct symmetry classes. In addition, one can include the compound symmetry that arises if both TRS and PHS are present, which is referred to sublattice symmetry (SLS) or chiral symmetry. Clearly, this is already included in the 9 combinations, however, there is an additional possibility that neither TRS nor PHS are present but the combination of both is. Including this additional unitary symmetry there are now 10 different symmetry classes that are conventionally assigned the labels given by Cartan [34].

The number of distinct topological classes is then determined by looking at the corresponding homotopy group that arises from the spatial dimension and the different constraints set for the Hamiltonian matrix at each k. The result is then that shown in Table 2.1, where 0 indicates no non-trivial topological class, $\mathbb{Z}$ a countable infinite number of non-trivial classes that can be defined through an integer topological invariant and $\mathbb{Z}_2$ one non-trivial class. Shown are only the first 8 dimensions, since the pattern repeats periodically as a consequence of the Bott periodicity theorem [39].

We note that the tenfold way classification discussed here is a consequence of anti-unitary symmetries only. A modified table taking into account inversion symmetry on top of these anti-unitary symmetries has been derived by Lu et al. [40]. Recently, also other advances have been made to describe a topological classification subject to certain spatial symmetries [40–42].

2.5 Interacting Topological Phases

Generally, the discussion of topology in condensed matter systems has mostly focused on non-interacting models, where a quadratic Hamiltonian can be written down in terms of a periodic quantum number—the momentum k—and the band index. The mapping from the Brillouin zone to the eigenspace of the Hamiltonian at momentum k then represents the fibre bundle that is characterized in terms of a Chern number. In interacting systems this is no longer possible since momentum itself is not a good quantum number and therefore the Hamiltonian cannot be written as a function of k.

The Chern number being a sum over individual contributions from occupied bands implies a related problem, namely that bands are not well-defined and in particular the interacting state contains not necessarily fully occupied single-particle bands. Thus, the existence of a simple mapping from an interacting state to a non-interacting topological index is not immediately clear. Nevertheless, we will introduce such a mapping in Sec. 2.5.2.

The topology of interacting systems is typically regarded as a property of the ground state and not of the Hamiltonian. In the non-interacting case both were related by the fact that the ground state corresponds simply to an occupation of the Hamiltonian's eigenstates up to the Fermi level. Therefore, the topology is determined by a subset of the single particle Hamiltonian's eigenstates. Considering a manybody Hamiltonian each eigenstate already contains the information about all particles and therefore the topology should be a property of a single eigenstate.

2.5.1 Hall Conductivity

In the following, we will quickly summarize the definition of a topological invariant for interacting systems that is in a way a generalization of the Chern number to a manybody ground state [43, 44]. During the discussion of the TKNN paper it was noted that a finite Chern number is basically a consequence of the impossibility to fix a global gauge throughout the Brillouin zone. Attempting to do so anyway will lead to a phase jump when going from $k = \pi$ to $-\pi$. A similar idea can be extended to the interacting ground state by considering different boundary conditions. A direct application of the non-interacting concept is not possible since states cannot be labeled with the crystal momentum quantum number. In a two-dimensional system we consider a lattice of size $L_1 \times L_2$. Quite generally one can formulate boundary conditions as [43]

$$\begin{aligned} T_i(L_1\mathbf{e}_x)\psi(x_i) &= e^{i\alpha L_1}\psi(x_i) =: e^{i\theta}\psi(x_i) \\ T_i(L_2\mathbf{e}_y)\psi(x_i) &= e^{i\beta L_2}\psi(x_i) =: e^{i\varphi}\psi(x_i), \end{aligned} \tag{2.44}$$

which bears resemblance to the Bloch condition for the eigenstates of the single particle Hamiltonian. T_i are the translation operators for electrons labeled by i. Using these boundary conditions it was shown that the Hall conductivity can be written in terms of the manybody ground state $|\psi_0\rangle$ as [43]

$$\sigma_{xy} = \frac{ie^2}{\hbar}\left[\langle\partial_\theta\phi_0|\partial_\varphi\phi_0\rangle - \langle\partial_\varphi\phi_0|\partial_\theta\phi_0\rangle\right], \tag{2.45}$$

where $|\phi_0\rangle$ is equal to $|\psi_0\rangle$ up to a unitary transformation $|\phi_0\rangle = e^{-i\alpha\sum_j x_j - i\beta\sum_j y_j}|\psi_0\rangle$. Eq. 2.45 has almost the same form as the TKNN invariant of Eq. 2.29. It is then motivated that this expression must yield the same value for any combination of phases θ, φ in the thermodynamic limit and also for finite systems under the condition that the spectral gap always remains finite when the phases are varied. Hence, the expression can be replaced by an average over all possible phases

$$\sigma_{xy} = \frac{ie^2}{\hbar}\int_0^{2\pi}\int_0^{2\pi}\left[\langle\partial_\theta\phi_0|\partial_\varphi\phi_0\rangle - \langle\partial_\varphi\phi_0|\partial_\theta\phi_0\rangle\right]\,\mathrm{d}\theta\mathrm{d}\varphi. \tag{2.46}$$

This expression for the Hall conductivity is now formally identical to the TKNN result by replacing $(k_1, k_2) \to (\theta, \varphi)$ and the sum over occupied Bloch states with the manybody ground state. The corresponding invariant is an integer due to the requirement that changing θ, φ along a path that starts and ends at 0 the ground state must be recovered up to a phase that is an integer multiple of 2π. This integer invariant is the integral of the Berry curvature on the torus parameterized by the boundary conditions θ, φ and the two-dimensional space spanned by θ, ϕ was later coined *twist space* in the context of a similar formulation for a $\mathbb{Z}_2$ invariant for time-reversal invariant systems [45–48]. The term "twist" refers to the twisted boundary conditions introduced in Eq. 2.44.

2.5.2 Topological Hamiltonian

A completely different approach to that found by Niu et al. has been discussed by Wang and Zhang [49–51]. This approach will be the starting point for some of the work presented in this thesis and therefore we will give a brief introduction. More details and an extensive discussion of the proof filling in all details that were cut from the paper can be found in Ref. [4].

The starting point for Wang et al. is another generalization of the single particle equations to interacting systems, where the knowledge about the quantization of the Hall conductivity is exploited to define a corresponding invariant for the interacting case. In 1986 Ishikawa and Matsuyama derived an expression for the Hall conductivity in terms of the single particle Green's function [52, 53]

$$\sigma_{xy} = \frac{e^2}{h}\frac{1}{24\pi^2}\iint \mathrm{d}k_0\mathrm{d}^2k\ \varepsilon_{\mu\nu\rho}\,\mathrm{tr}\left[\frac{\partial G^{-1}}{\partial k_\mu}G\frac{\partial G^{-1}}{\partial k_\nu}G\frac{\partial G^{-1}}{\partial k_\rho}G\right], \tag{2.47}$$

where $G = G(k_0 = i\omega, \mathbf{k})$ is the single particle imaginary frequency Green's function and $\varepsilon_{\mu\nu\rho}$ the Levi-Civita symbol. This expression is extremely difficult to compute in practice due to the required knowledge of the full frequency information on the imaginary axis. Implicitly it is assumed that $T = 0$, i.e., the discrete Matsubara spectrum becomes continuous, which allows for a definition of the derivative. In this limit, however, many of the established numerical methods fail due to either the inherent discreteness in the spectral information due to finite size (e.g., exact diagonalization) or the infamous sign problem of Monte Carlo methods [54, 55].

The major breakthrough by Wang et al. was the realization that for moderate interaction strengths, where the Green's function is analytic, a smooth connection from any finite frequency can be made to zero frequency via the definition

$$G(i\omega, \mathbf{k}, \lambda) = (1-\lambda)G(i\omega, \mathbf{k}) + \lambda\left[i\omega + G^{-1}(0, \mathbf{k})\right]^{-1}, \tag{2.48}$$

where $\lambda \in [0, 1]$. Clearly, at $\lambda = 0$ one recovers $G(i\omega, \mathbf{k})$ and at $\lambda = 1$ we have $\left[i\omega + G^{-1}(0, \mathbf{k})\right]^{-1}$, which is the single particle Green's function with the full frequency-dependent self-energy $\Sigma(i\omega, \mathbf{k})$ replaced with the value at $i\omega = 0$. With this definition it was shown that $G(i\omega, \mathbf{k}, \lambda)$ has no zero eigenvalues, i.e., remains invertible for any value of λ. This property together with the fact that the Ishikawa formula of Eq. 2.47 is a topological invariant leads to the conclusion that the value of σ_{xy} must be invariant w.r.t. changes of λ. Therefore, σ_{xy} can be evaluated with $G(i\omega, \mathbf{k}, \lambda = 1)$, removing the necessity of acquiring frequency information beyond that at $i\omega = 0$.

Given the sole dependence on $G(0, \mathbf{k})$ and its inverse, a representation in terms of the eigenvectors proves to be very convenient. In particular, it was shown that the Green's function is hermitian, i.e.,

$$[G^{-1}(0, \mathbf{k})]^\dagger = G^{-1}(0, \mathbf{k}), \tag{2.49}$$

and therefore the spectrum is real. From the corresponding eigenvalue equation

$$G^{-1}(0, \mathbf{k})|\alpha, \mathbf{k}\rangle = \mu_\alpha(\mathbf{k})|\alpha, \mathbf{k}\rangle, \tag{2.50}$$

one then obtains eigenvectors $|\alpha, \mathbf{k}\rangle$ that form two orthogonal subspaces: the "L-space" with $\mu_\alpha(\mathbf{k}) < 0$ and "R-space" with $\mu_\alpha(\mathbf{k}) > 0$. Writing out the definition of $G^{-1}(0, \mathbf{k})$ we find that

$$G^{-1}(0, \mathbf{k}) = \mu - H_0(\mathbf{k}) - \Sigma(0, \mathbf{k}), \tag{2.51}$$

where $H_0(\mathbf{k})$ is the non-interacting Bloch Hamiltonian and μ the chemical potential. Reversing the sign gives rise to the definition of the so-called *topological Hamiltonian*

$$h_t(\mathbf{k}) = -G^{-1}(0, \mathbf{k}) = H_0(\mathbf{k}) - \mu + \Sigma(0, \mathbf{k}), \tag{2.52}$$

which is essentially just the single particle Hamiltonian modulated with the zero-frequency self-energy. Occupied states, i.e., eigenstates of h_t with negative energy, correspond to the states from the R-space. Wang and Zhang have then shown that inserting the spectral decomposition of $G(i\omega, \mathbf{k}, \lambda = 1)$ into Eq. 2.47 yields an expression

$$\sigma_{xy} = \frac{e^2}{h} C_1, \tag{2.53}$$

with an invariant

$$C_1 = \frac{1}{2\pi i} \sum_{\alpha \in \text{R-space}} \int \mathrm{d}^2 k \ \left[\partial_x \langle \alpha, \mathbf{k} | \partial_y | \alpha, \mathbf{k} \rangle - \partial_y \langle \alpha, \mathbf{k} | \partial_x | \alpha, \mathbf{k} \rangle \right]. \tag{2.54}$$

Eq. 2.54 is apparently the Chern number and therefore the topological invariant is identical to the TKNN integer in the non-interacting limit. The major difference of this approach, though, is that C_1 is defined not in terms of the Bloch states of the non-interacting Hamiltonian but in terms of eigenstates of the topological Hamiltonian that is well-defined also for interacting systems provided that μ lies in a band gap and the Green's function is free of singularities, which is the case away from the Mott insulating phase. The fact that the interacting invariant has the same analytical form as the non-interacting invariant implies that all algorithms that are available for the computation of the Chern number can also be used to compute the topological invariant for the interacting system. Only the self-energy at zero frequency has to be provided in addition to the non-interacting Bloch Hamiltonian. Since C_1 is essentially the Chern number defined for a different fibre bundle we will not distinguish between C_1 and C in the following and leave it to the respective context to clarify whether we are discussing interacting or non-interacting systems.

The terminology of a topological Hamiltonian introduced above, that was not mentioned in the original paper by Wang and Zhang, was first used in another work by Zhong and Wang [51], where the usefulness of this quantity as an effective model was elaborated further. Unsurprisingly, applied to interacting systems, the topological Hamiltonian is only capable of describing the correct topological invariant. Using this effective description to obtain any other physical quantities at finite interaction strengths neglects the complete frequency dependence of the self-energy and is therefore likely to produce inaccurate results.

Another interesting approach was developed by Gurarie [56] based on earlier work by Volovik [57], where expressions for known topological invariants are formulated in terms of the single-particle Green's function. This allows a computation even for interacting systems, since the Green's function—unlike the Bloch states—is always defined. It is found that while the bulk-boundary correspondence guarantees gapless edge modes in the non-interacting case, it is in principle possible for interacting systems to avoid this such that the topological invariant changes across a boundary without edge states. This is likely to happen only in the strongly interacting case, though, since the spectral weight corresponding to the edge modes has to be gapped out by a cancellation of zeros and poles of the Green's function.

We note that with few exceptions such as the fractional quantum Hall effect [58, 59], the topological excitations of the Kitaev model [60,61] and the Kitaev toric code [62], the description of topological properties of interacting systems is—to this date and the best of our knowledge—mostly limited to generalizations of non-interacting theories to correlated systems [43,56,63–65]. At this point, a similarly developed theory for topological effects that is exclusive to interacting systems, i.e., does not rely on the existence of well-defined single-particle excitations, is not available, and the development of such ideas is an active field of research [66–70].

2.6 Fukui Algorithm

Finally, we review an efficient algorithm for the computation of the Chern number that will be used extensively throughout this thesis. In principle, the Chern number can be computed via the formula of Eq. 2.37. However, the occurring derivatives make a numerical calculation rather cumbersome as all derivatives have to be replaced by finite differences. This is possible, in principle, but it turns out that the resulting algorithm is highly sensitive to the number of k-points used, which is inversely proportional to the step size h in the finite difference rules that produce an error that is typically $\mathcal{O}(h^2)$. Luckily, Fukui et al. [71] were able to construct an algorithm that offers a much better scaling and in practice can be seen as almost independent of the number of k-points used, only requiring a minimal number that can usually be chosen somewhere around $N_{k_i} = 10$. In addition, the algorithm is gauge-invariant, which removes the necessity to choose a fixed gauge throughout the Brillouin zone.

Their algorithm is constructed as follows. Assuming that the Bloch states are denoted as $|n, \mathbf{k}\rangle$ with reciprocal lattice vectors $\mathbf{b}_1, \mathbf{b}_2$ that define "unit vectors" $\mathbf{e}_1 = \mathbf{b}_1/N_1$ and $\mathbf{e}_2 = \mathbf{b}_2/N_2$, where $N_{1,2}$ is the number of k-points in the two dimensions, respectively, we have

$$|n, \mathbf{k} + N_i \mathbf{e}_i\rangle = |n, \mathbf{k}\rangle \quad i \in \{1, 2\}. \tag{2.55}$$

The phase change of a state $|n, \mathbf{k}\rangle$ when moving on the discrete grid can be defined as

$$U_i(\mathbf{k}) = \frac{\langle \mathbf{k} | \mathbf{k} + \mathbf{e}_i \rangle}{|\langle \mathbf{k} | \mathbf{k} + \mathbf{e}_i \rangle|}, \tag{2.56}$$

which is rather clear if one considers a decomposition

$$|n, \mathbf{k} + \mathbf{e}_i\rangle = c_n |n, \mathbf{k}\rangle + \sum_{n' \neq n} c_{n'} |n', \mathbf{k}\rangle. \tag{2.57}$$

Then, due to the orthonormalization of Bloch states Eq. 2.56 reduces to $U_i(\mathbf{k}) = e^{i\phi}$, where we define $c_n = e^{i\phi}|c_n|$. The total phase obtained by going around a plaquette starting at $\mathbf{k}$ is then defined as

$$\tilde{F}_{12}(\mathbf{k}) = \log \left[U_1(\mathbf{k}) U_2(\mathbf{k} + \mathbf{e}_1) U_1(\mathbf{k} + \mathbf{e}_2)^{-1} U_2(\mathbf{k})^{-1} \right]. \tag{2.58}$$

An illustration of why this is indeed correct is shown in Fig. 2.3. With $\mathbf{k}$ in the lower left corner of the plaquette we move counter-clockwise around the square. The phase for a path along $\mathbf{e}_i$ starting at $\mathbf{q}$ is given by $U_i(\mathbf{q})$ as per Eq. 2.56 and therefore we obtain the phase factors indicated in the figure with arrows marking the direction of the respective subpath. These factors already correspond to the four terms in Eq. 2.58. The first two point along our chosen path, while the other two point in the opposite direction. We ameliorate this by reversing the phase, which corresponds to complex conjugation or the inverse.

Given the phase accumulated along a path around one plaquette the phase accumulated along the path that goes around the entire Brillouin zone can be expressed through $\tilde{F}_{12}$ as[1]

$$\tilde{c}_n = \frac{1}{2\pi i} \sum_{\mathbf{k}} \tilde{F}_{12}(\mathbf{k}). \tag{2.59}$$

The proof for this is rather obvious since the phase factors of inner paths cancel one another as every inner path is traversed exactly twice in opposite directions when $\tilde{F}_{12}$ is summed over all plaquettes.

[1] Note that the definition in Ref. [71] yields exactly $-C$, i.e., the opposite sign as the Chern number, cf. Eq. 2.37. This does not necessarily matter, since we can label equivalence classes any way we like, however, the result of Eq. 2.59 will not accurately reflect the sign of the Hall conductivity.

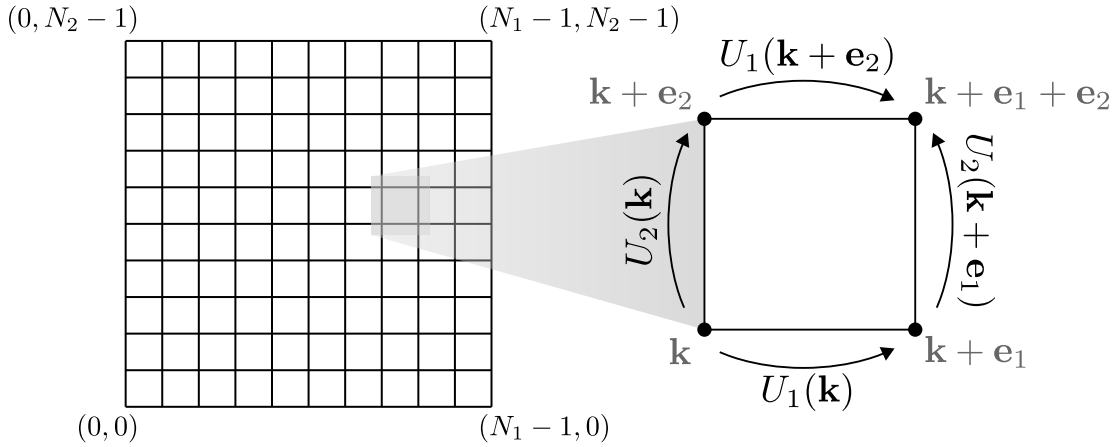

Figure 2.3: We illustrate the meaning of $\tilde{F}_{12}$ in Eq. 2.58. The lattice in k-space is given on the left, here we are only interested in the phase change for a path around one plaquette. We pick the plaquette with its lower left corner at $\mathbf{k}$. The other corners are then given by $\mathbf{k}+\mathbf{e}_1, \mathbf{k}+\mathbf{e}_2, \mathbf{k}+\mathbf{e}_1+\mathbf{e}_2$. With the phase factor along a path given by $U_i(\mathbf{k})$ we obtain the four phase factors marked to the right, where the arrows indicate the direction of the path. Apparently, for a counter-clockwise path we need to invert two of the phases and therefore arrive at Eq. 2.58.

Fukui et al. proceed by showing that $\tilde{c}_n$ is integer quantized for any choice of N_1, N_2, i.e., for arbitrary coarseness of the grid. In particular, $\tilde{c}_n$ approaches the correct Chern number for $N_1, N_2 \to \infty$ and it was shown that the correct Chern number is already obtained at very small $N_i \sim \mathcal{O}(1)$ in most cases.

Another important property is the gauge invariance of $\tilde{c}_n$, which can be proven by considering a general $U(1)$ gauge transformation g that is defined through

$$g : |n\rangle \mapsto e^{i\phi}|n\rangle. \tag{2.60}$$

g accounts for an arbitrary phase factor of a given eigenstate $|n\rangle$. Applying this transformation to all states with phases denoted as follows

$$|n, \mathbf{k}\rangle \mapsto e^{i\phi}|n, \mathbf{k}\rangle, \quad |n, \mathbf{k}+\mathbf{e}_i\rangle \mapsto e^{i\phi_i}|n, \mathbf{k}+\mathbf{e}_i\rangle, \tag{2.61}$$

we obtain

$$\begin{aligned}\tilde{F}'_{12}(\mathbf{k}) = {} & e^{-i\phi}e^{i\phi_1}\langle n, \mathbf{k}|n, \mathbf{k}+\mathbf{e}_1\rangle e^{-i\phi_1}e^{i\phi_{12}}\langle n, \mathbf{k}+\mathbf{e}_1|n, \mathbf{k}+\mathbf{e}_1+\mathbf{e}_2\rangle \\ & \times e^{i\phi_2}e^{-i\phi_{12}}\langle n, \mathbf{k}+\mathbf{e}_2|n, \mathbf{k}+\mathbf{e}_1+\mathbf{e}_2\rangle^{-1}e^{i\phi}e^{-i\phi_2}\langle n, \mathbf{k}|n, \mathbf{k}+\mathbf{e}_2\rangle^{-1}.\end{aligned} \tag{2.62}$$

All phase factors cancel, i.e.,

$$\tilde{F}'_{12}(\mathbf{k}) = \tilde{F}_{12}(\mathbf{k}). \tag{2.63}$$

Thus, $\tilde{F}_{12}(\mathbf{k})$ is invariant under $U(1)$ gauge transformations.

This has an important practical advantage. Since eigenvectors are by definition only defined up to a phase factor (assuming we require normalization), different algorithms will return different vectors. The gauge invariance of Eq. 2.58 guarantees the gauge invariance of the entire algorithm and therefore of the Chern number. It is therefore not necessary to fix a specific gauge.

In practice, one will often encounter cases where not all bands are separated from one another. This represents a serious problem to the algorithm discussed above, since the values $U_i(\mathbf{k})$ that are defined per band are no longer well-defined. In fact, a single degeneracy between a pair of

states between two bands means that at this point the phase has no well-defined value. This may not matter if the degeneracy lies somewhere between grid points or not on the path surrounding the Brillouin zone, however, this cannot be guaranteed in an arbitrary situation. In order to be able to treat the most general case, it can be shown that the simple modification

$$U_i(\mathbf{k}) = \frac{\det\left(\langle n, \mathbf{k} | m, \mathbf{k} + \mathbf{e}_i \rangle\right)}{|\det\left(\langle n, \mathbf{k} | m, \mathbf{k} + \mathbf{e}_i \rangle\right)|}, \tag{2.64}$$

instead of Eq. 2.56 leads to the correct result even in the generic non-abelian case [4,71–73]. Here, $\langle n, \mathbf{k} | m, \mathbf{k} + \mathbf{e}_i \rangle$ represents a matrix whose matrix elements are defined as $A_{nm} = \langle n, \mathbf{k} | m, \mathbf{k} + \mathbf{e}_i \rangle$.

Chapter 3

Methods

In this chapter we will review some important aspects of condensed matter theory that are of importance for this work. We start with a rather detailed discussion of the single particle picture, which is important also for understanding the general interacting case, since the framework of the topological Hamiltonian establishes a mapping between the two in the context of topological phases. By working out the details of the tight-binding description we set the stage for the discussion in Chapter 7, where we will build our approach onto this representation.

The review of many-body theory is kept intentionally short due to the abundance of good literature on the matter that we will point out instead.

We will review the method of exact diagonalization (ED) in Sec. 3.3 that will be used in Chapter 5, also focusing on some numerical details that we found interesting. As an extension to ED we discuss also cluster perturbation theory in Sec. 3.4, where we conclude that the method is not useful for the purpose of extracting details about the momentum-dependence of the self-energy that is discussed in Chapter 5 and Chapter 6.

This chapter concludes with a short primer on dynamical mean field theory, where we motivate that in this approximation the self-energy is momentum-independent, which is the starting point for the discussion in Chapter 5.

To anyone not interested in these rather basic concepts we recommend skipping this chapter. The discussion of topological phases will continue in Chapter 5, for which a deep understanding of most of the content discussed in the following is not required.

3.1 Single Particle Case

We begin with the discussion of the non-interacting limit, which is governed by the single particle Schrödinger equation

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V(x)\right)\psi_n(x) = E_n\psi_n(x), \tag{3.1}$$

where n represents a complete set of quantum numbers, $V(x) = V(x+a)$ is the periodic lattice potential and a the lattice constant. Eq. 3.1 can be solved in many different ways, the most straight-forward ones being integration via the shooting method or discretization of the Laplacian. Even analytically, the quantum number n can be made more concrete using the Bloch theorem [16].

3.1.1 Bloch Theorem

The periodicity of the potential $V(x) = V(x+G)$, where G is a lattice vector connecting any two lattice points, implies that V commutes with the operator that performs translations about

G:

$$T_G^{-1} V(x) T_G f(x) = V(x+G) f(x) = V(x) f(x), \tag{3.2}$$

i.e. $[T_G, V] = 0$. The translation operator is defined via

$$T_G^{-1} x T_G = x + G, \tag{3.3}$$

which implies the linearity in G

$$T_{G_1+G_2}^{-1} x T_{G_1+G_2} = x + G_1 + G_2 = T_{G_1}^{-1} T_{G_2}^{-1} x T_{G_2} T_{G_1}. \tag{3.4}$$

The only function satisfying Eq. 3.4 is the exponential function. With an operator-valued linear function $g : \mathbb{R}^n \to \mathrm{Lin}(\mathbb{R}^n \to \mathbb{R}^n)$ we have

$$T_G = e^{g(G)}. \tag{3.5}$$

We insert this into the definition

$$T_G^{-1} x T_G = e^{-g(G)} x e^{g(G)} = x + G \tag{3.6}$$

$$\Leftrightarrow \quad (x e^{g(G)} - e^{g(G)} x) = [x, e^{g(G)}] = G e^{g(G)} \tag{3.7}$$

and then use the series expansion of the exponential function

$$\sum_{n=0}^{\infty} \frac{1}{n!} [x, g(G)^n] = G e^{g(G)}. \tag{3.8}$$

Due to the linearity of g only the terms with exponent 1 of G can survive, i.e.

$$[x, g(G)] \stackrel{!}{=} G. \tag{3.9}$$

Using the canonical commutation relation $[x, p_x] = i\hbar$ we find a solution for g

$$g(G) = -\frac{i}{\hbar} p \cdot G, \tag{3.10}$$

that satisfies Eq. 3.9 and together with the property

$$[A, B^n] = B[A, B^{n-1}] + [A, B^{n-1}]B \tag{3.11}$$

also Eq. 3.8. Given the form of the translation operator

$$T_G = e^{-\frac{i}{\hbar} p \cdot G} \tag{3.12}$$

we can immediately conclude that $T_G^\dagger = T_G^{-1}$, i.e., T_G is unitary. As such it has a complete set of eigenvectors and its eigenvalues are of modulus 1, i.e.,

$$T_G |\lambda\rangle = \lambda |\lambda\rangle \tag{3.13}$$

with $|\lambda| = 1$. Apparently, all momentum eigenstates are also eigenstates of T_G with

$$\lambda_k = e^{-ik \cdot G}, \tag{3.14}$$

where $k = \lambda_p / \hbar$. Given the commutation between V and T_G we also have $[H, T_G] = 0$ and therefore the quantum numbers can be chosen as $n = (k, \alpha)$, where the band index α contains

the additional non-spatial degrees of freedom. With Eq. 3.3 we have $[x, T_G] = GT_G$ and thus with $\hat{x}|x\rangle = x|x\rangle$ (here the "hat" serves to distinguish the operator $\hat{x}$ from the eigenvalue x)

$$\hat{x}T_G|x\rangle = GT_G|x\rangle + T_G\hat{x}|x\rangle = (x+G)T_G|x\rangle. \tag{3.15}$$

This means that $T_G|x\rangle$ is also an eigenstate of x and therefore we define

$$T_G\psi(x) = \langle x|T_G|\psi\rangle = \langle x-G|\psi\rangle = \psi(x-G). \tag{3.16}$$

For the simultaneous eigenstates of H and T_G this means that

$$\psi_{k,\alpha}(x-G) = T_G\psi_{k,\alpha}(x) = e^{-ik\cdot G}\psi_{k,\alpha}(x) \tag{3.17}$$

or in the more conventional form by flipping the sign of G:

$$\psi_{k,\alpha}(x+G) = e^{ik\cdot G}\psi_{k,\alpha}(x). \tag{3.18}$$

Eq. 3.18 is known as Bloch's theorem. It states that the eigenstates of the Hamiltonian of a single particle on a periodic lattice are themselves periodic up to a phase factor. By defining

$$\psi_{k,\alpha}(x) = e^{ik\cdot x}u_{k,\alpha}(x) \tag{3.19}$$

we obtain

$$e^{ik\cdot(x+G)}u_{k,\alpha}(x+G) = e^{ik\cdot(G+x)}u_{k,\alpha}(x) \tag{3.20}$$

$$\Leftrightarrow \quad u_{k,\alpha}(x+G) = u_{k,\alpha}(x), \tag{3.21}$$

i.e., the wave function is composed of a periodic function $u_{k,\alpha}(x)$ times a plane wave. Eq. 3.19 is a direct consequence of the Bloch theorem and is therefore sometimes referred to under the same name.

3.1.2 Reciprocal Lattice and Brillouin Zone

We define the primitive lattice vectors a_i as the basis of the lattice that satisfies for any point G the relation

$$G = \sum_i n_i a_i \tag{3.22}$$

with $n_i \in \mathbb{Z}$. We defined the crystal momentum k in Eq. 3.14. Due to the superposition above we can write

$$\lambda_k = \prod_j e^{-in_j k\cdot a_j}, \tag{3.23}$$

where a certain arbitrariness in the definition of k is revealed due to the 2π periodicity of the exponential function. By defining

$$k = \sum_i y_i b_i \tag{3.24}$$

with $y_i \in \mathbb{R}$ and choosing $a_i \cdot b_j = 2\pi\delta_{ij}$ it is clear that

$$k \cdot a_i = y_i b_j \cdot a_i = 2\pi y_i \delta_{ij}, \tag{3.25}$$

i.e., y_i are defined only modulo 1, because integer shifts leave the value of λ_k invariant. This implies that k can be restricted to the volume of the parallelepiped spanned by the vectors

b_i. The set $\{b_i\}$ then forms the basis of the reciprocal lattice. As a consequence, we find for reciprocal lattice vectors $K = \sum_i m_i b_i$ that

$$\psi_{K,\alpha}(x+G) = e^{iK\cdot G}\psi_{K,\alpha}(x) = \psi_{K,\alpha}(x), \tag{3.26}$$

which indicates that eigenvalues of the translation operator for all reciprocal lattice vectors are degenerate. With fixed band index this means that $u_{k,\alpha}(x) = u_{k+K,\alpha}(x)$. In practice, we cannot interpret k as the real kinetic momentum of the electron from where it originated in Eq. 3.14, since due to the periodicity it is only well-defined modulo 2π. Instead, the kinetic momentum is equal to k only up to a reciprocal lattice vector. As a consequence, also conservation of the crystal momentum or quasi-momentum k is relaxed to a conservation modulo a reciprocal lattice vector.

In our calculations we will for simplicity always choose k in the unit parallelepiped spanned by the three reciprocal lattice vectors as given in Eq. 3.24 with $0 \leq y_i < 1$. While we will colloquially call this the first Brillouin zone this is technically not true. The first Brillouin zone is formally defined as the Wigner-Seitz cell around the origin in reciprocal space. We illustrate the difference, but also equivalence between the two representations in Fig. 3.1. Due to the periodicity of the Bloch states the unit cell in reciprocal space can be shifted around to place, e.g., $k = 0$ in the center of the cell. For the square lattice this produces the Wigner-Seitz cell, while for lattices with non-orthogonal lattice vectors the most symmetric cell is more complicated.

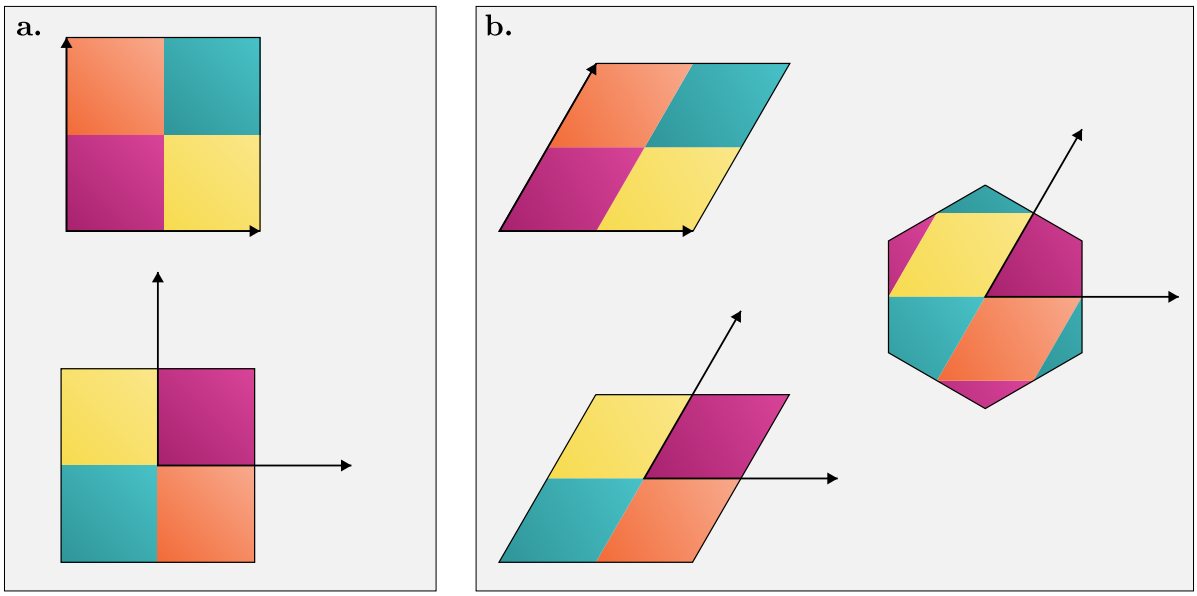

Figure 3.1: Brillouin zone vs. unit cell of the reciprocal lattice for **a.** the square lattice and **b.** the hexagonal lattice. For the square lattice, **a.**, we illustrate the unit cell (top), which is spanned by the two basis vectors, and the first Brillouin zone (bottom). Corresponding regions are marked in the same colors. It turns out that in this case the Brillouin zone is just a translation of the unit cell by $-(b_1+b_2)/2$. In **b.**, we show the same for the hexagonal lattice. The unit cell (top) can again be shifted, which results in the equivalent representation below. In this case (basically any case, where the lattice vectors are non-orthogonal) the shifted unit cell is not identical to the first Brillouin zone shown in the right-most picture.

For completeness, we note that the definition of b_i,

$$a_i \cdot b_j = 2\pi\delta_{ij}, \tag{3.27}$$

implies that b_i is orthogonal to a_j with $j \neq i$. This can be fulfilled in three dimensions with $b_i = ca_j \times a_k$, where $c \in \mathbb{R}$ is a constant and for uniqueness the tuple (ijk) is a cyclic permutation of (123). Due to $v \cdot (v \times w) = 0 \ \forall v, w$ the orthogonality is always satisfied and since $\{a_i\}$ are linearly independent there is always a finite overlap of b_i with a_i. Inserting the ansatz into the definition above we then obtain

$$a_i \cdot c(a_j \times a_k) = 2\pi, \tag{3.28}$$

and therefore

$$b_i = 2\pi \frac{a_j \times a_k}{a_i \cdot (a_j \times a_k)}, \tag{3.29}$$

which uniquely defines the reciprocal lattice vectors. In lower dimensions this formula can easily be adapted. For two dimensions one chooses $a_3 = (0, 0, 1)^T$, which then yields b_1, b_2 with the desired properties and since necessarily $b_1, b_2 \perp a_3$, only the first two components are nonzero, i.e.,

$$b_1^{\mathrm{2D}} = 2\pi \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} \\ -a_{21} \end{pmatrix}, \quad b_2^{\mathrm{2D}} = 2\pi \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} -a_{12} \\ a_{11} \end{pmatrix}. \tag{3.30}$$

In one dimension the same game can be repeated with $a_2 = (0, 1, 0)^T$ and $a_3 = (0, 0, 1)^T$ and one obtains

$$b_1^{\mathrm{1D}} = \frac{2\pi}{a_1}. \tag{3.31}$$

Incidentally, any function f that is periodic on the lattice can be represented by a Fourier series in terms of reciprocal lattice vectors

$$f(x) = \sum_K \tilde{f}_K e^{iK \cdot x}. \tag{3.32}$$

This is clear, since if we assume that K is not a reciprocal lattice vector and insert the periodicity requirement we find

$$f(x + G) = \sum_K \tilde{f}_K e^{iK \cdot x} e^{iK \cdot G} \stackrel{!}{=} f(x). \tag{3.33}$$

This is only satisfied if $K \cdot G = 2\pi n$, $n \in \mathbb{Z}$, which is a contradiction to the assumption that K is not a reciprocal lattice vector.

3.1.3 Solution of the Single Particle Schrödinger Equation

Typically we are only interested in the functions $u_{k,\alpha}$, since we can construct the wave function easily by multiplying a plane wave. Inserting the identity Eq. 3.19 into the Schrödinger equation [Eq. 3.1] we obtain

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V(x)\right) e^{ik \cdot x} u_{k,\alpha}(x) = E_{k,\alpha} e^{ik \cdot x} u_{k,\alpha}(x) \tag{3.34}$$

$$\left(-\frac{\hbar^2}{2m}\nabla(ike^{ik \cdot x} + e^{ik \cdot x}\nabla) + V(x)\right) u_{k,\alpha}(x) = E_{k,\alpha} e^{ik \cdot x} u_{k,\alpha}(x) \tag{3.35}$$

$$e^{ik \cdot x}\left(-\frac{\hbar^2}{2m}(-k^2 + 2ik\nabla + \nabla^2) + V(x)\right) u_{k,\alpha}(x) = E_{k,\alpha} e^{ik \cdot x} u_{k,\alpha}(x), \tag{3.36}$$

and after rearranging

$$\left(\nabla^2 + 2ik\nabla - \left[k^2 + \frac{2m}{\hbar^2}V(x)\right]\right) u_{k,\alpha}(x) = -\frac{2mE_{k,\alpha}}{\hbar^2} u_{k,\alpha}(x). \tag{3.37}$$

Apparently, the Hilbert space of the $u_{k,\alpha}$ functions is block diagonal in wave-vectors k, which means we can solve Eq. 3.37 for arbitrary values of k to compute all eigenstates of H. The technical procedure to accomplish this is to Fourier transform both $V(x)$ and $u_{k,\alpha}(x)$ and then compute the eigen-decomposition of the resulting matrix in terms of the Fourier coefficients. This is outlined below.

We first define the Fourier series for $u_{k,\alpha}(x)$ and $V(x)$ according to Eq. 3.32

$$u_{k,\alpha}(x) = \sum_K c_{K,k,\alpha} e^{iK\cdot x}, \tag{3.38}$$

$$V(x) = \sum_Q V_Q e^{iQ\cdot x}. \tag{3.39}$$

Inserting these into Eq. 3.37 yields

$$\sum_K \left(-K^2 - 2k\cdot K - k^2 - \frac{2m}{\hbar^2} \sum_Q V_Q e^{iQ\cdot x} \right) c_K e^{iK\cdot x} = -\frac{2mE_{k,\alpha}}{\hbar^2} \sum_K c_K e^{iK\cdot x}. \tag{3.40}$$

Due to the linear independence of the exponential functions with different K we must have an identity for any specific choice of K of

$$(K+k)^2 c_K e^{iK\cdot x} + \frac{2m}{\hbar^2} \sum_Q V_Q e^{i(K'+Q)\cdot x} \delta_{K,Q+K'} c_{K'} = \frac{2mE_{k,\alpha}}{\hbar^2} c_K e^{iK\cdot x}, \tag{3.41}$$

or

$$\sum_{K'} \left[(K+k)^2 \delta_{K,K'} + \frac{2m}{\hbar^2} \sum_Q V_Q \delta_{K,Q+K'} \right] c_{K'} = \frac{2mE_{k,\alpha}}{\hbar^2} c_K. \tag{3.42}$$

Eq. 3.42 has the form of a matrix vector product, where K plays the role of the row index and K' that of the column index. Given the potential V, the reciprocal lattice vectors K and the parameter vector k one can determine the matrix representation of the Hamiltonian. The eigenvalues are then related to the $E_{k,\alpha}$. In order to make the calculation unit-free it makes sense to define a length unit $a = \min_i\{\parallel a_i \parallel_2\}$ and with that the unit of energy $E_0 = \frac{\hbar^2}{2ma^2}$. Then, Eq. 3.42 reduces to

$$\sum_{K'} \left[(K+k)^2 \delta_{K,K'} + \sum_Q V_Q \delta_{K,Q+K'} \right] c_{K'} = E_{k,\alpha} c_K, \tag{3.43}$$

where all momenta K and k are measured in units of a^{-1} and all energies V_Q and $E_{k,\alpha}$ in E_0.

3.1.4 Wannier Basis

The Bloch functions that we have previously discussed are periodic on the lattice and therefore cannot be normalized over the entire lattice

$$\int_{-\infty}^{\infty} |\psi_{k,\alpha}(x)|^2 = \sum_G \int_{\text{unitcell}} |\psi_{k,\alpha}(x-G)|^2 = \sum_G \int_{\text{unitcell}} |\psi_{k,\alpha}(x)|^2 \tag{3.44}$$

$$= N \int_{\text{unitcell}} |\psi_{k,\alpha}(x)|^2, \tag{3.45}$$

where N is the number of unit cells. In practice, N is very large and the limit $N \to \infty$ is taken for infinite, i.e., fully periodic lattices. While this description works very well for metals,

where electrons are highly delocalized, the applicability to electronic wavefunctions in insulators is less obvious, since electrons are known to be localized. Therefore, it seems odd to think of a single electron as being distributed throughout the whole lattice. In fact, this counterintuitive result is a consequence of the single particle approximation. Remember that we constructed our Hamiltonian of the atomic potentials for all atoms in the crystal (these contain primarily the Coulomb potentials of the nuclei, but could contain effective screening due to bound electrons) and placed only a single electron into the system. From this reasoning it makes sense that the electron does not arbitrarily decide to sit in any one unit cell. In fact, if we placed N non-interacting electrons into the lattice we would expect them to distribute evenly throughout the crystal in thermal equilibrium, i.e., no external fields. This situation, however, effectively corresponds to the localized single electron picture due to the indistinguishability of quantum particles. There is simply no way to tell which electron is which and therefore the superposition of N localized states and the delocalized state are effectively equivalent. We illustrate this

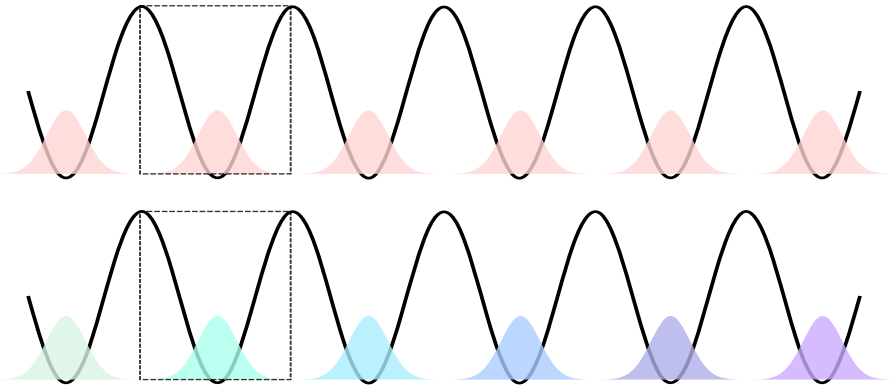

Figure 3.2: Illustration of the delocalized nature of Bloch states in deep lattices. The top figure represents the Bloch state, which is distributed throughout the entire crystal. The unit cell is indicated by a box. In the bottom plot we show the same function, however, now the wave function in each unit cell has a different color. Regarded as separate wave functions this implies separate electrons sitting in different unit cells, while maintaining the same overall translational symmetry.

equivalence in Fig. 3.2, where the top image shows a periodic amplitude of the wave function that is only finite at the atomic sites due to the Coulomb attraction between electrons and the nuclei. The bottom image represents a number of different wavefunctions that are only finite at one particular site. The combined electron density of this state cannot be distinguished from the delocalized picture above.

We now want to proceed with finding a set of functions that works with the second picture of individual electrons. From Fig. 3.2 it is clear that these functions should satisfy

$$w_G(r) = w_{G+G'}(r + G'). \tag{3.46}$$

Each function w_G is more or less confined to the unit cell at lattice vector G and therefore not periodic over the whole lattice. However, functions at different G should just be shifted versions of w_0. Taking into account that the Fourier transform of delocalized functions is localized we make the educated guess

$$w_{G,\alpha}(r) = \frac{1}{\sqrt{N}} \sum_k e^{-ik\cdot G} \psi_{k,\alpha}(r), \tag{3.47}$$

which yields

$$w_{G+G',\alpha}(r) = \frac{1}{\sqrt{N}} \sum_k e^{-ik\cdot(G+G')} \psi_{k,\alpha}(r) = \frac{1}{\sqrt{N}} \sum_k e^{-ik\cdot G} e^{ik\cdot(r-G')} u_{k,\alpha}(r) \tag{3.48}$$

$$= w_{G,\alpha}(r - G'). \tag{3.49}$$

For the last equality we used the fact that $u_{k,\alpha}$ are lattice periodic and G' is a lattice vector. Apparently, Eq. 3.47 fulfills our requirement and is therefore a suitable choice for the localized basis. The prefactor $1/\sqrt{N}$ ensures the normalization. We call $w_{G,\alpha}(r)$ the Wannier functions of band α [74, 75].

Given two Wannier functions $w_{G,\alpha}, w_{G',\beta}$ we have for the scalar product

$$\langle G, \alpha | G', \beta \rangle = \int_{-\infty}^{\infty} w^*_{G,\alpha}(r) w_{G',\beta}(r) \, \mathrm{d}r \tag{3.50}$$

$$= \frac{1}{N} \sum_{k,q} e^{-ik\cdot G} e^{iq\cdot G'} \int_{-\infty}^{\infty} \psi^*_{k,\alpha}(r) \psi_{q,\beta}(r) \, \mathrm{d}r \tag{3.51}$$

$$= \frac{1}{N} \sum_{k,q} e^{-ik\cdot G} e^{iq\cdot G'} \delta_{k,q} \delta_{\alpha,\beta} = \frac{1}{N} \sum_k e^{ik\cdot(G'-G)} \delta_{\alpha,\beta} \tag{3.52}$$

$$= \delta_{G,G'} \delta_{\alpha,\beta}. \tag{3.53}$$

Hence, the Wannier functions form a complete orthonormal basis. Unfortunately, the localization of $w_{G,\alpha}$ is not necessarily guaranteed by the definition Eq. 3.47. It turns out that every $\psi_{k,\alpha}$ can be multiplied with arbitrary phase factors that leave the state itself invariant, however, greatly affect a superposition such as the Fourier series. For simple systems it is usually enough to define a common gauge for all Bloch states by, e.g., requiring that the first Fourier coefficient $c_{K=0,k,\alpha}$ in the expansion of $u_{k,\alpha}(x)$ is real. In more complicated cases the localization has to be enforced by requiring a minimization of the spread

$$\int_{\mathbb{R}} x^2 |w_{0,\alpha}(x)|^2 \, \mathrm{d}^3x - \left| \int_{\mathbb{R}} x |w_{0,\alpha}(x)|^2 \, \mathrm{d}^3x \right|^2 . \tag{3.54}$$

Apparently, it is enough to perform this one minimization per band index α, since all other Wannier functions of the same band can be computed by exploiting the translational symmetry. This is, however, only possible if the bands are separable. Degenerate bands require a more complicated treatment [76, 77].

It has been shown by W. Kohn that in one dimension the maximally localized solution is unique under the condition that it is real and (anti-) symmetric w.r.t. reflections [78]. This can be generalized to arbitrary dimensions provided the potential is separable, i.e., $V(x) = \sum_i V(x_i)$. The question whether or not an exponentially localized Wannier function exists in the first place has been answered rather recently by Brouder et al. [79] and it was shown that for insulators the existence of such a function hinges on the topological properties of the system. While time-reversal symmetric systems are guaranteed to have localized Wannier functions, the opposite is the case for topological insulators with finite Chern number. In fact, Wannier functions are intricately related to the topological properties of a system via the theory of polarization, which can be described through the Berry phase when expressed in terms of the Bloch states or equivalently in terms of the Wannier centers [80, 81].

In one dimension, a convenient method for the computation of maximally localized Wannier functions based on the band projected position operator is possible [82]. Interestingly, this definition does not technically require translational symmetry and therefore also allows the description of disordered systems.

The previously studied Schrödinger equation of Eq. 3.1, which involved differential operators, can be reformulated in the Wannier basis by expanding the wave function as

$$\psi(x) = \sum_{G,\alpha} c_{G,\alpha} w_{G,\alpha}(x). \tag{3.55}$$

Then, the Schrödinger equation reads

$$\sum_{G,\alpha} \left(-\frac{\hbar^2}{2m}\nabla^2 + V(x) \right) w_{G,\alpha}(x) = \sum_{G,\alpha} E_n c_{G,\alpha} w_{G,\alpha}(x). \tag{3.56}$$

Multiplication with $w^*_{G',\beta}$ from the left and integration over x yields

$$\sum_{G,\alpha} \int w^*_{G',\beta}(x) \left(-\frac{\hbar^2}{2m}\nabla^2 + V(x) \right) w_{G,\alpha}(x)\ \mathrm{d}x c_{G,\alpha} = \sum_{G,\alpha} E_n c_{G,\alpha} \int w_{G',\beta}(x) w_{G,\alpha}(x)\ \mathrm{d}x,$$

or

$$\sum_{G,\alpha} t_{G',\beta,G,\alpha} c_{G,\alpha} = E_n c_{G',\beta}, \tag{3.57}$$

with the matrix elements

$$t_{G',\beta,G,\alpha} := \int w^*_{G',\beta}(x) \left(-\frac{\hbar^2}{2m}\nabla^2 + V(x) \right) w_{G,\alpha}(x)\ \mathrm{d}x. \tag{3.58}$$

Eq. 3.57 is apparently just a discrete eigenvalue equation, where $t = (t_{G',\beta,G,\alpha})$ plays the role of the matrix. In fact, from the periodicity of $w_{G,\alpha}$ and $V(x)$ it follows that $t_{G',\beta,G,\alpha} = t_{\beta,\alpha}(G - G')$ and therefore

$$\sum_{G,\alpha} t_{\beta,\alpha}(G - G') c_{G,\alpha} = E_n c_{G',\beta}. \tag{3.59}$$

The matrix t thus consists of blocks as shown in Fig. 3.3 whose matrix norm satisfies

$$\parallel t_{\beta,\alpha}(G - G') \parallel \to 0 \quad \text{for } |G - G'| \to \infty, \tag{3.60}$$

since matrix elements of the Hamiltonian vanish if the Wannier functions are localized to unit cells that are far apart because the product of Wannier functions (and their derivative) vanishes in this limit. Therefore, blocks that are far away from the diagonal contribute less and can be neglected with good approximation.

The process of leaving out matrix elements between Wannier functions localized to unit cells that are far apart is referred to as the tight-binding approximation. As the name suggests the assumption made is that electrons are tightly bound to an atom (i.e., to a unit cell) and therefore the neglect of long-range matrix elements is justified.

Many systems can therefore be represented in terms of a rather small number of *hopping amplitudes* $t_{\beta,\alpha}(G)$, which are obviously much easier to handle than the full continuous lattice potential. The term *hopping* refers to the fact that t describes the matrix element of the Hamiltonian w.r.t. different sites and its modulus squared is proportional to the probability for a particle to hop from one to the other according to Feynman's golden rule. In order to compute the energies one first computes the Fourier transform

$$[H(k)]_{\beta,\alpha} = t_{\beta,\alpha}(k) = \sum_G e^{-ik\cdot G} t_{\beta,\alpha}(G), \tag{3.61}$$

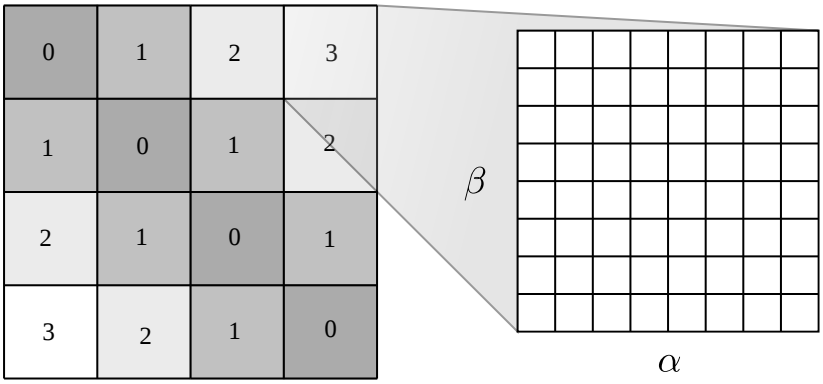

Figure 3.3: Hamiltonian matrix in the Wannier basis. The matrix consists of blocks, each labeled by the difference of lattice vectors $G - G'$. Due to the lattice translational symmetry the blocks with equal numbers are the same. Each block contains a matrix in band indices. Since the Wannier functions are localized, the norm of each block will tend to 0 for increasing distance of lattice vectors as indicated by the gray scale.

and obtains an equation

$$H(k)v_k = \epsilon(k)v_k, \tag{3.62}$$

where $H(k)$ is the so-called Bloch Hamiltonian and $\epsilon(k)$ the dispersion relation. Due to the unitarity of the Fourier transform Eq. 3.62 is equivalent to the original Schrödinger equation. In fact, Fourier transforming Eq. 3.59 yields

$$\frac{1}{\sqrt{N}} \sum_{G'} e^{-ik\cdot G'} \sum_{G,\alpha} t_{\beta,\alpha}(G - G')c_{G,\alpha} = E_n \frac{1}{\sqrt{N}} \sum_{G'} e^{-ik\cdot G'} c_{G',\beta} \tag{3.63}$$

$$\frac{1}{\sqrt{N}} \sum_{G'} e^{-ik\cdot (G'-G)} \sum_{G,\alpha} t_{\beta,\alpha}(G - G')e^{-ik\cdot G} c_{G,\alpha} = E_n c_{k,\beta} \tag{3.64}$$

$$\sum_{\alpha} \sum_{G'-G} e^{-ik\cdot (G'-G)} t_{\beta,\alpha}(G - G') \frac{1}{\sqrt{N}} \sum_{G} e^{-ik\cdot G} c_{G,\alpha} = E_n c_{k,\beta} \tag{3.65}$$

$$\sum_{\alpha} t_{\beta,\alpha}(k)c_{k,\alpha} = E_n c_{k,\beta}, \tag{3.66}$$

and by comparison of Eq. 3.66 with Eq. 3.62 we find that the dispersion relation is just a parameterization of the original energy eigenvalues. In the simplest case where the unit cell contains only one site with one orbital, Eq. 3.66 reduces to

$$t(k) = E_n = \epsilon(k), \tag{3.67}$$

i.e., the dispersion is the Fourier transform of the hopping matrix. One can illustrate the effect of the Fourier transform nicely as shown in Fig. 3.4, where we see that the block structure in lattice space has been reduced to a diagonal structure. This makes it fairly efficient to compute the eigenvalues, since the blocks in band space can be diagonalized separately for each k. If α, β correspond to the band indices then $H(k)$ is diagonal. We arrived at this definition by starting from the known Bloch states. We can easily obtain an alternate representation of the Bloch Hamiltonian in the basis of atomic orbitals that is related to the band basis through a unitary transformation by effectively defining the hopping matrix Eq. 3.58 through atomic orbitals.

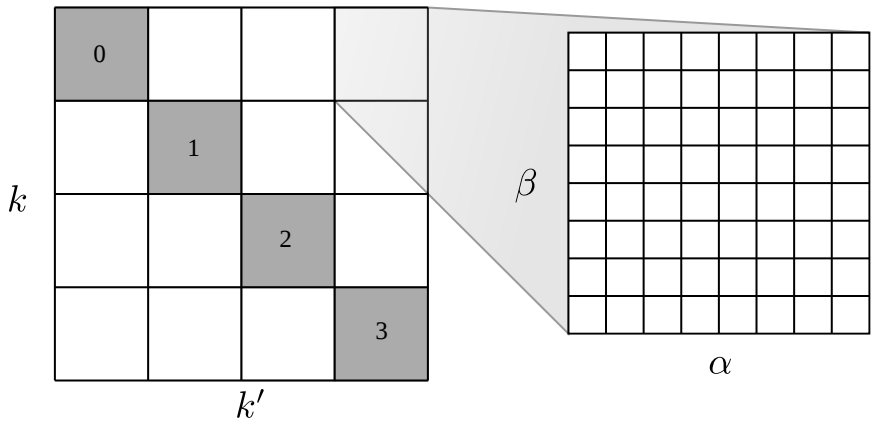

Figure 3.4: Illustration of the Hamiltonian matrix in the Bloch basis. The matrix is block diagonal in k allowing for a separate treatment of different momenta. Each block contains a matrix with site/orbital indices α, β. The diagonal blocks labeled by numbers correspond to different k.

Finally, we note that in the definition of $H(k)$ [Eq. 3.61] we performed the Fourier transform purely in the lattice space and with α, β corresponding to band indices the blocks in the Hamiltonian are diagonal, i.e. $\propto \delta_{\alpha\beta}$. In practice, one often has degeneracies between bands that result in a non-uniqueness of the Wannier functions. In this case the indices α, β of the maximally localized Wannier functions describe not the band index but sites within the unit cell that are displaced from the origin of the unit cell. In practice, two conventions for the treatment of these displacements exist. One as presented here and another that includes these displacement vectors δ_α as phase factors

$$[H(k)]_{\beta,\alpha} = \sum_G e^{-ik\cdot(G+\delta_\alpha-\delta_\beta)} t_{\beta,\alpha}(G). \tag{3.68}$$

This, however, corresponds merely to a unitary transformation as one can see from

$$\sum_G e^{-ik\cdot(G+\delta_\alpha-\delta_\beta)} t_{\beta,\alpha}(G) = e^{ik\cdot\delta_\beta} \sum_G e^{-ik\cdot G} t_{\beta,\alpha}(G) e^{-ik\cdot\delta_\alpha} = e^{ik\cdot\delta_\beta} [H(k)]_{\beta,\alpha} e^{-ik\cdot\delta_\alpha}, \tag{3.69}$$

which can be written equivalently as

$$e^{ik\cdot\delta_\beta} [H(k)]_{\beta,\alpha} e^{-ik\cdot\delta_\alpha} = U^\dagger(k) H(k) U(k), \tag{3.70}$$

with $U_{\alpha,\beta}(k) = e^{-ik\cdot\delta_\alpha}\delta_{\alpha\beta}$. Therefore, the dispersion obtained with any of the conventions is the same, only the eigenvectors differ by a unitary transformation.

Complexity Given an eigenvalue equation for a specific k, we have to solve

$$H(k) v_k = \epsilon(k) v_k, \tag{3.71}$$

with the Bloch Hamiltonian $H(k)$ and the unknown vector v_k. We can write equivalently

$$(H(k) - \epsilon(k) I)\, v_k = 0, \tag{3.72}$$

which, given $\epsilon(k)$, represents a homogeneous system of equations to be solved in $\mathcal{O}(n^3)$. In practice, both $\varepsilon(k)$ and v_k are unknown, which makes the computation more complicated. In addition to solving the system of equations, one has to find roots of the characteristic polynomial

$$P_c = \det(H(k) - \varepsilon(k) I) \stackrel{!}{=} 0. \tag{3.73}$$

There are more efficient algorithms, though, which rely on iterative procedures that perform the Schur decomposition of $H(k) = QUQ^{-1}$ [83], where Q is a unitary matrix and U an upper triangular matrix. The eigenvalues of the original matrix can be simply extracted from the diagonal of U. The complexity of the algorithm is $\sim n^3 + mn^2$, where m is the number of iterations and depends on the convergence speed for the particular matrix. See, e.g., Ref. [83] for more details.

3.2 Green's Functions

Here, we establish some notations and definitions for Green's functions and provide the essential equations that are required later on. For more information we refer to, e.g., the standard textbooks Refs. [84–86]. In condensed matter theory, the Green's function is defined as

$$G_{\alpha\beta}(x, x') = -i\langle \mathcal{T}\psi_\alpha(x)\psi^\dagger_\beta(x')\rangle, \tag{3.74}$$

where $\psi, \psi^\dagger$ are the fermionic field operators. Here and in the following discussion we use the notation $x = (t, \mathbf{r})$ that combines time and space into one vector. The expectation value is taken with respect to the partition function, i.e.,

$$\langle\cdot\rangle = \frac{\mathrm{tr}\left(e^{-\beta H}\cdot\right)}{Z}, \tag{3.75}$$

which in the limit $\beta \to \infty$ ($T \to 0$) equates to

$$\langle\cdot\rangle = \langle 0|\cdot|0\rangle, \tag{3.76}$$

with $|0\rangle$ being the ground state (assuming here that it is non-degenerate). $\mathcal{T}$ is the time ordering operator, whose action is defined as

$$\mathcal{T}\left[\psi(x)\psi^\dagger(x')\right] = \begin{cases} \psi(x)\psi^\dagger(x'), & \text{if} \quad x_0 > x'_0 \\ -\psi^\dagger(x')\psi(x), & \text{if} \quad x'_0 > x_0. \end{cases} \tag{3.77}$$

In addition to ensuring that earlier times are applied first to the right, $\mathcal{T}$ also respects the anti-commutation relations

$$\{\psi_\alpha(x), \psi_\beta(x')\} = 0, \quad \{\psi_\alpha(t, \mathbf{r}), \psi^\dagger_\beta(t, \mathbf{r})\} = \delta_{\alpha\beta}\delta(\mathbf{r} - \mathbf{r}'). \tag{3.78}$$

Apparently, the Green's function acts as a probe for studying the effect of adding an electron to the system, since it is essentially the overlap of a state with an additional electron added at $t = 0, \mathbf{r} = 0$ and one where the electron is added at a later time t at $\mathbf{r}$. As a function of t and $\mathbf{r}$, G encodes the likelihood that the added electron ends up at $\mathbf{r}$ after time t.

Inserting the definition of the time ordering operator into Eq. 3.74, we obtain the equivalent but more useful equation

$$G_{\alpha\beta}(x, x') = -i\left[\theta(t - t')\langle\psi_\alpha(x)\psi^\dagger_\beta(x')\rangle - \theta(t' - t)\langle\psi^\dagger_\beta(x')\psi_\alpha(x)\rangle\right], \tag{3.79}$$

where $\theta(t)$ is the Heaviside step function. We can use this expression to compute the equation of motion

$$\begin{aligned} \frac{\partial}{\partial t}G(x, x') = -i\Big[&\delta(t - t')\langle\psi_\alpha(x)\psi^\dagger_\beta(x')\rangle + \theta(t - t')\frac{\partial}{\partial t}\langle\psi_\alpha(x)\psi^\dagger_\beta(x')\rangle \\ &+\delta(t' - t)\langle\psi^\dagger_\beta(x')\psi_\alpha(x)\rangle - \theta(t' - t)\frac{\partial}{\partial t}\langle\psi^\dagger_\beta(x')\psi_\alpha(x)\rangle\Big]. \end{aligned} \tag{3.80}$$

The time derivatives still need to be evaluated. Since $x = (t, \mathbf{r}), x' = (t', \mathbf{r}')$, the time derivative acts only on x, such that we can insert the equation of motion for the field operators. For the following derivation we collect a few identities involving commutators

$$[A, BC] = B[A, C] + [A, B]C \tag{3.81}$$
$$[A, BC] = ABC - BCA + BAC - BAC = \{A, B\}C - B\{A, C\}. \tag{3.82}$$

Using the identities

$$\begin{aligned}[\psi_\alpha(x), \psi_\beta(x')] &= \psi_\alpha(x)\psi_\beta(x') - \psi_\beta(x')\psi_\alpha(x) \\ &= 2\psi_\alpha(x)\psi_\beta(x'),\end{aligned} \tag{3.83}$$

$$\begin{aligned}[\psi_\alpha(t, \mathbf{r}), \psi^\dagger_\beta(t, \mathbf{r}')] &= \psi_\alpha(t, \mathbf{r})\psi^\dagger_\beta(t, \mathbf{r}') - \psi^\dagger_\beta(t, \mathbf{r}')\psi_\alpha(t, \mathbf{r}) \\ &= 2\psi_\alpha(t, \mathbf{r})\psi^\dagger_\beta(t, \mathbf{r}') - \delta(\mathbf{r} - \mathbf{r}')\delta_{\alpha,\beta},\end{aligned} \tag{3.84}$$

that follow from the fermionic anti-commutation relations of Eq. 3.78 we obtain

$$[\psi_\alpha(t, \mathbf{r}), \psi_\gamma(t, \mathbf{r}')\psi^\dagger_\beta(t, \mathbf{r}')] = -\delta(\mathbf{r} - \mathbf{r}')\delta_{\alpha,\beta}\psi_\gamma(t, \mathbf{r}'), \tag{3.85}$$

$$[\psi_\alpha(t, \mathbf{r}), \psi^\dagger_\gamma(t, \mathbf{r}')\psi_\beta(t, \mathbf{r}')] = \delta(\mathbf{r} - \mathbf{r}')\delta_{\alpha,\gamma}\psi_\beta(t, \mathbf{r}'). \tag{3.86}$$

With the Heisenberg equation of motion

$$i\hbar\frac{\partial\psi(x)}{\partial t} = [\psi(x), H - \mu N], \tag{3.87}$$

and

$$H - \mu N = \sum_{\alpha,\beta}\int \mathrm{d}^3x\ \psi^\dagger_\alpha(x)[H_{\alpha\beta}(x) - \delta_{\alpha\beta}\mu]\psi_\beta(x) \tag{3.88}$$

we then have for the field operator

$$i\hbar\frac{\partial\psi_\alpha(x)}{\partial t} = \sum_\beta[H_{\alpha\beta}(x) - \delta_{\alpha,\beta}\mu]\psi_\beta(x). \tag{3.89}$$

This allows us to express the derivatives of the expectation values in Eq. 3.80 as

$$\frac{\partial}{\partial t}\langle\psi_\alpha(x)\psi^\dagger_\beta(x')\rangle = \frac{1}{i\hbar}\sum_\gamma[H_{\alpha\gamma}(x) - \delta_{\alpha,\gamma}\mu]\langle\psi_\gamma(x)\psi^\dagger_\beta(x')\rangle \tag{3.90}$$

and similarly for the reverse ordering. Therefore, we have

$$\begin{aligned}i\hbar\frac{\partial}{\partial t}G_{\alpha\beta}(x, x') &= \sum_\gamma[H_{\alpha\gamma}(x) - \delta_{\alpha,\gamma}\mu]G_{\gamma\beta}(x, x') \\ &\quad + \hbar\delta(t - t')\left[\langle\psi_\alpha(x)\psi^\dagger_\beta(x')\rangle + \langle\psi^\dagger_\beta(x')\psi_\alpha(x)\rangle\right] \qquad (3.91) \\ &= \sum_\gamma[H_{\alpha\gamma}(x) - \delta_{\alpha,\gamma}\mu]G_{\gamma\beta}(x, x') + \hbar\delta(x - x')\delta_{\alpha\beta}, \qquad (3.92)\end{aligned}$$

where we used the anti-commutation relation of Eq. 3.78. In matrix notation, introducing the grand canonical Hamiltonian $K = H - \mu N$, this can be written as

$$\left(i\hbar\frac{\partial}{\partial t} - K\right)G(x, x') = \hbar\delta(x - x'). \tag{3.93}$$

Eq. 3.93 immediately explains the name "Green's function" for the definition of Eq. 3.74, since G really is the mathematical Green's function for the Schrödinger operator $i\hbar\partial_t - K$. G is sometimes also called "propagator" due to its relation to the time-evolution operator, which takes an initial state and propagates it in time.

In thermodynamic equilibrium, propagation cannot depend on when it is initiated, since the Hamiltonian is time-independent. Therefore, the Green's function must depend only on the difference $t - t'$ between final and initial time. For translation invariant systems the same argument applies in terms of spatial coordinates, i.e., we have the identity

$$G(t, \mathbf{r}; t', \mathbf{r}') = G(t + t_0, \mathbf{r} + \mathbf{R}; t' + t_0, \mathbf{r}' + \mathbf{R}) \tag{3.94}$$

$\forall t_0$ and lattice vectors $\mathbf{R}$, so, in particular, also for $t_0 = -t', \mathbf{R} = -\mathbf{r}'$ which implies that

$$G(t, \mathbf{r}; t', \mathbf{r}') = G(t - t', \mathbf{r} - \mathbf{r}'; 0, 0) =: G(t - t', \mathbf{r} - \mathbf{r}'). \tag{3.95}$$

Therefore, the Green's function depends only on one time and position argument. We define the Fourier transform as

$$G(x) = G(t, \mathbf{r}) = \frac{1}{2\pi N} \sum_{\mathbf{k}} \int_{-\infty}^{\infty} \mathrm{d}\omega \; e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)} G(\omega, \mathbf{k}), \tag{3.96}$$

where

$$G(\omega, \mathbf{k}) = \int_{-\infty}^{\infty} \mathrm{d}t \int \mathrm{d}\mathbf{r} \; e^{-i(\mathbf{k}\cdot\mathbf{r} - \omega t)} G(t, \mathbf{r}). \tag{3.97}$$

Inserting the relation from Eq. 3.96 into Eq. 3.93 we obtain

$$\begin{aligned} &\frac{1}{2\pi N} \int_{-\infty}^{\infty} \sum_{\mathbf{k}} \left(i\hbar \frac{\partial}{\partial t} - K \right) e^{i(\mathbf{k}\cdot\mathbf{R} - \omega t)} G(\omega, \mathbf{k}) \; \mathrm{d}\omega \\ &\quad = \frac{1}{2\pi N} \int_{-\infty}^{\infty} \sum_{\mathbf{k}} \left(\hbar\omega - \varepsilon_{\mathbf{k}} + \mu \right) e^{i(\mathbf{k}\cdot\mathbf{R} - \omega t)} G(\omega, \mathbf{k}) \; \mathrm{d}\omega \qquad (3.98) \\ &\quad = \hbar\delta(x - x'). \qquad (3.99) \end{aligned}$$

Integrating this over $\mathbf{r}$ and t yields an expression for the Green's function in frequency-momentum space

$$G(\omega, \mathbf{k}) = \frac{1}{\hbar\omega + \mu - \varepsilon_{\mathbf{k}}}. \tag{3.100}$$

In addition to the time-ordered Green's function, one typically defines the so-called retarded Green's function as

$$G^R(t, x; t', x') = -i\theta(t - t') \langle \{ \psi(x, t), \psi^\dagger(x', t') \} \rangle. \tag{3.101}$$

In contrast to the time-ordered function the retarded function contains both terms $\psi\psi^\dagger$ and $\psi^\dagger\psi$, where the former amounts to inserting a particle and removing it at a later time, while the latter inserts a hole. As a result, also hole excitations are taken into account. The name "retarded" is motivated by the causality-respecting theta function that assures that the cause precedes the effect.

Quite generally we can define a Green's function in any basis $\{\alpha\}$ via the transformation of the fermionic operators. Let

$$c_\alpha = \sum_x \phi_\alpha(x)\psi(x), \quad c_\alpha^\dagger = \sum_x \phi_\alpha^*(x)\psi^\dagger(x), \tag{3.102}$$

which follows from the identity

$$c_\alpha^\dagger|\text{vac}\rangle = |\alpha\rangle = \int \mathrm{d}x|x\rangle\langle x|\alpha\rangle = \int \mathrm{d}x\phi_\alpha(x)|x\rangle = \int \mathrm{d}x\phi_\alpha(x)\psi^\dagger(x)|\text{vac}\rangle \tag{3.103}$$

and the hermitian conjugate

$$c_\alpha|\text{vac}\rangle = \int \mathrm{d}x\phi_\alpha^*(x)\psi(x)|\text{vac}\rangle. \tag{3.104}$$

Then, the retarded Green's function can be expressed as

$$\begin{aligned} G^R(t,x;t',x') &= -i\theta(t-t')\langle\{\psi(x,t),\psi^\dagger(x',t')\}\rangle & (3.105)\\ &= -i\theta(t-t')\sum_{\alpha,\alpha'}\phi_\alpha(x)\phi_{\alpha'}^*(x)\langle\{c_\alpha(t),c_{\alpha'}^\dagger(t')\}\rangle & (3.106)\\ &=: \sum_{\alpha,\alpha'}\phi_\alpha(x)\phi_{\alpha'}^*(x)G^R(t,\alpha;t',\alpha'). & (3.107)\end{aligned}$$

With $\phi_k(x) = e^{ikx}$ we obtain the momentum-space Green's function

$$\begin{aligned} G^R(t,k;t',k') &= \int \mathrm{d}x\int \mathrm{d}x' e^{i(k'x'-kx)}G^R(t,x;t',x') & (3.108)\\ &= \int \mathrm{d}x\int \mathrm{d}x' e^{i(k'-k)x'}e^{-ik(x-x')}G^R(t,x;t',x') & (3.109)\\ &= \int \mathrm{d}x\int \mathrm{d}x' e^{i(k'-k)x'}e^{-ik(x-x')}G^R(t,x-x';t',0) & (3.110)\\ &= \int \mathrm{d}x' e^{i(k'-k)x'}\int \mathrm{d}y e^{-iky}G^R(t,y;t',0) & (3.111)\\ &= \delta_{k,k'}\int \mathrm{d}y e^{-iky}G^R(t,y;t',0), & (3.112)\end{aligned}$$

that is indeed diagonal in k, i.e., it can be expressed as $G^R(t,k;t',k') =: G^R(t,t',k)$.

3.2.1 Spectral Representation

By evaluating the expectation value in the definition of the retarded Green's function we can derive a convenient representation that can be evaluated if the many-body ground state is known. We start from the definition

$$G_{\alpha\beta}^R(t-t') = -i\theta(t-t')\langle\{c_\alpha(t),c_\beta^\dagger(t')\}\rangle, \tag{3.113}$$

and insert the thermal average for the expectation value

$$G_{\alpha\beta}^R(t-t') = -i\theta(t-t')\frac{1}{Z}\sum_n e^{-\beta(E_n^N-\mu N)}{}_N\langle n|\{c_\alpha(t),c_\beta^\dagger(t')\}|n\rangle_N, \tag{3.114}$$

where $|n\rangle_N$ are eigenstates of H with particle number N and eigenvalues E_n^N. By inserting an identity in the many-body basis and the time evolution operators we can simplify

$$\begin{aligned}
G^R_{\alpha\beta}(t-t') &= -i\theta(t-t')\frac{1}{Z}\sum_{n,m} e^{-\beta(E_n^N-\mu N)}\Big[{}_N\langle n|e^{i(H-\mu N)t}c_\alpha e^{-i(H-\mu N)t}|m\rangle_{N+1} \\
&\quad\times {}_{N+1}\langle m|e^{i(H-\mu N)t'}c_\beta^\dagger e^{-i(H-\mu N)t'}|n\rangle_N + \text{c.c.}\Big] \\
&= -i\theta(t-t')\frac{1}{Z}\sum_{n,m} e^{-\beta(E_n^N-\mu N)}\Big[{}_N\langle n|e^{i(E_n^N-\mu N)t}c_\alpha e^{-i(E_m^{N+1}-\mu(N+1))t}|m\rangle_{N+1} \\
&\quad\times {}_{N+1}\langle m|e^{i(E_m^{N+1}-\mu(N+1))t'}c_\beta^\dagger e^{-i(E_n^N-\mu N)t'}|n\rangle_N + \text{c.c.}\Big] \\
&= -i\theta(t-t')\frac{1}{Z}\sum_{n,m} e^{-\beta(E_n^N-\mu N)}\Big[e^{i(E_n^N-E_m^{N+1}+\mu)(t-t')}\langle n|c_\alpha|m\rangle\langle m|c_\beta^\dagger|n\rangle + \text{c.c.}\Big],
\end{aligned}$$

where states $|m\rangle = |m\rangle_{N+1}$ and energies E_m^{N+1} correspond to a system with an additional particle with respect to $|n\rangle = |n\rangle_N$. In the second term abbreviated by c.c. we insert the identity with one particle less so that the matrix elements of the creation and annihilation operators can be finite. Explicitly, the second term is

$$e^{i(E_n^N-E_m^{N-1}-\mu)(t'-t)}\,{}_N\langle n|c_\beta^\dagger|m\rangle_{N-1}\;{}_{N-1}\langle m|c_\alpha|n\rangle_N. \tag{3.115}$$

We now rename the indices n, m and shift the particle number in the second term, which is allowed since we sum over all of them, so that it becomes

$$e^{i(E_n^N-E_m^{N+1}+\mu)(t-t')}\langle m|c_\beta^\dagger|n\rangle\langle n|c_\alpha|m\rangle. \tag{3.116}$$

Note that we now have the same matrix elements and exponents in both terms. This step also requires us to rename the energy in the Boltzmann factor:

$$\begin{aligned}
G^R_{\alpha\beta}(t-t') = -i\theta(t-t')\frac{1}{Z}\sum_{n,m}&\Big[e^{-\beta(E_n^N-\mu N)} + e^{-\beta(E_m^{N+1}-\mu(N+1))}\Big] \\
&\times e^{i(E_n^N-E_m^{N+1}+\mu)(t-t')}\,{}_N\langle n|c_\alpha|m\rangle_{N+1}\;{}_{N+1}\langle m|c_\beta^\dagger|n\rangle_N.
\end{aligned} \tag{3.117}$$

In order to arrive at $G(\omega)$ we Fourier transform this expression

$$G^R_{\alpha\beta}(\omega) = \int_{-\infty}^{\infty} \mathrm{d}t\; e^{i\omega t}G^R_{\alpha\beta}(t). \tag{3.118}$$

The terms can be rearranged so that we arrive at a simple integral

$$G^R_{\alpha\beta}(\omega) = -i\frac{1}{Z}\sum_{n,m}\Big[e^{-\beta(E_n-\mu N)} + e^{-\beta(E_m-\mu(N+1))}\Big]\int_{-\infty}^{\infty}\mathrm{d}t\;\theta(t)e^{i(\omega+E_n-E_m+\mu)t} \tag{3.119}$$

$$\times\langle n|c_\alpha|m\rangle\langle m|c_\beta^\dagger|n\rangle \tag{3.120}$$

$$= -i\frac{1}{Z}\sum_{n,m}\Big[e^{-\beta(E_n-\mu N)} + e^{-\beta(E_m-\mu(N+1))}\Big] \tag{3.121}$$

$$\times\int_0^{\infty}\mathrm{d}t\; e^{i(\omega+E_n-E_m+\mu)t}\Big]\langle n|c_\alpha|m\rangle\langle m|c_\beta^\dagger|n\rangle. \tag{3.122}$$

The time integrals do not converge since the limit of $e^{i\omega t}$ for $t \to \infty$ does not exist. We fix this by adding a small positive imaginary part to ω, i.e., $\omega \mapsto \omega + i\eta$, which guarantees that the exponent has a negative real part so that the integral converges:

$$\int_0^\infty \mathrm{d}t\ e^{i(\omega+i\eta+E_n-E_m)t} = \left[\frac{e^{i(\omega+i\eta+E_n-E_m)t}}{i(\omega+i\eta+E_n-E_m+\mu)}\right]_0^\infty \tag{3.123}$$

$$= \frac{i}{\omega+i\eta+E_n-E_m+\mu}. \tag{3.124}$$

Therefore, the retarded Green's function becomes

$$G^R_{\alpha\beta}(\omega) = \frac{1}{Z}\sum_{n,m}\left[e^{-\beta(E_n-\mu N)} + e^{-\beta(E_m-\mu(N+1))}\right]\frac{\langle n|c_\alpha|m\rangle\langle m|c^\dagger_\beta|n\rangle}{\omega+i\eta+\mu+E_n-E_m}. \tag{3.125}$$

The $T = 0$ result can be obtained similarly. Unfortunately, just removing the Boltzmann factors and Z and replacing the sum over $|n\rangle$ with the ground state is not enough, since the renaming of the labels n, m is no longer possible if we remove the sum over n. Therefore, we have to keep the form of Eq. 3.115. An analogous derivation then yields

$$G^R_{\alpha\beta}(\omega) = -i\Big[\sum_m \int_{-\infty}^\infty \mathrm{d}t\ \theta(t)e^{i(\omega+\mu+E_0^N-E_m^{N+1})t}\langle 0|c_\alpha|m\rangle\langle m|c^\dagger_\beta|0\rangle \tag{3.126}$$

$$+\sum_m \int_{-\infty}^\infty \mathrm{d}t\ \theta(t)e^{i(\omega+\mu+E_m^{N-1}-E_0^N)t}\langle 0|c^\dagger_\beta|m\rangle\langle m|c_\alpha|0\rangle\Big] \tag{3.127}$$

$$= \sum_m\left[\frac{\langle \mathrm{gs}|c_\alpha|m\rangle\langle m|c^\dagger_\beta|\mathrm{gs}\rangle}{\omega+i\eta+\mu+E_0^N-E_m^{N+1}} + \frac{\langle \mathrm{gs}|c^\dagger_\beta|m\rangle\langle m|c_\alpha|\mathrm{gs}\rangle}{\omega+i\eta+\mu+E_m^N-E_0^N}\right]. \tag{3.128}$$

3.2.2 Interacting Problem, Self-Energy & Dyson Equation

For the interacting problem some of the previous equations need to be modified, since the Hamiltonian contains non-quadratic terms. In general we have the lattice Hamiltonian

$$H = H_0 + V = \sum_{i,j} t_{ij}c^\dagger_i c_j + \frac{1}{2}\sum_{i_1,i_2,j_1,j_2} V_{i_1i_2j_1j_2}c^\dagger_{i_1}c^\dagger_{i_2}c_{j_2}c_{j_1}, \tag{3.129}$$

where the indices i, j contain all degrees of freedom including spin. The additional quartic term leads to a modified equation of motion for the Green's function, since the fermionic operator now satisfies

$$i\hbar\frac{\partial c_\alpha(t)}{\partial t} = \sum_j [H_{\alpha j} - \delta_{\alpha j}\mu]c_j(t) + \sum_{i_1,j_1,j_2} V_{\alpha i_1 j_1 j_2}c^\dagger_{i_1}c_{j_2}c_{j_1}. \tag{3.130}$$

For the Green's function we obtain

$$\begin{aligned} i\hbar\partial_t G_{\alpha\beta}(t;t') - [H_{\alpha\gamma} - \delta_{\alpha\gamma}\mu]G_{\alpha\gamma}(t;t') &= \delta(t-t')\delta_{\alpha\beta} \\ &\quad - i\sum_{i_1,j_1,j_2} V_{\alpha i_1 j_1 j_2}\langle \mathcal{T}c^\dagger_{i_1}(t)c_{j_2}(t)c_{j_1}(t)c^\dagger_\beta(t')\rangle, \end{aligned} \tag{3.131}$$

where the last term is the only addition compared to Eq. 3.93. In a similar matrix notation we can write this result as

$$[i\hbar\partial_t - (H-\mu)]G(t;t') = \delta(t-t') + \int \mathrm{d}t''\ \Sigma(t;t'')G(t'';t'), \tag{3.132}$$

or in terms of frequency

$$G(\omega) = \frac{1}{\hbar\omega + \mu - H - \Sigma(\omega)}, \tag{3.133}$$

where the so-called self-energy $\Sigma(\omega)$ contains all effects of the interaction. Implicitly, the above equation relates the full Green's function to the non-interacting Green's function G_0 through

$$G^{-1} = G_0^{-1} - \Sigma. \tag{3.134}$$

Eq. 3.134 is called Dyson's equation. Multiplication from the left with G_0 and from the right with G and rearranging the terms yields

$$G = G_0 + G_0 \Sigma G, \tag{3.135}$$

which defines the so-called Dyson series through continued insertion of the equation on the right hand side.

3.3 Exact Diagonalization

The most straight-forward way to solve the many-body problem is to write down the full Hamiltonian in a suitable basis and obtain all eigenstates and eigenvalues through the diagonalization of the matrix. This method is exact, since it provides the exact ground state of the problem and therefore all observables can be computed without the need for approximations. On the other hand, the dimension of the Hilbert space is 4^N, where N is the number of single particle states. This exponential scaling unfortunately reduces the applicability of the exact diagonalization scheme to small clusters of size $N \sim \mathcal{O}(10)$. Nonetheless, one can approximate the solution of the full problem with that of a small cluster.

In order to diagonalize the Hamiltonian one first has to determine its matrix elements. For a given model or material the hopping matrix t_{ij} needs to be known. For simplicity, we assume here a Hubbard interaction [87]

$$H = \sum_{ij} t_{ij} c_i^\dagger c_j + U \sum_i n_{i\uparrow} n_{i\downarrow}. \tag{3.136}$$

We compute the matrix elements in the occupation number basis, which is defined by

$$|n_0, n_1, n_2, \ldots\rangle = \prod_i (c_i^\dagger)^{n_i} |\text{vac}\rangle, \tag{3.137}$$

where n_i are the occupation numbers of single particle states $|i\rangle$. We find immediately that the matrix element

$$\langle \vec{n}|T_{ij}|\vec{m}\rangle = t_{ij}\langle \vec{n}|c_i^\dagger c_j|\vec{m}\rangle \tag{3.138}$$

of the kinetic energy operator T in this basis can only be nonzero if $m_j = 1$ and $n_i = 0$, since $c_i|\ldots, n_i = 0, \ldots\rangle = c_i^\dagger|\ldots, n_i = 1, \ldots\rangle = 0$. Assuming this is the case we take a look at $c_j|\vec{m}\rangle$, which can be written as

$$c_j|\vec{m}\rangle = c_j \prod_i (c_i^\dagger)^{m_i}|\text{vac}\rangle \tag{3.139}$$

$$= (-1)^{\sum_{i=0}^{j-1} m_i} \prod_{i=0}^{j-1} (c_i^\dagger)^{m_i} c_j \prod_{i=j}^{N} (c_i^\dagger)^{m_i}|\text{vac}\rangle \tag{3.140}$$

$$= (-1)^{\sum_{i=0}^{j-1} m_i}|m_0, m_1, \ldots, m_{j-1}, 0, m_{j+1}, \ldots\rangle, \tag{3.141}$$

where we used $\{c_i, c_j^\dagger\} = \delta_{ij}$. Since $\langle\vec{n}|c_i^\dagger = (c_i|\vec{b}\rangle)^\dagger$, we obtain an analogous result for the other side. In total, the matrix element for arbitrary basis states $|\vec{n}\rangle, |\vec{m}\rangle$ can then be written as

$$t_{ij}^{-1}\langle\vec{n}|T_{ij}|\vec{m}\rangle = \delta_{n_i,0}\delta_{m_j,1}\delta_{n_j,1}\delta_{m_i,}\prod_{k\neq i,j}\delta_{n_k,m_k}(-1)^{\sum_{i=0}^{j-1}m_i}(-1)^{\sum_{k=0}^{i-1}n_k} \tag{3.142}$$

$$= \delta_{n_i,0}\delta_{m_j,1}\delta_{n_j,1}\delta_{m_i,}\prod_{k\neq i,j}\delta_{n_k,m_k}(-1)^{\sum_{i=0}^{j-1}m_i+\sum_{k=0}^{i-1}n_k} \tag{3.143}$$

$$= \delta_{n_i,0}\delta_{m_j,1}\delta_{n_j,1}\delta_{m_i,}\prod_{k\neq i,j}\delta_{n_k,m_k}(-1)^{2\sum_{i=0}^{\min(\{j-1,i-1\})}n_i+\sum_{k=\min(\{j,i\})+1}^{\max(\{j-1,i-1\})}n_k} \tag{3.144}$$

$$= \delta_{n_i,0}\delta_{m_j,1}\delta_{n_j,1}\delta_{m_i,}\prod_{k\neq i,j}\delta_{n_k,m_k}\begin{cases}(-1)^{\sum_{k=j+1}^{i-1}n_k} & \text{if} \quad i>j\\ 1 & \text{if} \quad i=j\\ (-1)^{\sum_{k=i+1}^{j-1}m_k} & \text{if} \quad i<j.\end{cases} \tag{3.145}$$

Note that the Kronecker deltas imply $m_j = n_i = 1$ and $n_j = m_i = 0$, so that the sum over k skips the first term. We can summarize that the matrix element of a single hopping term contributes t_{ij} times a sign if the sum of occupied states between i and j is odd, provided that the basis states differ only in two positions i, j with $m_j = 1$ and $m_i = 0$.

Seeing that the vast majority of matrix elements vanishes it seems rather inefficient to determine H with an algorithm of the following kind:

```
for k=1 to N
{
  for l=1 to N
  {
    n_i = calc_occupation_at(k, i)
    n_j = calc_occupation_at(k, j)
    m_i = calc_occupation_at(l, i)
    m_j = calc_occupation_at(l, j)

    tmp = swap_occupation(swap_occupation(k, i), j)
    s = calc_occupation_between(k, i, j)
    H[k, l] = (tmp==l) * m_j * (m_i-1)* n_i * (n_j-1) * (-1)**s * t[i,j]
  }
}
```

where k, l label the many-body states, N is the dimension of the Hilbert space and we assume an additional outer loop over quantum numbers i, j. While this is certainly the most straightforward way to build a matrix, the problem with this approach is the double loop over the dimension of the Hilbert space, i.e., every matrix element is visited while no consideration is given to the fact that the matrix is extremely sparse because the number of nonzero matrix elements is much smaller than the matrix size. In particular,

$$\#\text{nonzero} \sim \mathcal{O}(N(\log N)^2). \tag{3.146}$$

We can show this by considering the hopping matrix elements t_{ij} of which there are at most $(\log_2 N)^2$. For each one we have two occupations fixed and the remaining $\log_2 N - 2$ can take arbitrary values in $\{0, 1\}$, which makes for a total of $2^{\log_2 N-2} = N/4$ combinations. Therefore, we end up with a worst case of $N(\log_2 N)^2/4$. Usually, each site is only connected to few neighbors and therefore the number of finite matrix elements is practically $\mathcal{O}(N)$. Here, we

assumed that the total Hilbert space is taken into account, while usually one restricts to only a subspace with a particular filling. In that case the occupation numbers cannot be freely distributed and for n single particle orbitals and $k \leq n-2$ zeros we have $\binom{n-2}{k}$ permutations. The worst case is at half filling with $k = n/2 - 1$

$$\binom{n-2}{n/2-1} = \frac{(n-2)!}{(n/2-1)!(n/2-1)!} \approx \frac{\sqrt{n-2}}{n/2-1} \frac{(n-2)^{n-2} e^{n-2}}{e^{n-2}(n/2-1)^{n-2})} \tag{3.147}$$

$$= \frac{2^n}{2\sqrt{n-2}}, \tag{3.148}$$

where we used Stirling's approximation. With $n \sim \log_2 N$ we find a similar relation as Eq. 3.146. Going away from half filling the number of states determined by the number of permutations of 0s and 1s becomes much smaller, and approaches $\log_2 N$. It may be more natural to write Eq. 3.146 in terms of the number of spin-orbitals n as

$$\#\text{nonzero} \sim \mathcal{O}(4^n n^2). \tag{3.149}$$

This is surely much smaller than 8^n, which is the complexity of the algorithm presented earlier.

A more efficient algorithm can then be formulated using only a single loop (again, the loops over i, j are assumed to be external):

```
for k=1 to N
{
  n_i = calc_occupation_at(k, i)
  n_j = calc_occupation_at(k, j)

  m = swap_occupation(k, i, j)
  s = calc_occupation_between(k, i, j)
  H[m, k] = n_i * (1-n_j) * (-1)**s * t[i,j]
}
```

In practice, it is even convenient to loop only over the pairs (i, j) with $i \leq j$ and nonzero t_{ij} and immediately add the complex conjugate (for $i \neq j$).

For the matrix elements of density-density interactions no sign appears, since all basis vectors are eigenstates of these operators. One can therefore add all of these matrix elements in a single loop over all N many-body basis states. The total complexity of the construction of the many-body Hamiltonian is therefore $\mathcal{O}(N)$, which is drastically smaller than the number of matrix elements. For exchange interaction operators of the form

$$c^\dagger_{i\uparrow} c_{i\downarrow} c^\dagger_{j\downarrow} c_{j\uparrow} \quad \text{and} \quad c^\dagger_{i\uparrow} c^\dagger_{i\downarrow} c_{j\downarrow} c_{j\uparrow} \tag{3.150}$$

a similar computation as for the hopping terms can be performed and one obtains a sign determined by the number of occupations between $i\uparrow, j\uparrow$ and $i\downarrow, j\downarrow$, respectively.

3.3.1 Memory representation of integers

For the efficient storage and manipulation of fermionic Fock states it is very convenient to make use of the representation of integer numbers in memory. A fermionic state is a superposition of anti-symmetrized product states. In the occupation number basis the basis states can be parameterized by the occupation numbers of single particle states

$$|e_i\rangle = |n^i_0, n^i_1, n^i_2, \ldots\rangle. \tag{3.151}$$

Figure 3.5: Bit order of unsigned integer numbers on different architectures. The memory pointer is illustrated by an arrow pointing to the first bit that is read. Left: big-endian, i.e., the first bit accessed is the one corresponding to the largest value. Right: little-endian, the least significant bit is stored in the first position. The difference is irrelevant for usual integer and floating-point arithmetic. However, using algorithms that operate on the integers' bits directly requires awareness of the system's architecture.

Due to the anti-symmetry, each n_j^i can only take values 0 or 1. The entire state can therefore be written as a bitstring of the form, e.g., `010001`$= |0,1,0,0,0,1\rangle$.

Computers work with binary arithmetics, i.e., for each unit of storage, the Bit, only two values 0 and 1 are available. The direct representation of a decimal number is therefore impossible. One can instead transform the number into its binary equivalent. In a particular basis b, every non-negative integer number n can be represented as a finite sequence of digits d_i

$$n = d_{N-1}d_{N-2}\dots d_1 d_0 = \sum_i d_i b^i, \tag{3.152}$$

where $d_i < b\ \forall i$. Hence, considering the case of binary representation we have $d_i \in \{0,1\}$, such that the integer is simply a sequence of bits. Instead of storing a list of values for each basis state it is therefore sufficient to store only one integer number.

In order to do manipulations of quantum states it is necessary to assign certain bits to single particle states. This procedure is straight-forward on paper, since we can simply enumerate the bits as in Eq. 3.152. However, the order of bits in memory depends on the computer architecture. Most office computers and many compute clusters run processors of the x86 architecture. The architecture defines a set of instructions that the processor can perform. Originally, x86 was introduced by Intel for its 8086 processor, released in 1978. Over time things such as the address width and the size of the registers have been increased from initially 16 to now 64 bits to accommodate demands for larger computations.

When storing a bit string in memory one is faced with a choice between essentially two equal possibilities. Either one puts the zeroth digit that contains the smallest power of the base on the first or the last bit. The first case is referred to as little-endian, the second as big-endian, since the numbering of bits in each byte starts from the least/most significant bit, cf. Fig. 3.5.

All x86 systems, such as the modern Intel 64 and AMD64, are little endian, which means that the memory address of the integer points to the least significant bit, cf., e.g., [88]. Since memory addresses are only allocated to bytes not bits one has to use boolean arithmetic to obtain information about the individual bits. The value of, e.g., the fourth bit is obtained with `n AND 0x8`.

A 32 bit or 4 byte unsigned integer can represent a basis state with up to 16 single particle states (two spins). Here, the limitation of simulating quantum physics on classical computers becomes apparent, since the dimension of the Hilbert space scales exponentially with the size of the system, i.e., the number n of single particle states. While each basis state can be represented in terms of a sequence of n bits the coefficients require $N = 4^n$ complex numbers. Storing such a vector quickly approaches the limitations of any computer as for, e.g., $n = 100$ one has $N \approx 10^{60}$, which amounts to $\sim 10^{50}$GB. The dimension of the Hilbert space can be reduced significantly by considering symmetries and therefore decomposing the Hamiltonian into a block-diagonal matrix, where each block corresponds to a particular symmetric subspace. In practice, however, this approach is also limited to small system sizes of up to ~ 50 for spin systems [89, 90], whose

Hilbert space scales only as 2^n. We expect that the simulation of large-scale quantum systems will become much simpler once quantum computers reach a certain maturity, since for a system of size n only n qubits are required. At the present time devices with 100 qubits already exist. Provided that a sufficient level of reliability can be achieved, much larger system sizes will become available for study in the near future.

3.3.2 Bitwise operations

We discuss here some interesting bit hacks that are commonly used in ED algorithms to determine the occupation numbers of Fock states from their integer representation.

Number of set bits The most trivial algorithm to count ones in a bit array is to first convert the integer to a string of 1s and 0s:

```
100110 => "100110"
```

Then, we can simply loop through the string and increase a counter each time the comparison with `"1"` returns true. With this approach we will lose most of what we gained with the efficient representation of Fock states, since strings themselves are typically stored as arrays of integers. In the following, we will discuss a much more efficient algorithm [91]. For simplicity we focus here on 8 Bit integers, which are defined, e.g., in C++ as the `unsigned char` type. We start by creating two copies of the integer `n`, one where only odd bits are copied from the original and all even bits are `0`

```
n1 = n & 01010101
```

and one with only even bits from `n`:

```
n2 = n & 10101010
```

Clearly, the total number of set bits in these two variables is still the same as that in `n`. If we now shift `n2` to the right by one bit we have the set bits in the same positions so that adding up the two copies will provide the number of the set bits in each block of two. We can prove this by constructing the algorithm in Table 3.1.

n	00	01	10	11
n1	00	01	00	01
n2	00	00	10	10
n2>>1	00	00	01	01
n1+(n2>>1)	00	01	01	10

Table 3.1: Method of counting bits in an 8 Bit unsigned integer `n`. The initial integer has 4 set bits. `n1` and `n2` are copies where only odd/even bits (counting from the right, i.e., least significant bit is number 1) are copied over, respectively. Right-shifting `n2` moves all the bits to odd positions. Addition with `n1` then yields the total number of set bits in each block of two bits, since `01`=1 and `10`=2.

This algorithm is done at this point for 2 Bit integers and can be generalized to larger integers by interpreting the full bit array as a sequence of two-bit arrays. Applying the above algorithm once returns the number of set bits in each array that we have to add up to obtain the final count. Instead of adding up single bits we now want to add up two-bit integers. We proceed again like before and define

```
n1 = n & 00110011
n2 = n & 11001100
n = n1 + (n2 >> 2)
```

where now only odd/even pairs are copied, respectively. By freeing up every even bit pair we have effectively four bits to store the result of the sum, which itself is facilitated by the bit shift by 2. We then arrive at a structure that contains a sequence of four bit blocks containing the counts of set bits in the corresponding positions in the original integer. The goal is now clear: Iterate this until we reach the breaking condition, that is, until the length of the sequence reaches 1. For a 32 Bit integer this takes exactly five steps, since in every step the size of the list entries is doubled at constant total number of bits. This means the length of the list is divided by two each time

$$L = S\,2^{-N}, \tag{3.153}$$

where S is the size of the integer in bits and N is the number of steps. $L = 1$ is therefore reached at

$$N = \log_2 S. \tag{3.154}$$

For a 64 Bit integer (`unsigned long int` on Linux and other systems implementing the LP64 data model) we would have the following algorithm:

```
function bits_set(n) {

n = (n & 0x5555555555555555) + ((n & 0xAAAAAAAAAAAAAAAA) >> 1)
n = (n & 0x3333333333333333) + ((n & 0xCCCCCCCCCCCCCCCC) >> 2)
n = (n & 0x0F0F0F0F0F0F0F0F) + ((n & 0xF0F0F0F0F0F0F0F0) >> 4)
n = (n & 0x00FF00FF00FF00FF) + ((n & 0xFF00FF00FF00FF00) >> 8)
n = (n & 0x0000FFFF0000FFFF) + ((n & 0xFFFF0000FFFF0000) >> 16)
n = (n & 0x00000000FFFFFFFF) + ((n & 0xFFFFFFFF00000000) >> 32)

return n
}
```

Here, we use a hexadecimal representation of the integers to cut the number of digits from 64 down to 16. Since the hexadecimals are converted as `0x5=0b0101` and `0xA=0b1010`, respectively, we clearly recognize the generalization of the previous discussion in the first line. This algorithm requires 18 operations and is therefore fairly efficient. Note that some compilers provide builtin implementations, e.g., `__builtin_popcount` in GCC [92] or `_popcnt` in ICC [93], that perform hardware instructions if available and are therefore more efficient on systems supporting the SSE4 instruction set.

Computing the occupation number The computation of occupation numbers is now merely an application of the previous algorithm. The total occupation is simply given by the total number of set bits. Separate occupations for up- and down-spins can be computed by applying certain bit masks beforehand:

```
up_mask = 0xFFFFFFFF00000000
dn_mask = 0x00000000FFFFFFFF
```

The occupation numbers are then

```
n_up = bits_set(n & up_mask)
n_dn = bits_set(n & dn_mask)
```

In a completely analogous way we define masks

```
1_mask = 1
2_mask = 2
3_mask = 4
.
.
.
32_mask = 1 << 32
```

For each spin-orbital. The local occupation for spin-orbital m is then easily obtained by

```
(n & m_mask) || 0
```

which returns `true` iff the m-th bit is set in n.

3.3.3 The QR Algorithm

Often the QR algorithm [94–96] is used to perform the diagonalization when calling the LAPACK [97] routines `*geev` with `*` either `s,d,c` or `z` for real single or double precision, and complex single and double precision, respectively. Starting with a matrix H, this algorithm performs a QR decomposition defined through

$$H = QR, \tag{3.155}$$

where Q is a unitary matrix and R an upper triangular matrix. Then, by computing

$$H_1 = RQ = Q^{-1}QRQ = Q^{-1}HQ, \tag{3.156}$$

a matrix is obtained that has the same eigenvalues as H. Performing this step iteratively, the sequence of matrices H_i converges to an upper triangular matrix, for which the eigenvalues are given by the diagonal matrix elements.

Unfortunately, the matrix products that have to be computed in every step of the iteration are rather expensive and therefore linear algebra packages like LAPACK often use the Schur decomposition of the initial matrix to construct a triangular matrix. One such method typically used is the the Householder transformation [98], which aims at bringing the matrix H into tridiagonal shape. This reduces the cost of both the QR decomposition and the matrix product to $\mathcal{O}(n)$ [99].

Since the full many-body Hamiltonian is rather sparse—as shown earlier—special algorithms that make use of this property can be used. In particular, sparse matrices can be represented in memory, e.g., as a tuple of three arrays

$$A = ([r_1, r_2, r_3, \ldots], [c_1, c_2, c_3, \ldots], [a_1, a_2, a_3, \ldots]), \tag{3.157}$$

where $r_i, c_i \in \mathbb{N}$ are, respectively, the row and column indices of the matrix elements a_i. Instead of a total size of $N^2 \times 16\text{B}$ (16Bytes double precision complex numbers) this representation requires only $n \times 16\text{B} + 2n \times 8\text{B}$, where n is much smaller than N^2. An efficient sparse memory representation facilitates the treatment of larger systems that would otherwise easily reach the memory limitations of the hardware. In our calculations we use the `scipy.sparse.linalg.eigsh` routine from the SciPy [100] package that is essentially a wrapper for the Arpack library. The algorithm implemented under the hood is the "implicitly restarted Lanczos method" [101].

3.3.4 Iterative Approximate Methods

Instead of the numerically exact diagonalization of the Hamiltonian matrix (or blocks thereof), approximate iterative schemes exist that provide results with good accuracy at much smaller

computational cost. The most commonly applied method is the so-called Lanczos method [83, 102, 103].

The main idea is the representation of the Hamiltonian in a smaller Hilbert space—the so-called Krylov subspace, in which the matrix assumes a tridiagonal shape. The Krylov subspace if formally defined as [83]

$$\mathcal{K}(A, q_1, k) = \mathrm{Span}\{q_1, Aq_1, A^2q_1, \dots A^{k-1}q_1\}, \tag{3.158}$$

$A \in \mathbb{C}^{n\times n}$, $q_1 \in \mathbb{C}^n$ is a random initial vector and $k < n$ is the dimension of the Krylov subspace. The vectors $v_i = A^i q_1$ are then computed in an iterative scheme, during which one applies the Gram-Schmidt algorithm for orthogonalization. Intermediate results can be used to assemble the representation of A in $\mathcal{K}$ without requiring any additional calculation. The resulting decomposition has the form $T = Q^\dagger AQ$, with

$$T = \begin{pmatrix} \alpha_1 & \beta_1 & 0 & \dots & 0 \\ \beta_1 & \alpha_2 & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & \beta_{n-1} \\ 0 & \dots & & \beta_{n-1} & \alpha_n \end{pmatrix}. \tag{3.159}$$

The better scaling in terms of faster convergence of dense matrix algorithms such as the QR algorithm when applied to tridiagonal shape is then exploited to obtain a number k of eigenvalues and eigenvectors of the original problem.

In addition to zero-temperature calculations, the computation of finite temperature observables is possible with the finite-temperature Lanczos method [104, 105].

3.4 Cluster Perturbation Theory

Neglecting numerical instabilities, the exact diagonalization scheme suffers mainly from finite size effects, meaning that while the result obtained accurately reflects the physics of the finite size system, not all quantities scale well to the original large system that one is interested in. Consequently, it is possible that the phyisics observed are merely artifacts of the finite size and therefore a scaling analysis is always required in order to build trust in the result.

While ED can essentially be improved only by facilitating the treatment of larger systems, there are other approaches that try to alleviate the shortcomings of the bare method. One of these is the so-called Cluster Perturbation Theory (CPT) [106–109], with which one aims at minimizing finite size effects by building a lattice out of the initial finite cluster by virtue of perturbation theory in the couplings between adjacent clusters. CPT is part of a whole family of quantum cluster approaches [110–112] that cover various degrees of sophistication. Among these, CPT can be regarded as the simplest as it provides a "single shot" result and does not require a self-consistent solution. Recent applications include also topological systems [113–115].

3.4.1 Lattice Definition

Usually, we describe a lattice in terms of lattice vectors $\mathbf{a}_1, \mathbf{a}_2$, through which every point on the lattice can be expressed as a set of integer indices n_1, n_2 as

$$\mathbf{r} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2. \tag{3.160}$$

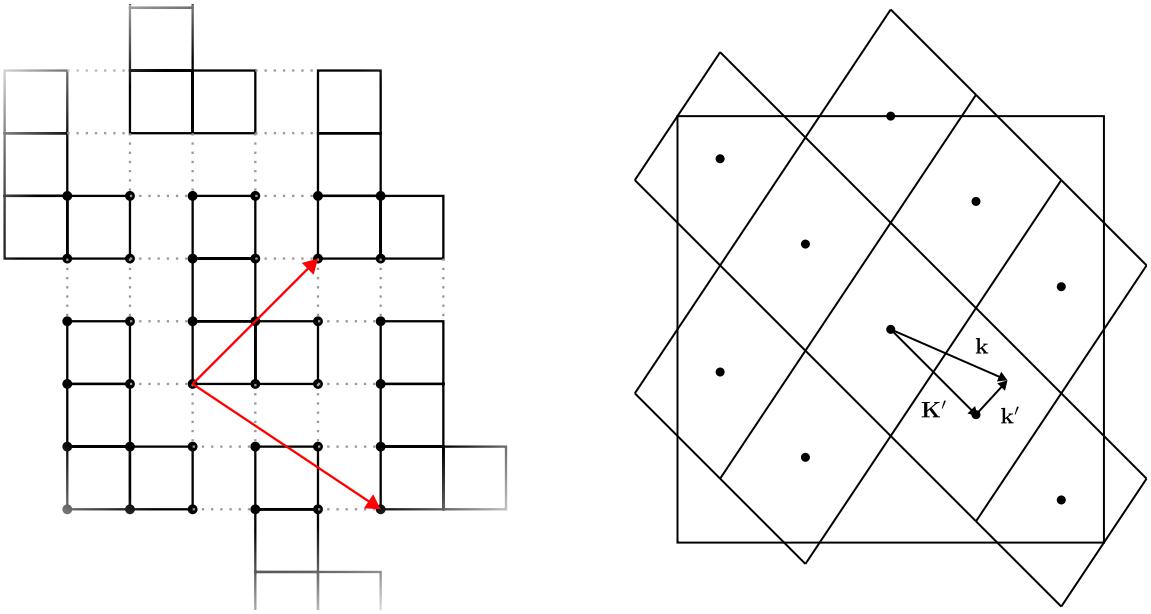

Figure 3.6: Superlattice mapping for an arbitrary cluster choice on the square lattice. Left: The original square lattice is cut up into L-shaped clusters. Solid bonds indicate hoppings within clusters, while dashed bonds correspond to the couplings between clusters. A choice for the two lattice vectors on the superlattice is drawn in red. Right: First Brillouin zone of the original lattice (large square) and overlap with the shrunken Brillouin zone of the superlattice. Any point in the original Brillouin zone has a unique representation in terms of $\mathbf{K}'$ and $\mathbf{k}'$, given by Eq. 3.165.

By first combining a number of adjacent sites into a cluster and treating this cluster as an enlarged unit cell we can define new lattice vectors $\mathbf{s}_1, \mathbf{s}_2$ that span the *superlattice*. A lattice point is then expressed as

$$\mathbf{r} = n_1'\mathbf{s}_1 + n_2'\mathbf{s}_2 + \delta\mathbf{r}, \tag{3.161}$$

i.e., an integer linear combination of the superlattice vectors plus a residual intra cluster displacement $\delta\mathbf{r}$ that connects the point with the origin of the cluster. We write in short

$$\mathbf{r} = \mathbf{r}' + \delta\mathbf{r}, \tag{3.162}$$

where $\mathbf{r}'$ is the coordinate of the host cluster in the superlattice. In Fourier space we have

$$\mathbf{K}_i \cdot \mathbf{R}_j = 2\pi\delta_{ij} \tag{3.163}$$

for reciprocal lattice vectors $\mathbf{K}_i$ in the fully expanded lattice and define

$$\mathbf{K}_i' \cdot \mathbf{R}_j' = 2\pi\delta_{ij} \tag{3.164}$$

for the superlattice. In this notation primed quantities refer to the superlattice. We can express every momentum vector in the full Brillouin zone as

$$\mathbf{k} = \mathbf{K}_i' + \mathbf{k}', \tag{3.165}$$

where $\mathbf{k}'$ is a vector in the superlattice Brillouin zone. This construction is illustrated in Fig. 3.6.

A site can be labeled on the full lattice with an index i that determines its position $\mathbf{r}_i$. We have seen already that we can split up the position vector into the cluster coordinates and the

residual displacement within the cluster $\mathbf{r}_i = \mathbf{r}'_i + \delta\mathbf{r}$. We now define indices for cluster and atom via

$$\mathbf{r}_i = \mathbf{r}_m + \mathbf{r}_a, \tag{3.166}$$

where m labels the cluster and a the atom within the cluster. Within this notation the Fourier transform of the fermionic operators is defined as

$$c_{\mathbf{k}} = \frac{1}{\sqrt{L}} \sum_i e^{-i\mathbf{k}\cdot\mathbf{r}_i} c_i \tag{3.167}$$

$$= \frac{1}{\sqrt{L}} \sum_{m,a} e^{-i\mathbf{k}\cdot(\mathbf{r}_m+\mathbf{r}_a)} c_{m,a}, \tag{3.168}$$

with the total number of sites L. In addition we can have the mixed representation of the Fourier transform to the superlattice

$$c_{\mathbf{k}',a} = \frac{1}{\sqrt{L_\mathrm{C}}} \sum_m e^{-i\mathbf{k}'\cdot\mathbf{r}_m} c_{m,a}, \tag{3.169}$$

where L_C is the number of clusters, and

$$c_{\mathbf{K}',\mathbf{k}'} = \frac{1}{\sqrt{L}} \sum_{m,a} e^{-i(\mathbf{k}'\cdot\mathbf{r}_m+\mathbf{K}'\cdot\mathbf{r}_a)} c_{m,a}. \tag{3.170}$$

Apparently, Eq. 3.168 and Eq. 3.170 are not the same, since

$$\mathbf{k}\cdot(\mathbf{r}_m+\mathbf{r}_a) = (\mathbf{K}'+\mathbf{k}')\cdot\mathbf{r}_m + (\mathbf{K}'+\mathbf{k}')\cdot\mathbf{r}_a \neq \mathbf{k}'\cdot\mathbf{r}_m + \mathbf{K}'\cdot\mathbf{r}_a. \tag{3.171}$$

With $\mathbf{K}'\cdot\mathbf{r}_m = 2\pi n$ the phase difference is given by $\mathbf{k}'\cdot\mathbf{r}_a$. Due to

$$\frac{L_\mathrm{C}}{L} \sum_{\mathbf{Q}'} \sum_a e^{-i(\mathbf{k}'+\mathbf{K}'-\mathbf{Q}')\cdot\mathbf{r}_a} c_{\mathbf{Q}',\mathbf{k}'} = \frac{L_\mathrm{C}}{L} \sum_{\mathbf{Q}'} \sum_a e^{-i(\mathbf{k}'+\mathbf{K}'-\mathbf{Q}')\cdot\mathbf{r}_a} \frac{1}{\sqrt{L}} \sum_{m,b} e^{-i(\mathbf{k}'\cdot\mathbf{r}_m+\mathbf{Q}'\cdot\mathbf{r}_b)} c_{m,b} \tag{3.172}$$

$$= \frac{1}{\sqrt{L}} \sum_{m,a,b} e^{-i(\mathbf{k}'+\mathbf{K}')\cdot\mathbf{r}_a} e^{-i\mathbf{k}'\cdot\mathbf{r}_m} c_{m,b} \frac{L_\mathrm{C}}{L} \sum_{\mathbf{Q}'} e^{i\mathbf{Q}'\cdot(\mathbf{r}_a-\mathbf{r}_b)} \tag{3.173}$$

$$= \frac{1}{\sqrt{L}} \sum_{m,a} e^{-i(\mathbf{k}'+\mathbf{K}')\cdot\mathbf{r}_a} e^{-i\mathbf{k}'\cdot\mathbf{r}_m} c_{m,a} \tag{3.174}$$

$$= \frac{1}{\sqrt{L}} \sum_{m,a} e^{-i\mathbf{k}\cdot(\mathbf{r}_a+\mathbf{r}_m)} c_{m,a} \tag{3.175}$$

$$= c_{\mathbf{k}} = c_{\mathbf{K}'+\mathbf{k}'}, \tag{3.176}$$

this phase factor is a consequence of a unitary transformation

$$U_{\mathbf{K}'\mathbf{Q}'}(\mathbf{k}') = \frac{L_\mathrm{C}}{L} \sum_a e^{-i(\mathbf{k}'+\mathbf{K}'-\mathbf{Q}')\cdot\mathbf{r}_a}, \tag{3.177}$$

which satisfies

$$[U^\dagger U]_{\mathbf{P}'\mathbf{Q}'} = \sum_{\mathbf{K}'} U^*_{\mathbf{P}'\mathbf{K}'}(\mathbf{k}')U_{\mathbf{K}'\mathbf{Q}'}(\mathbf{k}') \tag{3.178}$$

$$= \frac{L_\mathrm{C}^2}{L^2}\sum_{a,b}\sum_{\mathbf{K}'} e^{i(\mathbf{k}'+\mathbf{P}'-\mathbf{K}')\cdot\mathbf{r}_b}e^{-i(\mathbf{k}'+\mathbf{K}'-\mathbf{Q}')\cdot\mathbf{r}_a} \tag{3.179}$$

$$= \frac{L_\mathrm{C}}{L^2}\sum_{a,b} e^{i\mathbf{k}'\cdot(\mathbf{r}_b-\mathbf{r}_a)}e^{i(\mathbf{P}'\cdot\mathbf{r}_b+\mathbf{Q}'\cdot\mathbf{r}_a)}\sum_{\mathbf{K}'} e^{i\mathbf{K}'\cdot(\mathbf{r}_b-\mathbf{r}_a)} \tag{3.180}$$

$$= \frac{L_\mathrm{C}}{L}\sum_{a} e^{i(\mathbf{P}'+\mathbf{Q}')\cdot\mathbf{r}_a} \tag{3.181}$$

$$= \delta_{\mathbf{P}'\mathbf{Q}'}. \tag{3.182}$$

For the hopping matrix we then obtain the representation

$$\sum_{ij} t_{ij}c_i^\dagger c_j = \sum_{\mathbf{k}',\mathbf{q}'}\sum_{m,n} t_{m,a;n,b}c^\dagger_{m,a}c_{n,b} \tag{3.183}$$

$$= \frac{1}{L_\mathrm{C}}\sum_{\mathbf{k}',\mathbf{q}'}\sum_{m,n} t_{m,a;n,b}e^{-i\mathbf{k}'\cdot\mathbf{r}_m}e^{i\mathbf{q}'\cdot\mathbf{r}_n}c^\dagger_{\mathbf{k}',a}c_{\mathbf{q}',b}, \tag{3.184}$$

and with $t_{m,a;n,b} = t_{m-n,a;0,b}$ due to periodicity

$$\sum_{ij} t_{ij}c_i^\dagger c_j = \frac{1}{L_\mathrm{C}}\sum_{\mathbf{k}',\mathbf{q}'}\sum_{m,n} t_{m-n,a;0,b}e^{-i\mathbf{k}'\cdot(\mathbf{r}_m-\mathbf{r}_n)}e^{i(\mathbf{q}'-\mathbf{k}')\cdot\mathbf{r}_n}c^\dagger_{\mathbf{k}',a}c_{\mathbf{q}',b} \tag{3.185}$$

$$= \sum_{\mathbf{k}',\mathbf{q}'}\sum_{m} t_{m,a;0,b}e^{-i\mathbf{k}'\cdot\mathbf{r}_m}c^\dagger_{\mathbf{k}',a}c_{\mathbf{q}',b}\frac{1}{L_\mathrm{C}}\sum_{n} e^{i(\mathbf{q}'-\mathbf{k}')\cdot\mathbf{r}_n} \tag{3.186}$$

$$= \sum_{\mathbf{k}'}\sum_{m} t_{m,a;0,b}e^{-i\mathbf{k}'\cdot\mathbf{r}_m}c^\dagger_{\mathbf{k}',a}c_{\mathbf{k}',b}. \tag{3.187}$$

The hopping matrix in terms of the superlattice momentum $\mathbf{k}'$ is therefore given by

$$t_{a,b}(\mathbf{k}') = \sum_{m} t_{m,a;0,b}e^{-i\mathbf{k}'\cdot\mathbf{r}_m}. \tag{3.188}$$

3.4.2 Cluster Solution

The full many-body Hamiltonian is solved with exact diagonalization on a cluster with open boundary conditions. This means that the links that couple different clusters are left out and the respective hoppings are set to zero. In principle, the Green's function can be computed from Eq. 3.128 knowing the ground state $|0\rangle$ of H, however, the chemical potential is not known. Therefore, one derives an analogous Lehmann representation with the time-evolution given by e^{-iHt} instead, that is given by

$$G^R_{ab}(\omega) = \sum_n \frac{\langle 0|c_a|n\rangle\langle n|c_b^\dagger|0\rangle}{\omega + i\eta - E_n + E_0} + \sum_n \frac{\langle 0|c_b^\dagger|n\rangle\langle n|c_a|0\rangle}{\omega + i\eta + E_n - E_0}. \tag{3.189}$$

A detailed description of the exact procedure is given in [109] and shall not be repeated at this point.

3.4.3 Restoring the Lattice

In order to compute the cluster solution we had to cut certain links between adjacent clusters. The next step involves adding these terms back in, which is of course easier said than done. Formally, the Hamiltonian has been decomposed into two parts

$$H = \underbrace{\sum_{a,b} t'_{a,b} c_a^\dagger c_b + H_U}_{H_\mathrm{C}} + \underbrace{\sum_{a,b} V_{a,b} c_a^\dagger c_b}_{H_V}, \tag{3.190}$$

where H_C is the cluster Hamiltonian and therefore $t'_{a,b}$ is missing links between clusters that instead comprise H_V. One then treats the term H_V as a perturbation to H_C and obtains for the Green's function to first order in H_V [107, 108, 116, 117]

$$G^\mathrm{CPT}(\omega, \mathbf{k}') = [(G^\mathrm{C})^{-1}(\omega) - H_V(\mathbf{k}')]^{-1}, \tag{3.191}$$

where G^C is the cluster Green's function. This corresponds exactly to the following approximation [106]:

$$\Sigma^\mathrm{CPT}(\omega) = \Sigma^\mathrm{C}(\omega), \tag{3.192}$$

i.e., the lattice self-energy is approximated through the cluster self-energy.

3.4.4 Periodization

The CPT-Green's function G^CPT is a matrix-valued function of $\mathbf{k}'$, i.e., in superlattice coordinates, in the space of sites a, b in the clusters. In order to compare with other methods and momentum-resolved experiments we need to express this Green's function as a function of $\mathbf{k}$, which corresponds to mapping the object G^CPT to the fully periodic lattice through a process called *periodization*. The name becomes clear if we remember that the solution to the cluster breaks translational symmetry in a specific way, since we cut certain links on the fully periodic lattice. Perturbation theory brought these links back, however, since the lattice self-energy still breaks the translational symmetry so does the CPT Green's function. Unfortunately, there are only approximate methods to achieve a periodic result, the most commonly used is described below.

Since translational symmetry was only present on the superlattice, the Fourier transform has been carried out only with respect to superlattice coordinates $\mathbf{R}'$. We complete the transform by applying

$$G_{\mathbf{K}'\mathbf{Q}'}(\omega, \mathbf{k}') = \frac{L_\mathrm{C}}{L} \sum_{a,b} e^{-i(\mathbf{K}'\cdot\mathbf{r}_a - \mathbf{Q}'\cdot\mathbf{r}_b)} G^\mathrm{CPT}_{ab}(\omega, \mathbf{k}'). \tag{3.193}$$

Now, since G^CPT was computed for a system with broken translational symmetry it is also likely to be unsymmetric and hence, it is not diagonal in $\mathbf{K}, \mathbf{K}'$. As an approximation we can choose to neglect the off-diagonal terms and define

$$G_{\mathbf{K}'}(\omega, \mathbf{k}') = \frac{L_\mathrm{C}}{L} \sum_{a,b \in \mathrm{cluster}} e^{-i\mathbf{K}'\cdot(\mathbf{r}_a - \mathbf{r}_b)} G^\mathrm{CPT}_{a,b}(\omega, \mathbf{k}'). \tag{3.194}$$

With Eq. 3.176 this can also be written in terms of $\mathbf{k}$ as

$$G^\mathrm{periodic}(\omega, \mathbf{k}) = \frac{L_\mathrm{C}}{L} \sum_{a,b \in \mathrm{cluster}} e^{-i\mathbf{k}\cdot(\mathbf{r}_a - \mathbf{r}_b)} G^\mathrm{CPT}_{a,b}(\omega, \mathbf{k}), \tag{3.195}$$

where we used that $t_{a,b}(\mathbf{k}') = t_{a,b}(\mathbf{k})$.

We note that the approximation above is exact if the cluster contains only one site, i.e., translational symmetry was not broken and there are no off-diagonal terms, or when the self-energy vanishes, i.e., in the non-interacting limit. Additionally, if the perturbation H_V is very small, i.e., for a strongly interacting system with $U \gg t$, the error becomes very small and vanishes in the atomic limit. Of course, the limit of infinite cluster size also produces the exact result, however, this is irrelevant in practice due to the strong limitations to small clusters that result from the exponential scaling of the problem with the system size.

Secondly, since the spectral function is defined as the trace over the Green's function, only the $G_{\mathbf{KK}}$ diagonal terms contribute in any case. The periodization approximation thus has no negative influence on the computation of the spectral function. For other observables that are not translationally invariant this is not the case, however, and it is therefore recommended to use the periodization cautiously. In case the full Green's function or self-energy is desired, e.g., for comparison with other methods, the approximation should be treated with suspicion.

Since the functional dependence of the self-energy is of great interest to us in the discussion presented in Chapters 5 and 6, the periodization error is unavoidable and therefore CPT did not find much application throughout the rest of this work. We will nonetheless spend a little more time on the discussion of the error.

Let us investigate the important case $U = 0$. Since an exact diagonalization solver is used we have obviously no systematic error in

$$G^{\text{cluster}}(\omega) = [\omega - H_{\text{C}}]^{-1} . \tag{3.196}$$

The CPT equation then gives

$$G^{\text{CPT}}(\omega, \mathbf{k}') = \left[\omega - H_{\text{C}} - H_V(\mathbf{k}')\right]^{-1} , \tag{3.197}$$

where we note that $H_{\text{C}} + H_V(\mathbf{k}')$ is simply $H(\mathbf{k}')$, i.e., the Hamiltonian in the mixed representation with superlattice momentum $\mathbf{k}'$. This corresponds to a Green's function definition in terms of $c_{\mathbf{k}',a}$ as in Eq. 3.169. The additional Fourier transform w.r.t. the site index yields the Green's function in the representation of Eq. 3.170, however, we have already seen that these fermionic operators are related to the usual operators on the periodic lattice of Eq. 3.168 through a unitary transformation that is given in Eq. 3.177. For the Green's function one obtains [109]

$$G(\omega, \mathbf{K}' + \mathbf{k}', \mathbf{Q}' + \mathbf{k}') = \frac{L_{\text{C}}}{L} \sum_{a,b} e^{-i(\mathbf{K}'+\mathbf{k}')\cdot\mathbf{r}_a} e^{i(\mathbf{Q}'+\mathbf{k}')\cdot\mathbf{r}_b} G^{\text{CPT}}_{ab}(\omega, \mathbf{k}'). \tag{3.198}$$

The off-diagonal terms for $\mathbf{K}' \neq \mathbf{Q}'$ generally do not vanish, however, since $G^{\text{CPT}}_{ab}(\omega, \mathbf{k}')$ is exact at $U = 0$ we know that setting $\mathbf{K}' = \mathbf{Q}'$ yields the correct result. This is not true, however, if $G^{\text{CPT}}_{ab}(\omega, \mathbf{k}')$ breaks the translational symmetry, which happens precisely if $\Sigma \neq 0$.

We now discuss these issues at the example of results for the 1D single band Hubbard model at half filling. In Fig. 3.7 we show the spectral function obtained with CPT at $U/t = 4$ for various cluster sizes. In all cases we obtain the same qualitative picture of an insulator with a finite spectral gap at the Fermi level. However, the smallest cluster produces rather sharp bands indicating long quasiparticle lifetimes. This is changed once the cluster size is increased and we observe a much broader distribution of spectral weight indicating non-Fermiliquid behavior. The incremental changes in the spectral function become rather small for cluster sizes > 10, indicating that we have already reached the level of convergence that is easily accessible numerically.

In order to investigate the validity of the periodization we analyze the periodicity of the cluster self-energy. The perfectly periodic system would satisfy $\Sigma_{ab} = \Sigma_{a-b,0}\ \forall a, b$, which implies equal values along diagonals of the self-energy matrix. In fact, we expect the self-energy matrix

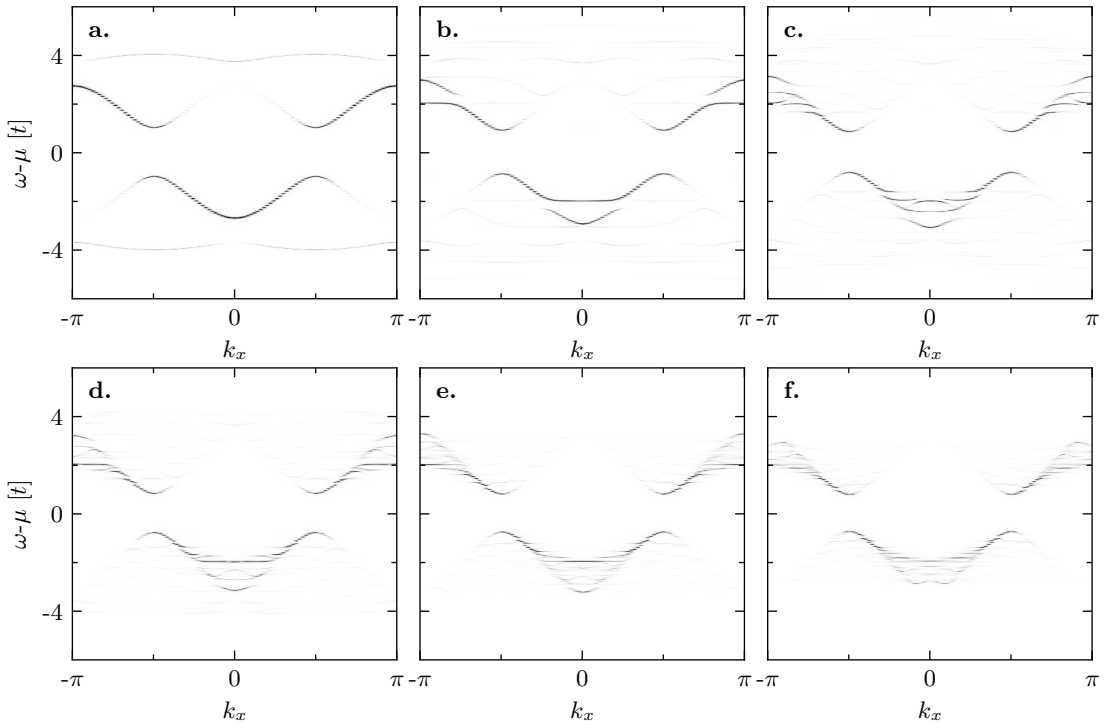

Figure 3.7: Spectral function $A(\omega, k)$ for the half filled one band Hubbard model in 1D for $U/t = 4$. Cluster sizes are **a.** 2, **b.** 4, **c.** 6, **d.** 8, **e.** 10, **f.** 12. Already the smallest non-trivial cluster with two sites shows the opening of the spectral gap at large U, however, the individual bands remain rather sharp. Increasing the cluster size does not change the location of the bands, only the width, reflecting the finite lifetime of quasiparticles.

to have the following structure:

$$\Sigma(\omega) = \begin{pmatrix} A & B & C & D & \dots \\ B' & A & B & C & \dots \\ C' & \ddots & \ddots & \ddots & \\ & & & & \\ \vdots & & & & A \end{pmatrix}. \tag{3.199}$$

Breaking translational symmetry will result in a deviation from this general form meaning that different values are encountered along individual diagonals. This can be measured in terms of the variance

$$\epsilon^2(\omega) = \max_l \Big\{ \mathrm{Var}\big[\, \{\Sigma_{a,a+l}(\omega)\ \forall a\} \,\big] \ \Big|\ l = 0, 1, 2 \dots \Big\}. \tag{3.200}$$

where l indicates the l-th diagonal. We note that for cluster sizes of $L/L_\mathrm{C} = 1, 2$, translational symmetry is not broken in this way since

$$\Sigma^1_{ab} = \begin{pmatrix} A \end{pmatrix}, \qquad \Sigma^2_{ab} = \begin{pmatrix} A & B \\ B' & A \end{pmatrix}, \tag{3.201}$$

i.e., there cannot be a finite variance since either there is only one value or the two values are necessarily identical as they both refer to boundary sites. Translational symmetry is broken explicitly in the self-energy for larger clusters where inner sites are distinct from boundary sites.

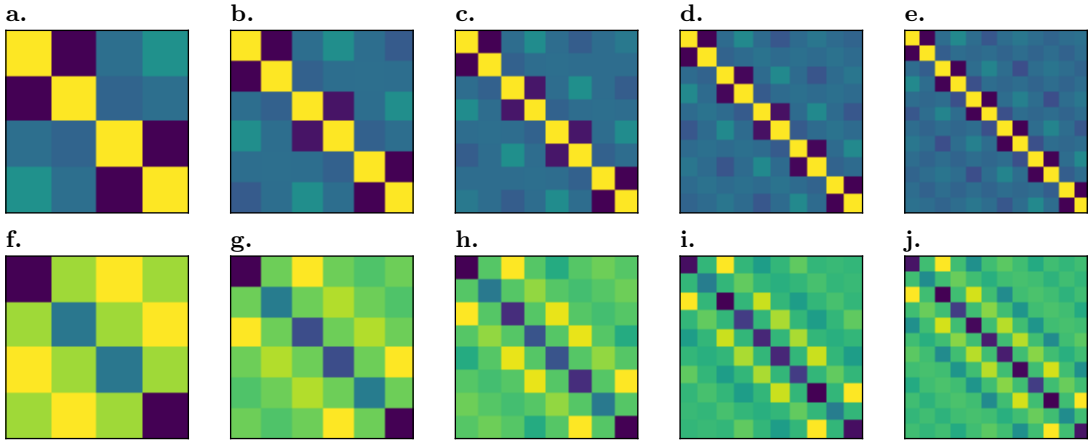

Figure 3.8: Matrix elements of the CPT self-energy $\Sigma(\omega = \mu)$ for the half-filled one band Hubbard model in 1D at $U/t = 4$ for different cluster sizes: 4 (**a.**,**f.**), 6 (**b.**,**g.**), 8 (**c.**,**h.**), 10 (**d.**,**i.**), 12 (**e.**,**j.**). The real part of $\Sigma_{ab}(\omega = \mu)$ is shown in the top row, the imaginary part in the bottom row, both in arbitrary units. The two-site cluster trivially has translational symmetry due to a lack of inner sites and is not shown. All other cases clearly break translational symmetry and we find that for large clusters the self-energy becomes more and more periodic across the inner sites, while the boundary sites break the symmetry. The periodization approximation should therefore become better the less significant boundary sites are compared to the whole.

We show the real and imaginary part of the cluster self-energy at $\omega = \mu$ for different cluster sizes in Fig. 3.8. All clusters with sizes larger than 2 clearly break the translational symmetry, since values along the diagonals differ significantly. It is expected that the translational symmetry will be restored along the inner sites for large clusters, where the boundary sites are negligible compared to the whole cluster. An onset of this can already be observed in the 10-site cluster, where the imaginary part of the self-energy is more or less periodic across the inner-most four sites, which is not observed at $L/L_\mathrm{C} = 8$ (cf. Fig. 3.8**i**,**j**).

The deviation of the self-energy from a periodic solution, described by $\epsilon(\omega)$ of Eq. 3.200, shows a strong frequency dependence due to the strong peaks usually encountered in finite-size calculations at real frequencies. Here, we chose $\eta = 10^{-2}$ for the regularization of the Green's function that adds minimal broadening to the delta-peaks. In order to remove the frequency dependence from the error we define instead

$$\epsilon_l^2 = \int_{-\infty}^{\infty} \mathrm{d}\omega \ \mathrm{Var}\big[\,\{\Sigma_{a,a+l}(\omega)\ \forall a\}\,\big]. \tag{3.202}$$

The result for the self-energies that we obtained before are shown in Fig. 3.9. As shown before, the self-energy of the two-site cluster is periodic. We can apply the same argument to the diagonal of order $l = L/L_\mathrm{C} - 1$ and $l = L/L_\mathrm{C} - 2$, i.e., the highest two values of l in all other cases, since there we compute either the variance of only one value, which vanishes necessarily, or the variance of two equal values. Concentrating on the remainder of the values we find that ϵ_l is generally largest for smaller l—a consequence of larger values of Σ_{ll} compared to off-diagonal matrix elements. Furthermore, we find that the the deviation from a periodic solution decreases as the cluster size increases, as expected.

We note that the results shown here indicate that clusters of size 12 are still aperiodic enough to cast doubts on the applicability of the periodization scheme to compute a lattice self-energy. Moreover, the slowdown in the convergence rate with increasing cluster size makes reaching an

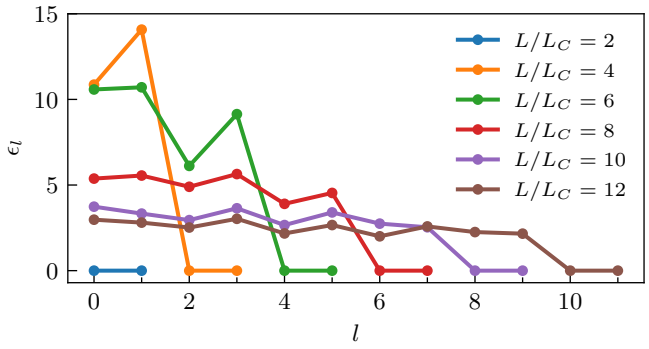

Figure 3.9: Periodicity error [Eq. 3.202] of the finite size self-energy of the half-filled one band Hubbard model in 1D at $U/t = 4$ for different cluster sizes L/L_C. ϵ_l is a function of the order l of the diagonal, i.e. $\Sigma_{i,i+l}$. The two-site cluster is by definition of our measure periodic and therefore $\epsilon_l = 0$. All other clusters show a finite value, while the values for the largest two l necessarily vanish in all cases for the same reason as for $L/L_C = 2$. For all other values we find that ϵ_l generally decreases as a function of the cluster size L/L_C, i.e. the solution becomes more periodic. The speed of the convergence towards zero decreases with increasing cluster size.

approximately periodic solution impossible in practice. This becomes even worse for higher-dimensional systems where generally a much larger number of sites per cluster is required to achieve the same results. In 2D, for instance, we would expect a similar periodicity for a 12×12 cluster, which is numerically infeasible. While CPT remains relevant for the computation of the spectral function and other physical observables including topological invariants, we cannot compute the lattice self-energy $\Sigma(\mathbf{k})$ for comparison with other methods that operate on the fully periodic lattice. For this reason we will not make much use of CPT during the discussions in the later chapters.

3.5 Dynamical Mean Field Theory (DMFT)

For a period of roughly 30 years since its invention in the late 1980s and early 1990s [118–121] DMFT has been the main workhorse for studying strongly correlated systems in condensed matter physics. This huge success is mainly due to the capability of describing the Mott-Hubbard transition.

The basic idea underlying the mean field approach is illustrated in Fig. 3.10. One chooses a single site as a starting point and removes it from the lattice. In the following the remainder of the lattice is described as a non-interacting bath that couples to the removed site (where the Hubbard interaction is still present) through amplitudes $V_{k\sigma}$. While static mean field takes into account only interactions in the limit of infinite time, where the system is equilibrated, DMFT treats the dynamical processes of electron exchanges with the bath exactly, effectively leading to a more accurate description of the local interaction, since higher-order processes like the one shown in the figure, where the interaction affects the initially empty site after several electron exchanges with the bath. In terms of equations one starts from the Hubbard model [87]

$$H_{\text{Hubbard}} = \sum_{\substack{i,j \\ \sigma}} t_{ij} c^\dagger_{i\sigma} c_{j\sigma} + U \sum_i n_{i\uparrow} n_{i\downarrow}, \tag{3.203}$$

which is mapped to

$$H_{\text{AIM}} = \sum_{k,\sigma} \varepsilon_k c^\dagger_{k\sigma} c_{k\sigma} + \sum_\sigma \varepsilon_d d^\dagger_\sigma d_\sigma + U d^\dagger_\uparrow d_\uparrow d^\dagger_\downarrow d_\downarrow + \sum_{k,\sigma} V_{k\sigma} (d^\dagger_\sigma c_{k\sigma} + c^\dagger_{k\sigma} d_\sigma). \tag{3.204}$$

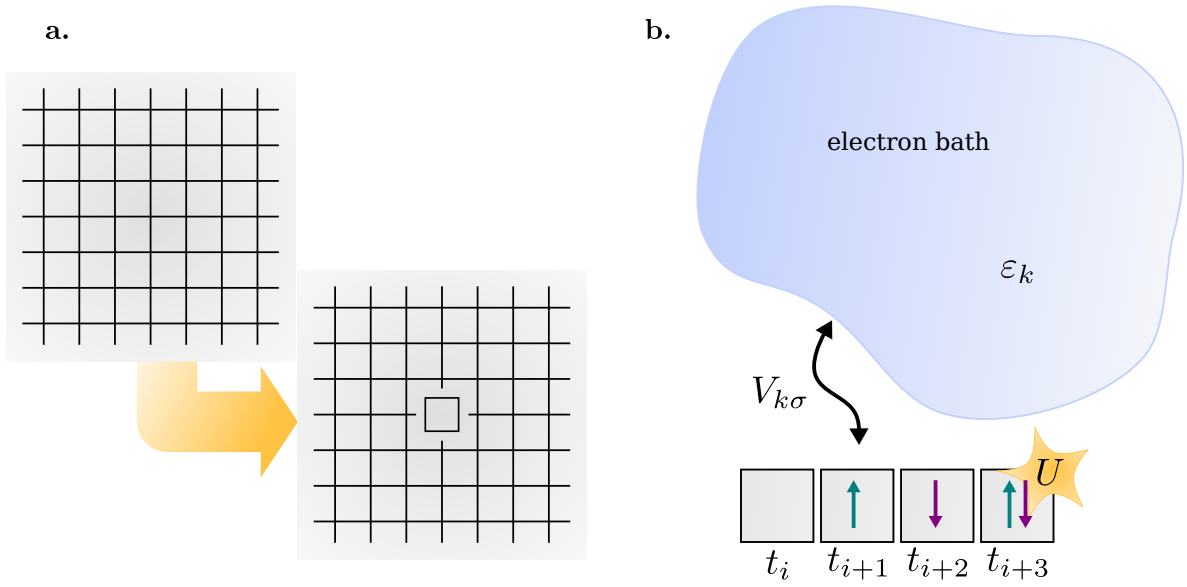

Figure 3.10: Illustration of DMFT via the cavity construction. **a.** Starting from the complete lattice a single site is picked out and removed. The remainder of the lattice is then considered as a non-interacting bath that couples to the cavity site and whose parameters are determined self-consistently. This interpretation is shown in **b.**, where the single site and the coupling $V_{k\sigma}$ to the bath are treated exactly, while the correlations on bath sites are neglected. In comparison to static mean field, DMFT takes into account the full dynamical information of the interaction.

Eq. 3.204 is the Anderson impurity model (AIM) [122] describing a single site denoted by $d, d^\dagger$ operators with a Hubbard interaction U and a non-interacting bath denoted by $c, c^\dagger$. The coupling between the two is described in terms of amplitudes V_k. Since in the original Hubbard model all sites were interacting sites one cannot easily perform this mapping. In fact, it turns out that there is no closed solution for V_k. Instead, the two models are related via self-consistent equations such that the parameters V_k need to be determined self-consistently. The latter requires an iterative solution of the Anderson impurity model. Fortunately, an exact solution to the AIM became attainable with the development of continuous-time quantum Monte Carlo methods [55, 123] that produce the correct self-energy in the limit of large sample sizes or long run times of the program. Alternatively, the AIM can be solved with exact diagonalization, however, this limits the number of bath sites and is therefore less accurate. Since in DMFT the self-energy of the lattice is approximated by the self-energy of the AIM we obtain the important equation

$$\Sigma_{\mathrm{DMFT}}(\omega, k) = \Sigma_{\mathrm{DMFT}}(\omega), \tag{3.205}$$

i.e., the reduction of complexity to a single site problem comes at the loss of the momentum-dependence of the self-energy. We will discus implications following this approximation on the topological classification in Chapters 5 and 6.

The self-consistent mapping underlying DMFT is approximate in nature and becomes exact in the important cases $U/t = 0$ and $t/U = 0$. In addition, it has been shown that DMFT becomes exact in the limit of infinite dimensions [118]. Since DMFT has already been covered in detail in an earlier thesis [4] we do not want to go into more details here and point instead at the large quantity of good literature on the subject [124–128].

Chapter 4

Statistics, Information Theory & Machine Learning

Since our statistical method presented in Chapter 6 and Chapter 7 requires basic knowledge of probability theory, we will give a brief introduction thereof in this chapter. This review is based on the axiomatic definition of Kolmogorov [129] and the book on statistics by Casella and Berger [130]. Furthermore, we will discuss the fundamentals of information theory (Sec. 4.2) and machine learning (Sec. 4.3) that are important for the statistical method presented in Chapter 7. We strongly recommend the books by Bishop and MacKay [131, 132] for more details on these topics.

4.1 Probability

While statistics deals with the interpretation of data, probability theory provides the means to make predictions based on this prior knowledge. This knowledge is encoded in the most important quantity in probability theory—the name-giving *probability*.

Definition 1 (Probability)**.** *Given a set E, a map $p : E \to [0, 1]$ is called probability distribution iff $\| p \| = 1$. A value $p(e)$ for $e \in E$ is called the probability of e.*

Here, the norm is defined differently for countable and uncountable sets E:

$$\| p \| = \begin{cases} \sum_{e \in E} p(e), & E \text{ countable}, \\ \| p \| = \int_E p(x)\, \mathrm{d}x, & E \text{ uncountable}, \end{cases} \tag{4.1}$$

where in the second case we demand that the function p is integrable over the domain E. From a physicist's perspective we can immediately identify that the units of p must differ between the two cases, since $\| p \| = 1$ implies that p is dimensionless in the case of countable E. For uncountable E, however, $[p] = [1/x]$ and therefore we call $p(x)$ the *probability density function*.

The elements of the set E are called *events*, each of which is assigned a probability to occur through p, such that an event $e_1 \in E$ with $p(e_1) = 0$ will never occur, while $e_2 \in E$ with $p(e_2) = 1$ will occur with such certainty that all other events can not take place due to the normalization of p.

With $p(E) := \sum_{e \in E} p(e) = 1$ it is guaranteed that any event will take place. Implicitly, this defines the sum rule for probabilities, i.e.,

$$p(e_1 \vee e_2) = p(e_1) + p(e_2). \tag{4.2}$$

In many situations the probability is not known. It is therefore necessary to determine $p(e)$ through some means. The easiest way is to set up an experiment, where an action with different possible outcomes $e \in E$ is repeated a number $N \in \mathbb{N}$ times and each outcome is recorded. The probability can then be approximated as the fraction of realizations of event e w.r.t. all recorded events

$$p(e) \approx n_e/N, \tag{4.3}$$

with n_e being the number of occurrences of event e. Due to $\sum_{e\in E} n_e = N$ and $0 \leq n_e \leq 1$ the definition above does indeed satisfy the requirements of a probability.

It is important to underline that probability is merely a theoretical concept that can only be applied in an approximate manner. In rigorous terms one calls the measured probability of Eq. 4.3 *frequency*. In this work, though, we use the word probability rather loosely to refer to either the exact concept or the measured value depending on the context.

In order to associate random events better with a particular experiment we define the so-called random variable X that can assume values in E. With our chance experiment in mind, X is the outcome of the next iteration, i.e., it does not have a fixed value since in each experiment all events $e \in E$ are allowed to occur. However, based on the probability $p(e)$ that is assigned to each possible outcome we can make a prediction for the next observable values of X. We then denote the probability for X to take on the value e with $p(X = e)$. This notation makes sense in the case of multiple random variables, where events may be overlapping. Let $X, Y \in E$ be two different random variables. The set of events when we observe both experiments is then $\tilde{E} = E \times E$ and we can analogously define the *joint probability* $p(X = e_1, Y = e_2)$, which defines the probability for both events e_1 and e_2 to occur at the same time. Again, we require $p \in [0, 1]$ and the normalization condition now means that the sum over all possible values of $X, Y \in \tilde{E}$ equals 1. The same notion can be generalized to an arbitrary number n of random variables X_i and the joint probability is conveniently described by Definition 1 with $E = E_1 \times \ldots \times E_n$ with $X = (X_1, \ldots, X_n) \in E$.

Using a set theoretic notation, where $p : \mathcal{P}(E) \to [0, 1]$ and $\mathcal{P}(E)$ denotes the power set of E, i.e., the set of all subsets, we can derive several important statements about probabilities. Using the sum rule of Eq. 4.2 we find for $A \subset E$, $p(A) = \sum_{e\in A} p(e)$. Let $E = \cup_i A_i$ with A_i mutually disjoint, i.e., A_i are non-overlapping subsets of E and therefore contain different events. With $A_i = E \setminus \cup_{j\neq i} A_j$ we find an important relation between the combined probability of the events in A_i and the probability of the complement $\bar{A}_i = \cup_j A_j$

$$p(A_i) = p(E) - p(\cup_j A_j) = 1 - p(\bar{A}_i). \tag{4.4}$$

Similarly, the sum rule immediately implies that

$$p(A_i \cup A_j) = p(A_i) + p(A_j), \tag{4.5}$$

and therefore

$$p(\cup_i A_i) = \sum_i p(A_i) = 1 \tag{4.6}$$

The situation is more complicated if the subsets A_i of E are not mutually disjoint, since then $p(E) = p(\cup_i A_i) < \sum_i p(A_i)$. In this case we perform another random experiment and use the resulting number of recorded events n_i for events from subsets A_i to define

$$p(A_i) \approx \frac{n_i}{n}, \quad p(A_j) \approx \frac{n_j}{n}, \tag{4.7}$$

where "$\approx \quad \to \quad =$" in the limit of an infinite number of repetitions. With $n_{ij} \leq n_i + n_j$ the number of samples that are contained in both subsets A_i, A_j, i.e., $e \in A_i \cap A_j$, we can define also

$$p(A_i \cap A_j) \approx \frac{n_{ij}}{n}. \tag{4.8}$$

This immediately leads to the definition of the probabilities

$$p(A_j|A_i) = \frac{n_{ij}}{n_i} \qquad \text{and} \qquad p(A_i|A_j) = \frac{n_{ij}}{n_j}, \tag{4.9}$$

which can be related to the joint probability through

$$p(A_j|A_i) = \frac{n_{ij}}{n}\frac{n}{n_i} = \frac{p(A_j \cap A_i)}{p(A_i)} \tag{4.10}$$

or equivalently

$$p(A_j \cap A_i) = p(A_j|A_i)p(A_i). \tag{4.11}$$

We call Eq. 4.9 the *conditional probability* of events in A_j w.r.t. A_i. Eq. 4.11 is called the multiplication law for probabilities, since it establishes that the joint probability of two events is given by the product of the probability of one event and the conditional probability that the second event takes place under the assumption that the first did. Note that in general $p(A_j|A_i) \neq P(A_j)$, i.e., $p(A_j \cap A_i) \neq p(A_j)p(A_i)$.

4.1.1 Bayes' Theorem

We now want to elaborate more on these conditional probabilities and derive an important relation between them. Assume that we have two random variables $X \in E_1, Y \in E_2$. We can define the so-called marginal probabilities $p(X), p(Y)$ that describe the outcome of one of X and Y irrespective of the other and the joint probability $p(X, Y)$ that describes the combination of both outcomes. The conditional probabilities on the other hand describe one random variable under the condition that the value of the other is already known. The notation $p(X = e|Y = f)$ means the probability of X assuming the value $e \in E_1$ *given that* $Y = f \in E_2$ is realized in the same random experiment. Since we need to retain normalization for $p(X)$ and $p(Y)$, we have $\sum_{e\in E_1} p(X = e) = 1$ and with $\sum_{e\in E_1, f\in E_2} p(X = e, Y = f) = 1$ this requires

$$p(X) = \sum_{e\in E_2} p(X, Y = e), \tag{4.12}$$

and

$$p(Y) = \sum_{e\in E_1} p(X = e, Y). \tag{4.13}$$

Due to Eq. 4.11 the conditional probability can be related to the joint probability via

$$p(X, Y) = p(X|Y)p(Y). \tag{4.14}$$

Of course, one can simply substitute X and Y and obtain

$$p(Y, X) = p(Y|X)p(X). \tag{4.15}$$

Due the symmetry of the $p(X, Y)$ we arrive at Bayes' theorem

$$p(X|Y)p(Y) = p(Y|X)p(X), \tag{4.16}$$

by simply equating the two expressions for the joint probability. Eq. 4.16 is also known as the definition of the inverse probability, since it relates $p(X|Y)$ to its inverse $p(Y|X)$.

4.1.2 Statistical (In-) Dependence

Suppose we have two random variables X, Y that correspond to two unrelated random experiments. Unrelated could here mean that you and your friend both performed a random experiment at home and you later met to compare results. We can assign marginal probabilities $p(X), p(Y)$ to both experiments separately as we usually would, since their outcomes are entirely unrelated to each another. Then, the joint probability is the probability that specific events in the two experiments *both* take place. When performing the random experiment each person records the counts of occurrences for each random variable X, Y and events $e_1 \in E_1, e_2 \in E_2$, respectively, to determine an estimate for the marginal probabilities

$$p(X = e_1) = n_{X=e_1}/n_X, \quad p(Y = e_2) = n_{Y=e_1}/n_Y. \tag{4.17}$$

For the joint probability $p(X = e_1, Y = e_2)$ it is intuitively clear that out of a total of n_X and n_Y separate events, respectively, we can have a total of $n_X n_Y$ combinations, while the samples restricted to our specific two events e_1, e_2 can be arranged in $n_{X=e_1} n_{Y=e_2}$ possible ways. Thus,

$$p(X = e_1, Y = e_2) = \frac{n_{X=e_1} n_{Y=e_2}}{n_X n_Y} = p(X = e_1)p(Y = e_2). \tag{4.18}$$

In the derivation above we have made the assumption that the joint probability is defined through the number of combinations $n_X n_Y$. One could instead try to define the joint probability as $(n_{X=e_1} + n_{Y=e_2})/(n_X + n_Y)$, which is also normalized to 1 by summing over all e_1, e_2. However, for an event e_2 that is impossible, i.e., $n_{e_2} = 0$ we obtain $n_{X=e_1}/(n_X + n_Y)$, which can generally be finite. On the other hand, according to Eq. 4.11 the joint probability of an impossible event and any other event must vanish. This is only satisfied by the multiplicative ansatz we used above.

Based on this thought experiment with two unrelated random experiments and the resulting identity relating the joint probability to the marginal probabilities of the individual random variables we arrive at a definition for statistical independence:

Definition 2 (Independence)**.** *Two random variables X, Y are said to be independent iff the outcome of one does not influence the outcome of the other, i.e., $p(X|Y) = p(X)$ or equivalently $p(X, Y) = p(X)p(Y)$.*

Apparently, statistical independence refers to the special case where the joint probability is just the product of the two marginal probabilities. Statistical dependence on the other hand is the opposite case and can be defined through $p(X, Y) \neq p(X)p(Y)$.

4.1.3 Expectation and Moments

The mathematical "expectation" of a random variable X, denoted by $\mathrm{E}[X]$, is the average value that the random variable assumes over a large number of repeated random experiments under the same conditions, i.e., the random variable obeys the same probability distribution, or in other words the probabilities $p(e)$ do not change over the course of the repeated experiments.

The expectation is defined as

$$\mathrm{E}[X] = \sum_i p(X = e_i)e_i, \tag{4.19}$$

where $e_i \in E$ denotes the value associated with a particular event. The expectation value is

linear in the random variable X, i.e.,

$$\begin{aligned}
\mathrm{E}[X+Y] &= \sum_{i,j} p(X=x_i, Y=y_j)(x_i+y_j) && (4.20)\\
&= \sum_{i,j} p(X=x_i, Y=y_j)x_i + \sum_{i,j} p(X=x_i, Y=y_j)y_j && (4.21)\\
&= \sum_{i} p(X=x_i)x_i + \sum_{j} p(Y=y_j)y_j && (4.22)\\
&= \mathrm{E}[X] + \mathrm{E}[Y]. && (4.23)
\end{aligned}$$

It is obvious that

$$\mathrm{E}[aX] = a\mathrm{E}[X], \tag{4.24}$$

and due to the normalization of probability

$$\mathrm{E}[a] = a, \tag{4.25}$$

i.e., the expectation of a constant is the constant itself. In case of independent random variables one can also find a simplified expression for the expectation value of the product of random variables

$$\begin{aligned}
\mathrm{E}[XY] &= \sum_{i,j} p(X=x_i, Y=y_j)x_i y_j && (4.26)\\
&= \sum_{i,j} p(X=x_i)p(Y=y_j)x_i y_j && (4.27)\\
&= \sum_{i} p(X=x_i)x_i \sum_{j} p(Y=y_j)y_j && (4.28)\\
&= \mathrm{E}[X]\mathrm{E}[Y]. && (4.29)
\end{aligned}$$

For finite sets of samples we can define the so-called sample average as the sum over all observed values x_i divided by the total number of observations

$$\mathrm{E}[X] \approx \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{4.30}$$

This is motivated by the fact that we can rewrite the equation above as

$$\mathrm{E}[X] \approx \frac{1}{N}\sum_{e\in E} n_e x_e \approx \sum_{e\in E} p(e)x_e, \tag{4.31}$$

with x_e the value associated with event e and $p(e) \approx n_e/N$, cf. Eq. 4.7. For large N we therefore have equality

$$\mathrm{E}[X] = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{4.32}$$

As a measure of the variability of the random variable X, i.e., its average deviation from the typical value, we define the *variance* as the expected square deviation from the mean, i.e.,

$$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]. \tag{4.33}$$

Using the linearity of E we can derive the alternative form

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{E}[X^2 - 2X\mathrm{E}[X] + \mathrm{E}[X]^2] && (4.34)\\ &= \mathrm{E}[X^2] - 2\mathrm{E}[X]^2 + \mathrm{E}[X]^2 && (4.35)\\ &= \mathrm{E}[X^2] - \mathrm{E}[X]^2, && (4.36)\end{aligned}$$

that relates the variance to the means of the squared random variable and X itself. For a sum $X = Y + Z$ of two independent random variables we can define

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{E}[(Y+Z)^2] - \mathrm{E}[Y+Z]^2 \\ &= \mathrm{E}[Y^2] + \mathrm{E}[Z^2] + 2\mathrm{E}[Y]\mathrm{E}[Z] - \mathrm{E}[Y]^2 - \mathrm{E}[Z]^2 - 2\mathrm{E}[Y]\mathrm{E}[Z] && (4.37)\\ &= \mathrm{Var}[Y] + \mathrm{Var}[Z].\end{aligned}$$

This implies, in particular, for the sample average that

$$\mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}[X_i], \tag{4.38}$$

where we used that $\mathrm{Var}[aX] = a^2\mathrm{Var}[X]$. The variance of the sample average is therefore proportional to the sum of individual variances.

As a straight-forward generalization of the variance we define the *covariance* of two random variables as the expectation of the product of their deviations from the respective means

$$\mathrm{Cov}[X,Y] = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]. \tag{4.39}$$

Clearly, the variance is just a special case of the covariance, since $\mathrm{Cov}[X,X] = \mathrm{Var}[X]$, and if one interprets $C_{X,Y} = \mathrm{Cov}[X,Y]$ as a matrix containing the covariances between any number of random variables X, Y, the variances of the individual random variables are found on the diagonal of this covariance matrix. In an analogous calculation as for the variance we show that

$$\begin{aligned}\mathrm{Cov}[X,Y] &= \mathrm{E}[XY - X\mathrm{E}[Y] - Y\mathrm{E}[X] + \mathrm{E}[X]\mathrm{E}[Y]] && (4.40)\\ &= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y], && (4.41)\end{aligned}$$

and with Eq. 4.29 we have for the expectation value of products of independent random variables

$$\mathrm{Cov}[X,Y] = 0. \tag{4.42}$$

Thus, the covariance of independently distributed random variables vanishes. This gives rise to the interpretation of the covariance as a measure of *correlation*. Note that the opposite is not true, i.e., $\mathrm{Cov}[X,Y] = 0$ does not necessarily imply that X, Y are independent, since $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$ can also be fulfilled by accident with $p(X,Y) \neq p(X)p(Y)$.

Having discussed the most important expectations of one and two random variables we note that one calls these quantities the *moments* of a probability distribution. The n-th moment is defined in general as

$$M_n = \mathrm{E}[X^n], \tag{4.43}$$

and we can see immediately, that $M_0 = 1$ due to the normalization of probability. M_1 corresponds to the mean and $M_2 = \mathrm{E}[X^2] = \mathrm{Var}[X] + E[X]^2$ is related to the variance. Therefore, we have already discussed the moments of lowest orders. The entire series of moments is conveniently described in terms of the so-called generating function of moments

$$M(t) = \mathrm{E}[e^{tX}], \tag{4.44}$$

which satisfies $\frac{\mathrm{d}^n M}{\mathrm{d}t^n}\big|_{t=0} = M_n$. This property follows immediately from the series expansion of the exponential function. Given $M(t)$ one therefore automatically has knowledge of all moments of a probability distribution p.

4.1.4 Continuous Random Variables

A random variable is continuous if instead of a finite countable set the target space is uncountable. Let us here assume that $E = \mathbb{R}$. It is always possible to extend $E \subset \mathbb{R}$ to $\mathbb{R}$ by making every event in $\mathbb{R} \setminus E$ impossible. The probability distribution p, or probability density function (PDF), of a continuous random variable X is normalized such that

$$\int_{-\infty}^{\infty} p(x)\ \mathrm{d}x = 1, \tag{4.45}$$

which implies that some event $e \in \mathbb{R}$ will come to pass with certainty. We now define the cumulative distribution function

$$F(x) = P(X < x) = \int_{-\infty}^{x} p(y)\ \mathrm{d}y. \tag{4.46}$$

Apparently, $F(-\infty) = 0$ and $F(\infty) = 1$ and F is a monotonously increasing function, since

$$\frac{\mathrm{d}F}{\mathrm{d}x} = p(x) \geq 0. \tag{4.47}$$

From the definition it is immediately clear that F describes the probability that X assumes a value below the threshold x. Since $p(x)$ is the derivative of a probability we call it a probability density. If we take another look at the definition of F and p we find that $F : \mathbb{R} \to [0, 1]$ while $p : \mathbb{R} \to \mathbb{R}^+$, i.e., the target space of the cumulative distribution function is bounded as opposed to the probability density function. It is therefore often convenient to work with the cumulative distribution function instead.

We can apply the same trick as in defining F in Eq. 4.46 once more to obtain the probability for X to lie in an interval (x_1, x_2)

$$\tilde{F}(x_1, x_2) = P(x_1 < X < x_2) = \int_{x_1}^{x_2} p(y)\ \mathrm{d}y. \tag{4.48}$$

Note that this is generally equal to the probability for the corresponding closed interval $[x_1, x_2]$, since the probability assigned to each individual event is vanishingly small. The usefulness of Eq. 4.48 becomes clear if we partition $\mathbb{R}$ into mutually disjoint intervals I_j such that $\mathbb{R} = \cup_j I_j$ and $I_j \cap I_k = \emptyset\ \forall j, k$. We can now use Eq. 4.48 to assign probabilities $p(x \in I_j) = \tilde{F}(\min(I_j), \max(I_j)) > 0$ to each interval that obey

$$\sum_j p(x \in I_j) = 1. \tag{4.49}$$

Thus, at the loss of precision one can always transform a continuous random variable into a discrete random variable, which is convenient for numerical treatments.

The issue of determining random distributions of compound variables comes up rather frequently. For instance, given a random variable X and a function f we can construct a new random variable $Y = f(X)$ whose distribution depends somehow on the distribution of X and the function f. To determine this dependence we make use of the fact that the moments of distributions with finite support are unique, i.e., the sequence of all moments $M_n\ \forall n$ uniquely determines the distribution [130]. If all moments are the same then necessarily also the generating function must be the same, since the moments are the coefficients of the corresponding Taylor series. We begin with the generating function of moments from Eq. 4.44 that for the

continuous random variable Y has the form

$$\begin{aligned} M_Y(t) &= \int_{-\infty}^{\infty} e^{ty} p(y)\ \mathrm{d}y && (4.50)\\ &= \int_{-\infty}^{\infty} e^{tf(x)} p(f(x)) \frac{\mathrm{d}y}{\mathrm{d}x}\ \mathrm{d}x && (4.51)\\ &= \int_{-\infty}^{\infty} e^{tf(x)} p(x)\ \mathrm{d}x. && (4.52) \end{aligned}$$

This implies immediately that $p(y) = (\frac{\mathrm{d}y}{\mathrm{d}x})^{-1} p(x(y)) = (\frac{\mathrm{d}f}{\mathrm{d}x})^{-1} p(x(y))$. A special case of this is the case $Y = F(X)$ with F the cumulative distribution function of X, cf. Eq. 4.46. We compute

$$\begin{aligned} p(y) &= \left(\frac{\mathrm{d}F}{\mathrm{d}x}\right)^{-1} p(x(y)) \\ &= \frac{p(x(y))}{p(x(y))} = 1, \end{aligned} \tag{4.53}$$

which implies that Y is uniformly distributed on $[0, 1]$. Similarly, one obtains for $Y = F^{-1}(Z)$ with Z distributed over $[0, 1]$ the result

$$p(y) = p(x(y))p(F(y)). \tag{4.54}$$

If the distribution of Z is uniform, then $p(F(y)) = p(z) = 1$ and therefore $p(y) = p(x(y))$, i.e., the two distributions are the same. This result is of importance for numerically sampling from a particular random number distribution, since given F^{-1}, only a uniform random number generator is required.

We explain in the following one particular example of this that is relevant for the discussion in Chapter 7. There, the necessity arises to sample from a uniform distribution on a circle. Given a uniform random number generator, one can easily construct a random distribution in two dimensions by defining $Z = (X, Y)$, where $X, Y \in [0, 1)$ are both uniformly distributed random numbers. Z is then uniformly distributed on $[0, 1) \times [0, 1)$, i.e., on a square. The most trivial algorithm to achieve a uniform distribution on a circle only would be to embed said circle in a square and throw away all samples that do not lie within the circle. The method based on the cumulative distribution function described above, however, offers a much more elegant solution. Given the area of the circle with radius R as $A = \pi R^2$, we can formally write the probability distribution function as $p(x, y) = \frac{1}{\pi R^2}$. Normalization now demands that

$$1 \stackrel{!}{=} \int_{\bigcirc} p(x, y)\ \mathrm{d}x\mathrm{d}y = \int_0^{2\pi} \left[\int_0^R p(x(r, \phi), y(r, \phi))\ r\mathrm{d}r \right] \mathrm{d}\phi, \tag{4.55}$$

where we assume that the circle is centered at the origin. Due to the constant expression for p, the integral is readily computed

$$1 \stackrel{!}{=} \int_0^R \frac{2r}{R^2} \mathrm{d}r. \tag{4.56}$$

Considering the problem in polar coordinates from the start, uniformity in x, y within the circle immediately implies uniformity in ϕ, so that we can write $p(r, \phi) = p(r)p(\phi) = \frac{p(r)}{2\pi}$. It then follows that

$$1 \stackrel{!}{=} \int_0^R p(r)\ \mathrm{d}r, \tag{4.57}$$

and by comparison $p(r) = \frac{2r}{R^2}$. The cumulative distribution function is then $F(r) = \frac{r^2}{R^2}$, and the inverse yields $F^{-1}(z) = R\sqrt{z}$. Sampling this with z uniformly distributed on $[0, 1)$ then generates the desired distribution of $p(r)$, and with $(x, y) = (r\cos(\phi), r\sin(\phi))$ we obtain uniformly distributed samples in the circle with radius R.

4.2 Information Theory

Information theory is a highly interdisciplinary field living in the intersection of mathematics, physics and computer science. Borrowing ideas from Boltzmann's statistical mechanics it was first established by Claude Shannon out of the desire to quantify the amount of information in a given message. The usefulness of such considerations is apparent considering the need for data compression, error-robust communication and cryptography.

The foundation of information theory lies in the definition of Shannon entropy

$$H = -\sum_i p_i \log_2 p_i, \tag{4.58}$$

where p_i are the probabilities/frequencies of letters in the underlying alphabet. The base 2 of the logarithm defines the units of entropy, here bits. It has been shown by Shannon in his source coding theorem [133] that the potential for data compression, i.e., the minimal length of a given message, is given by the Shannon entropy.

The form of Eq. 4.58 resembles that of an expectation value, i.e.,

$$H = -\mathrm{E}_p[\log_2 p]. \tag{4.59}$$

The quantity $I(e_i) = -\log_2 p_i$ is called the *information content* of the letter with index i. Let us assume that a letter appears with probability 1. Then, $I(e_i) = 0$, i.e., the letter contains no information since it is the only possibility. Therefore, information cannot be transmitted via a single message that contains only one letter (no spaces) without providing additional information that equips the length of the message with a specific meaning. In the opposite case where the probability p_i approaches zero the information content diverges to positive infinity. This behavior supports the following interpretation. The information content is a measure of the amount of surprise the reader experiences when encountering the corresponding letter. Letters that appear all the time generate no surprise, unlike those that appear only rarely.

One can show that the definition of the information content is the only function that combines these properties. To recapitulate, I is a function of $p \in [0, 1]$, where an event with $p = 1$ contains no information, and I is inversely proportional to p. Moreover, the information content of any two independent events must equal the sum of the information contained in each of the individual events. These conditions can be written more formally as:

i) $p = 1 \Rightarrow I = 0$,

ii) $\frac{\mathrm{d}I}{\mathrm{d}p} < 0$,

iii) x, y independent $\Rightarrow I(x, y) = I(x) + I(y)$.

The logarithm is the only function satisfying condition iii) and while there is freedom in the choice of the base, this is only a multiplicative factor due to

$$\log_n x = \frac{\log x}{\log n}. \tag{4.60}$$

Typically, one chooses base 2, which defines the unit of information to be the *bit*.

Since entropy is the expectation value of the information content, it can be interpreted as the mean information contained in a single letter. The possible values of H are bounded, since $p \leq 1$, $\log p \leq 0$ and therefore $H \geq 0$. We show now that there is also an upper bound. Let p_i be the probabilities for events e_i, such that

$$\sum_i p_i = 1. \tag{4.61}$$

Then, we define the Lagrangian function

$$L = H(p_i) - \lambda \left(\sum_i p_i - 1 \right). \tag{4.62}$$

An extremum is obtained at

$$\text{grad } L = 0, \tag{4.63}$$

which evaluates to

$$-\log p_i - 1 - \lambda \stackrel{!}{=} 0, \tag{4.64}$$

$$\left(\sum_i p_i - 1 \right) \stackrel{!}{=} 0. \tag{4.65}$$

The first equation is satisfied for $\log p_i = -(1+\lambda)$ or $p_i = 2^{-(1+\lambda)}$. Using the second equation we have

$$\sum_i 2^{-(1+\lambda)} = N 2^{-(1+\lambda)} \stackrel{!}{=} 1, \tag{4.66}$$

and therefore $\lambda = -\log 1/n - 1 = \log n - 1$, which yields the final result $p_i = 1/N$. This means that the maximal entropy is obtained for a uniform probability distribution, which makes sense considering that we defined it as the average information content or average "surprise". This is clearly maximal if all events are equally likely to occur. The maximal value is given by

$$H_{\max} = H(p_{\text{unif}}) = -\sum_{i=1}^{N} \frac{1}{N} \log \frac{1}{N} = \log N. \tag{4.67}$$

Hence, the more possible events exist, the higher the average information becomes, since each individual event is less likely given the normalization of the uniform distribution. This is therefore consistent with condition ii) above. We can now summarize that H is bounded by $0 \leq H \leq \log N$.

By replacing the probability for the single variable X with the joint probability $p(x, y)$ of two random variables X, Y in the definition of entropy, cf. Eq. 4.58, we can define the joint entropy as

$$H(X, Y) = -\sum_{x,y} p(x, y) \log(p(x, y)). \tag{4.68}$$

If X and Y are independent, the joint probability is the product of the marginal probabilities $p(x, y) = p(x)p(y)$ and therefore

$$H(X, Y) = -\sum_{x,y} p(x)p(y) \log(p(x)p(y)) = -\sum_{x,y} p(x)p(y) \log(p(x)) - \sum_{x,y} p(x)p(y) \log(p(y)) \tag{4.69}$$

$$= -\sum_x p(x) \log(p(x)) - \sum_y p(y) \log(p(y)) = H(X) + H(Y), \tag{4.70}$$

which is not surprising, since the information content is additive and H is simply the expectation value of I.

In the opposite case, where X and Y are not independent Eq. 4.70 is not satisfied. We therefore define the difference between the sum of marginal entropies and the joint entropy as

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{4.71}$$

$$= -\sum_{x,y} p(x,y)\log(p(x)p(y)) + \sum_{x,y} p(x,y)\log(p(x,y)) \tag{4.72}$$

$$= -\sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right), \tag{4.73}$$

where we used the identities

$$H(X) = -\sum_{x} p(x)\log(p(x)) = -\sum_{x,y} p(x,y)\log(p(x)) \tag{4.74}$$

$$H(X) + H(X) = -\sum_{x,y} p(x,y)\log(p(x)p(y)). \tag{4.75}$$

$I(X;Y)$ is called the *mutual information* of the random variables X, Y and encodes the degree to which X, Y are dependent on each other.

It follows immediately from Eq. 4.70 and Eq. 4.71 that $I(X;Y) = 0$ for independent random variables, which supports the name "mutual information" that we prematurely assigned. In order to give meaning to the name we have to understand the relationship between $H(X,Y)$ and $H(X)+H(Y)$. We show now that $H(X,Y)$ is bounded from above by $H(X)+H(Y)$. Since $H(X,Y)$ is mathematically the same as $H(X)$, the only difference being that it is defined with respect to the joint probability, we already know the maximum value that $H(X,Y)$ can take to be $\log(N_x N_y)$, where $N_i = |E_i|$ for $i \in \{x,y\}$. Incidentally, this is just the sum of the maxima of $H(X)$ and $H(Y)$. However, this does not guarantee $H(X,Y) \leq H(X) + H(Y)$ for any case other than the uniform distribution. For the general proof we use Jensen's inequality [132], which states that for a convex function f and a function g the following inequality holds

$$f\left(\int g(x)\ \mathrm{d}x\right) \leq \int f\left(g(x)\right)\ \mathrm{d}x. \tag{4.76}$$

In a discretized version the same holds true for sums instead of integrals, which can be directly related to an inequality for expectation values

$$f\left(E[g]\right) \leq \mathrm{E}\left[f \circ g\right]. \tag{4.77}$$

Noting that the definition of the mutual information can be written in terms of expectation values like so

$$I(X;Y) = -\mathrm{E}_{x,y}\left[\log\left(\frac{p(x,y)}{p(x)p(y)}\right)\right], \tag{4.78}$$

where $\mathrm{E}_{x,y}$ implies the expectation value over the joint probability distribution $p(x,y)$, we can immediately conclude

$$I(X;Y) = \mathrm{E}_{x,y}\left[\log\left(\frac{p(x)p(y)}{p(x,y)}\right)\right] \tag{4.79}$$

$$\geq \log\left(E_{x,y}\left[\left(\frac{p(x)p(y)}{p(x,y)}\right)\right]\right) \tag{4.80}$$

$$= \log\left(\sum_{x,y} p(x,y)\frac{p(x)p(y)}{p(x,y)}\right) \tag{4.81}$$

$$= 0. \tag{4.82}$$

The positivity of the mutual information together with its definition, cf. Eq. 4.71, immediately implies that

$$H(X,Y) \leq H(X) + H(X), \tag{4.83}$$

where equality holds only if $p(x,y) = p(x)p(y)$. This result can be used to motivate a more intuitive understanding of these information measures in terms of a pictorial representation, see Fig. 4.1. The marginal entropies $H(X)$ and $H(Y)$ encode the information contained in each single random variable represented by two circles, while the joint entropy $H(X,Y)$ encodes the total information content of both, which is illustrated as the total area covered by the two circles. The overlap of information between single random variables, i.e., the information that one random variable contains about the other is the mutual information $I(X;Y)$. In the figure, this is represented as the intersection between the two circles.

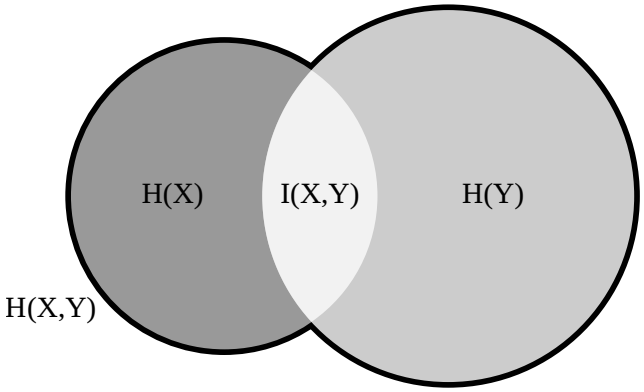

Figure 4.1: Illustration of the relationship between the different information measures. The joint entropy $H(X,Y)$ measures the information in both random variables X,Y. Each individual random variable encodes information $H(X)$ and $H(Y)$, respectively. The overlap is the information that either one random variable contains about the other. This is the mutual information $I(X;Y)$. The relationships between all four quantities are encoded in the shaded areas in the figure.

4.3 Machine Learning

Machine learning is the process of extracting information from data that is not obvious a priori. In particular, machine learning can be used to make predictions or to find structure in large data sets. In the latter context the term "big data" has become rather popular. Two large fields in machine learning are supervised and unsupervised learning, which are distinguished by the type of data that is available. In supervised learning the data set is comprised of a pair (X,Y), where X are data points and Y are corresponding labels that assign some value to these data point. These labels could, e.g., be measured values of some quantity that is measured according to the parameters given in X, or categories, e.g., names of flowers if X contains properties of plants. In the former case we are faced with a regression task, where we try to obtain a continuous function that satisfies $Y = f(X)$. The latter case is a classification task, where the function f is discrete-valued. Unsupervised learning on the other hand, stands out through the fact that the label Y is not available. Therefore, the idea is to find an organization of the data X into different categories without knowing their names and which information in X they refer to. Some of these ideas will be discussed in Chapter 7 in the context of topological phase diagrams.

Possibly the simplest example of a machine learning application is the classic coin toss experiment. The setup is as follows: a coin is tossed a number N times in a controlled environment

and each time the observed one of the two possible outcomes (heads or tails) is recorded. This creates a data set that can be used to approximate the probabilities $p(\text{“heads”}), p(\text{“tails”})$.

We start from our hypothesis that the coin is fair, i.e., that it is equally likely to land on either side. Therefore, our hypothesis is $p(\text{“heads”}) = p(\text{“tails”}) = 0.5$. Given this hypothesis we can now compute the so-called likelihood of the recorded data and vary the hypothesis such that the data is most likely. This approach is known as the method of *maximum likelihood* and will be explained in the following.

4.3.1 Bayesian Statistics

We have learned about the information theoretical surprise earlier in the context of Shannon's fundamental definition of the information content. In Bayesian statistics there is a similar concept of surprise. Given a set of possible hypotheses h, the so-called prior probability $p(h)$ encodes the prior knowledge of the observer. Given new data x, the probability that the observer confides in a specific hypothesis is given by $p(h|x)$, where

$$p(h|x) = \frac{p(x|h)p(h)}{p(x)}, \tag{4.84}$$

according to Bayes' theorem (Eq. 4.16). If the data is entirely unsurprising to the observer it contains no new information. In that case, the posterior probability distribution should coincide with the prior distribution, i.e.,

$$p(h|x) \stackrel{!}{=} p(h). \tag{4.85}$$

Therefore, the amount of surprise contained in the data x can be formalized in terms of a distance function d between two probability distributions

$$I(x) = d[p(h|x), p(h)]. \tag{4.86}$$

To be an adequate descriptor of the information contained in the data x, the function $d(p, q)$ should have the following properties:

i) $p = q \Leftrightarrow d(p, q) = 0$,

ii) $d \geq 0$.

One example for a function satisfying these properties is the Kullback-Leibler divergence [134]

$$D_{\mathrm{KL}}(p \parallel q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right). \tag{4.87}$$

We can confirm one direction of the first property above easily, since

$$\begin{aligned} D_{\mathrm{KL}}(p \parallel p) &= \sum_x p(x) \log\left(\frac{p(x)}{p(x)}\right) \\ &= \sum_x p(x) \log(1) \\ &= 0. \end{aligned}$$

The other direction is not as simple to see and, in fact, follows from the proof of the second property, which requires

$$\sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right) \geq 0. \tag{4.88}$$

We can show this using the inequality $x-1 \geq \log(x)$, that can be proven by applying the strictly monotonous exponential function to both sides and defining

$$g(x) = e^{x-1} - x. \tag{4.89}$$

Clearly, $g(x)$ vanishes only for $x = 1$. Since $g'(x) = e^{x-1} - 1$ is > 0 for $x > 1$ and < 0 for $x < 1$, $g(x)$ has a global minimum at $x = 1$, i.e., $g(x) \geq 0$, and therefore $x - 1 \geq \log(x)$. This proves the upper bound for the logarithm. Therefore, we have

$$D_{\mathrm{KL}}(p \parallel q) \geq -\sum_x p(x) \left(\frac{q(x)}{p(x)} - 1 \right) \tag{4.90}$$

$$= \sum_x (p(x) - q(x)) = 0, \tag{4.91}$$

since $\sum_x p(x) = \sum_x q(x) = 1$. Equality is found only if $p(x)/q(x) = 1$ for all x as shown above, which also proves the first property.

Having found a suitable expression for the function d, the Bayesian measure of information can thus be expressed as the Kullback-Leibler divergence between the posterior and the prior distributions

$$I(x) = D_{\mathrm{KL}}(p(h|x) \parallel p(h)). \tag{4.92}$$

Note that D_{KL} is not symmetric under exchanging p and q, since

$$D_{\mathrm{KL}}(q \parallel p) = \sum_x q(x) \log \left(\frac{q(x)}{p(x)} \right) \tag{4.93}$$

$$= \sum_x q(x) \log \left(q(x) \right) - \sum_x q(x) \log \left(p(x) \right) \tag{4.94}$$

$$\neq \sum_x p(x) \log \left(p(x) \right) - \sum_x p(x) \log \left(q(x) \right). \tag{4.95}$$

Therefore, the Kullback-Leibler divergence is not a distance in the strictly mathematical sense.

4.3.2 Bayesian Inference

Surprisingly, there is more than one possible interpretation of probability, see, e.g., Ref. [131]. In addition to the so-called "frequentist" viewpoint that we introduced in the beginning of this chapter following Kolmogorov's axiomatic definition, the Bayesian viewpoint is considered more general. In the following we briefly explain the conceptual differences between both approaches.

The frequentist interpretation of statistics assumes that a probability is always the theoretical manifestation of a repeatable random experiment, meaning that it can be measured as the limit of a series of random experiments. This obviously excludes common colloquial uses of the word *probability* like "what are the odds that it rains today?" as it is described, e.g., in [131, 132]. Obviously, the weather that presents itself on a given day is not a repeatable random experiment, since many parameters will have changed on the next day or even on the same day one year later. One can therefore only take into account past facts to arrive at some sort of expectation. The Bayesian view on the other hand includes all of these other use cases with a more loose definition of probability as a quantified measure of one's belief in a particular scenario [131].

As the name suggests, Bayesian inference is a prediction scheme based on Bayes' theorem. In order to introduce the conventional nomenclature we briefly recapitulate. Bayes' theorem establishes a relation between the conditional probabilities $p(X|Y)$ and $p(Y|X)$. Here, we define

two random variables x and h, where x represents the data and h the hypothesis that is supposed to explain the data. We then apply Bayes' theorem (Eq. 4.16) and obtain

$$p(h|x)p(x) = p(x|h)p(h). \tag{4.96}$$

Here, $p(h|x)$ is called the *posterior* probability that describes the odds that the hypothesis is correct given the data, $p(x)$ is the *evidence* that is independent of the hypothesis and therefore simply a constant normalization factor, $p(x|h)$ is the *likelihood* that the data x can be explained by the hypothesis h and $p(h)$ is the *prior*, i.e., the prior belief in the hypothesis. Apparently, the posterior probability can be inferred from a reasonably large data set by counting the number of data points compatible with the hypothesis. In order to make predictions one needs to obtain a generative model, i.e., a model that can produce new data points that follow the same underlying law as the original data. This is achieved by finding the optimal hypothesis in terms of the best possible description of the available data. We can quantify the quality of a given model h through the likelihood as follows

$$p(h|x) = \frac{p(x|h)p(h)}{p(x)}, \tag{4.97}$$

which in words is often expressed as "posterior equals likelihood times prior", where the constant evidence is simply a normalization factor. Note that instead of a probability as a function of the data x, the likelihood is actually considered a function of the hypothesis h, since the data is usually a constant. In practice, a hypothesis is represented by a statistical distribution that depends on a set of parameters θ.

A typical update scheme can be put into place by iteratively computing the posterior probability from Eq. 4.97 as new data is coming to light. In this case, the prior $p(h)$ is computed without the knowledge of the new data and subsequently updated in every step. The information gain that is achieved as new data is added is given by the Kullback-Leibler divergence according to Eq. 4.92.

The predictive quality of the model is exploited by asking for the distribution of a new data point (x', y') that can be expressed in terms of the posterior distribution

$$p(x', y'|x, y) = \sum_h p(y'|x', h)p(h|x, y). \tag{4.98}$$

Here, the difference between the Bayesian and frequentist approaches becomes clear. Instead of selecting a "best fit" hypothesis and basing all further consideration onto this choice, the Bayesian prediction marginalizes over all possible hypotheses, thereby reflecting the uncertainty in the choice of the model.

4.3.3 Regression

Regression problems are equivalent to the problem of fitting data to a model. The basic task that one is attempting to achieve is to find a model or hypothesis h, that depends on weights w such that

$$y = h_w(x). \tag{4.99}$$

Ideally, this relation holds for all x, y, however, in practice one can only hope to achieve a fraction close to 1 due to an in practice almost guaranteed imperfect choice of the model. In most cases, the purpose of the model is not to reproduce the data, but to generalize to new data x' to make predictions

$$y' = h_w(x') \tag{4.100}$$

for unknown $x' \notin X$.

In order to be explicit we look at a specific hypothesis and assume for simplicity a one-dimensional (i.e., $x, y \in \mathbb{R}$) polynomial model

$$h_w(x) = \sum_{i=0}^{n} w_i x^i, \tag{4.101}$$

where n is the degree of the polynomial. Given data $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, Eq. 4.100 represents a linear system of equations with $n+1$ unknowns and N equations. It is clear that in case $N \leq n+1$ a solution exists, where equality can be satisfied for all x_i, which reduces the problem to a simple interpolation. In more practical examples the number of data points is larger than the number of model parameters, which also reflects the desire to find a simple explanation for the data. In that case we define the regression error as

$$E(w) = f[\mathbf{y} - h_w(\mathbf{x})], \tag{4.102}$$

where f is a positive, monotonous function that satisfies $f(0) = 0$. Intuitively, the sum of squares seems like a good choice, i.e.,

$$E(w) = \sum_{i=1}^{N} (y_i - h_w(x_i))^2, \tag{4.103}$$

since it penalizes, in particular, large deviations from the correct solutions. The solution to the regression problem, i.e., the best fit, is to minimize the error $E(w)$, since this reveals the model that represents the data best w.r.t. the chosen error function. For the sum of squares error function this problem is mathematically well-defined as we shall see below.

Let us write out the system of equations that we want to solve (Eq. 4.100) for all data points (x, y)

$$\mathbf{y} = Mw, \tag{4.104}$$

with $M_{ij} = x_i^{j-1}$ for $1 \leq i \leq N$ and $1 \leq j \leq n+1$. We thus have to minimize the equation

$$E(w) = (\mathbf{y} - Mw)^T(\mathbf{y} - Mw) \tag{4.105}$$

$$= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T Mw - w^T M^T \mathbf{y} + w^T M^T Mw \tag{4.106}$$

with respect to w. The first term is a constant and we have for the others

$$\nabla (wM\mathbf{y}) = \nabla \left(\mathbf{y}^T Mw\right) = \nabla \left(\sum_{i,j} y_i M_{ij} w_j \right) = y_i M_{ij} \mathbf{e}_j = M^T \mathbf{y} \tag{4.107}$$

$$\nabla \left(w^T M^T Mw\right) = \nabla \left(\sum_{i,j,k} w_i M_{ji} M_{jk} w_k \right) = \sum_{i,j,k} \mathbf{e}_i M_{ji} M_{jk} w_k + \sum_{i,j,k} w_i M_{ji} M_{jk} \mathbf{e}_k \tag{4.108}$$

$$= 2 \sum_{i,j,k} \mathbf{e}_i M_{ji} M_{jk} w_k = 2M^T Mw. \tag{4.109}$$

Combining everything into one equation then yields

$$0 \stackrel{!}{=} 2(M^T Mw - M^T \mathbf{y}). \tag{4.110}$$

The solution to the regression problem is thus given by the solution to the system of equations

$$M^T Mw = M^T \mathbf{y}. \tag{4.111}$$

If $M^T M\mathbf{a} = 0$ for some $\mathbf{a}$, then $0 = \mathbf{a}^T M^T M\mathbf{a} = \| M\mathbf{a} \|^2$, which requires $\mathbf{a} = 0$ due to M being rank $n+1$ (linear independence of monomials). Hence, $M^T M$ is invertible and Eq. 4.111 has a unique solution.

The solution of Eq. 4.111 seems like the straight-forward way to do linear regression. However, it is very specific to the least squares error function, which is often subsituted against other functions.

We will discuss a different approach now that is usually applied in machine learning. To this end we go back to Eq. 4.102 and try to compute the minimum iteratively. Starting from an arbitrary point w_0 we update the weights in a greedy fashion by moving in the direction of the largest descent of the error function. Therefore, we compute the gradient with respect to w and since it always points in the direction of the largest ascent of $E(w)$ we define

$$w_{m+1} = w_m - \alpha \nabla E(w_m). \tag{4.112}$$

Here, $\alpha > 0$ is the learning rate, that determines the speed of convergence. Large (small) α means large (small) steps in w in each iteration. However, too large of a value for the learning rate leads to overshooting and can impede performance. In general, this iterative approach, while being applicable to any error function, comes with all the disadvantages of high dimensional nonlinear minimization methods, such as getting stuck in local minima. As a possible remedy to this problem one uses the so-called hyperparameter optimization, which in this case means repeating the iteration for different values of w_0, α. This avoids running into the same local minima and provides a better estimate of the global minimum.

4.3.4 Loss Function

We take a closer look at the error function (loss function), that we had defined previously as the averaged squared difference between the feature input and the target output. It is common to denote the loss function with the letter J, i.e.,

$$J_{\text{MSE}}(w) = \frac{1}{2N} \sum_{i=1}^{N} (h_w(x) - y)^2, \tag{4.113}$$

where the factor $1/2N$ is for convenience only and does not change the location of the minimum. Dividing by N has the benefit that errors for data sets of different sizes can be compared with one another. For the gradient-descent algorithm we need to define the gradient of Eq. 4.113

$$\frac{\partial J_{\text{MSE}}}{\partial w_j} = \frac{1}{N} \sum_{i=1}^{N} (h_w(x_i) - y_i) \frac{\partial h_w(x_i)}{\partial w_j}, \tag{4.114}$$

which for the polynomial model can be expressed as

$$\frac{\partial J_{\text{MSE}}}{\partial w_j} = \frac{1}{N} \sum_{i=1}^{N} (h_w(x_i) - y_i) x_i^j. \tag{4.115}$$

The update scheme per iteration now reduces to a matrix-vector product, which is $\mathcal{O}(nN)$, where n is the number of features and N the number of data points.

4.3.5 Classification

In contrast to regression, where feature vectors are mapped to a continuous spectrum of labels, the process of classification assigns each feature vector a unique class label, where the set of class

labels is discrete. This has the immediate consequence that the prediction of the model must be rather stable, since for sensible applications the prediction shouldn't change if the feature vector is modified by only a small amount.

The simplest case is a classification problem with only two classes, for example, say, "cat" and "dog". The set of class labels is then $\{\text{cat}, \text{dog}\}$ and the classification problem is such that a hypothesis $h_w(x)$ parameterized with weights w is optimized such that $h_w(x) = y$ for the known data. For regression any hypothesis was valid in principle. Here, it turns out that we have to be more restrictive. Let, e.g., h_w be linear in x. Then, we have

$$h_w(x + \delta x) = h_w(x) + h_w(\delta x) = y_x + y_{\delta x} \stackrel{!}{=} y. \tag{4.116}$$

This is, however, wildly unreasonable, since it implies that the label for small feature vectors ($|\delta x| \ll 1$) must be 0. This follows by assuming x lies far away from the transition to another class. Therefore, the hypothesis must be non-linear. Assuming a mapping dog = 1 and cat = 0 we can argue that any function $\sigma : x \mapsto [0, 1]$ should be able to serve as a hypothesis via the transformation

$$h_w(x) = \text{round}(\sigma(\tilde{h}_w(x))) = \begin{cases} 0 & \text{if} \quad \sigma(\tilde{h}_w(x)) < 0.5 \\ 1 & \text{if} \quad \sigma(\tilde{h}_w(x)) \geq 0.5. \end{cases} \tag{4.117}$$

The function $\tilde{h}_w : \mathbb{R}^n \to \mathbb{R}$ can again be any function and for the *activation function* σ we can choose, e.g., the Heaviside step function $\theta(x)$. It is more common, though, to relax the classification by allowing intermediate values to account for falsely labeled data. In this case we define

$$h_w(x) = \sigma(\tilde{h}_w(x)) = \frac{1}{1 + e^{-\tilde{h}_w(x)}}, \tag{4.118}$$

where σ is a sigmoid function. Clearly, $h_w(x) \in [0, 1]\ \forall x$. Depending on the steepness of $\tilde{h}_w$ there will be a broad or narrow regime with $h_w(x)$, significantly larger than 0 and smaller than 1, that represents the uncertainty of the model around the decision boundary between the two classes.

In order to quantify the fit of the model to the data we define the loss function, i.e., the classification error as

$$J(x) = \frac{1}{N} \sum_{i=1}^{N} \left(-y_i \log(h_w(x_i)) - (1 - y_i) \log(1 - h_w(x_i)) \right), \tag{4.119}$$

which guarantees that each correct classification contributes 0, since for $y_i \in \{0, 1\}$ we have

$$J_i = -y_i \log(y_i) - (1 - y_i) \log(1 - y_i) = 0. \tag{4.120}$$

The maximal value is taken if $h_w(x) = 0.5$, i.e., if the prediction falls right in the middle between the two classes

$$J_i = \log(2). \tag{4.121}$$

In principle, one has a freedom in the choice of the loss function, and of course we could have chosen also the sum of squares. However, as we will see in the following, the choice of Eq. 4.119 leads to a very convenient representation of the gradient descent update formula for the sigmoid activation function. For the gradient we obtain

$$\frac{\partial J}{\partial w_j} = \frac{1}{N} \sum_{i=1}^{N} \left(-\frac{y_i}{h_w(x_i)} + \frac{(1 - y_i)}{1 - h_w(x_i)} \right) \frac{\partial h_w(x_i)}{\partial w_j}, \tag{4.122}$$

and with $\partial\sigma/\partial x = -e^{-x}/(1+e^{-x})^2 = \sigma(x)(1-\sigma(x))$ we have

$$\frac{\partial J}{\partial w_j} = \frac{1}{N}\sum_{i=1}^{N}\left(-\frac{y_i}{h_w(x_i)} + \frac{(1-y_i)}{1-h_w(x_i)}\right) h_w(x_i)(1-h_w(x_i))\frac{\partial \tilde{h}_w(x_i)}{\partial w_j} \tag{4.123}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(-y_i(1-h_w(x_i)) + (1-y_i)h_w(x_i)\right)\frac{\partial \tilde{h}_w(x_i)}{\partial w_j} \tag{4.124}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(h_w(x_i) - y_i\right)\frac{\partial \tilde{h}_w(x_i)}{\partial w_j}. \tag{4.125}$$

Assuming a linear hypothesis $\tilde{h}_w(x) = w^T x$ the gradient simplifies to

$$\frac{\partial J}{\partial w_j} = \frac{1}{N}\sum_{i=1}^{N}\left((h_w(x_i) - y_i)x_i\right), \tag{4.126}$$

which is the same result as for the mean square error and can therefore also be implemented as a simple matrix-vector product.

4.3.6 Neural Networks

Finally, we want to quickly introduce the concept of neural networks for supervised learning, mainly to motivate why this is not used in our analysis later on. Generally speaking, a neural network represents just another parameterization of a fit function (hypothesis) so that most of the ideas from the previous sections still apply. The aim is again to minimize a loss function in order to find the optimal hypothesis that provides the best description of the training data while allowing for the best possible amount of generalization.

The idea is perhaps best described visually, therefore we show a graphical illustration of a neural network in Fig. 4.2. Subfigure **a** shows the full neural network which is composed of individual neurons represented by yellow circles that are arranged in columns. Each column bounded by a blue box corresponds to one layer of the network with n being the dimension of the first layer, i.e., the number of neurons. Neurons from adjacent layers are connected by lines that represent the flow of information through the network. In a feed-forward network, which we are describing here, information can flow only in one direction, i.e., from the left to the right. Consequently, the first layer on the left is the input layer where the original data is inserted into the model. The number of neurons n must therefore correspond to the dimension of the vector space in which the data is represented. On the other hand, the right-most layer with dimension m must be the output layer, where the data exits the neural network. m must therefore be equal to the dimension of the label vectors. For a regression task with a scalar label we would have, e.g., $m = 1$, for classification the number of output neurons is usually chosen equal to the number of different class labels such that each entry in the output vector corresponds to the probability for the input data to correspond to that particular class. All layers in between the input and output layers are called hidden layers since they provide no interface to the outside and can only be accessed from within the neural network. We indicated one hidden layer in the figure and added a gray box representing an arbitrary number of additional hidden layers.

In Fig. 4.2 **b**, we show how each neuron operates on the data, i.e., how the information flow through the network is affected by individual neurons. Here, w_{ij} are matrix elements of matrices W^l, where $l = 1, 2, \ldots$ is the number of the layer where the data originates, e.g., W^1 corresponds to the data flow from the input layer to the first hidden layer. Given an input vector $\mathbf{x}$, the data

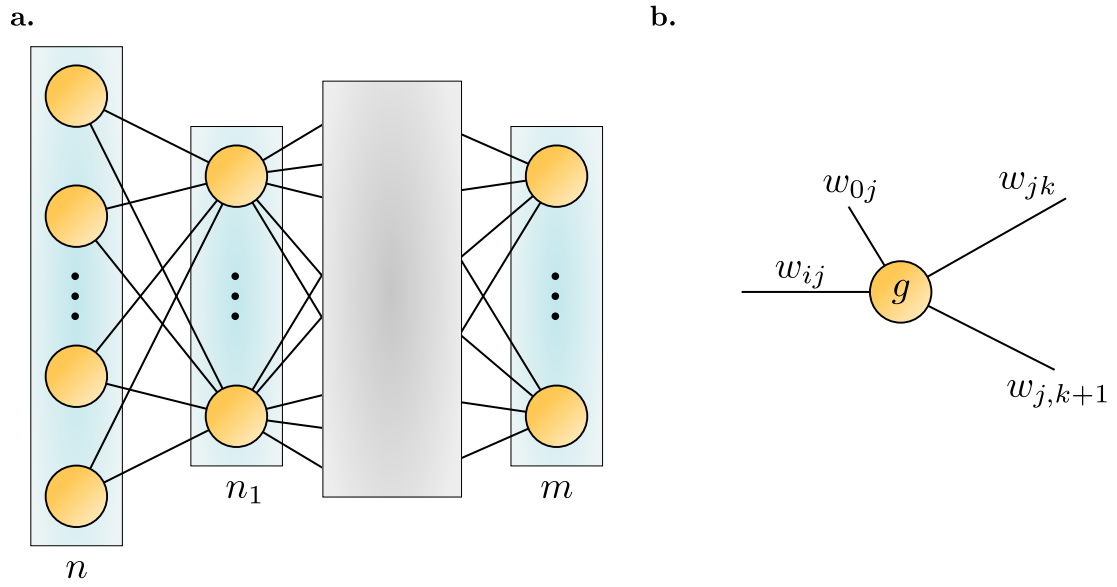

Figure 4.2: We illustrate the general architecture of artificial neural networks. **a.** General layout featuring a number of layers (blue boxes) consisting of a number of neurons (yellow circles) each. Here, we denote n as the number of neurons in the input layer and m the number of neurons in the output layer. The gray box is a placeholder for any number of additional hidden layers. Neurons from adjacent layers are connected by lines indicating information flow. In a feed-forward network the information flow is always directed from the left to the right, i.e., input towards output. **b.** A single neuron performs several operations on the data. The neuron j receives data from neuron i with weight w_{ij} and adds a bias w_{0j}. Then, an activation function g is applied and the result is output to neurons k and $k+1$ with their respective weights. The free parameters of the neural network are matrices $W^l = w^l_{ij}$, where l labels the layer where the data originates.

that appears at the first hidden layer can be represented as another vector $\mathbf{x}^1$ of dimension n_1 that is determined as

$$\mathbf{x}^1 = {W^1}^T \mathbf{x} = \sum_{i=0}^{n} \sum_{j=1}^{n_1} w_{ij} x_i \mathbf{e}_j, \tag{4.127}$$

where $\mathbf{e}_j$ is the j-th unit vector and $x_0 = 1$. The latter is just a convenient way of accounting for the bias w_{0j} that represents a constant offset of the data. At neuron j in the second layer the incoming data is therefore determined by

$$x^1_j = \sum_{i=1}^{n} w_{ij} x_i + w_{0j}, \tag{4.128}$$

which corresponds simply to a linear model. A neural network without a hidden layer is therefore identical to a linear model. Adding a hidden layer repeats the same process a second time, which results in a redundancy since the same linear model is expressed through more independent variables. This can be repeated many times, however, it is straight-forward to show that the model will always remain linear. In order to account for non-linearities, each neuron in hidden layers performs an additional task similar to what we saw in logistic regression, that is, to apply a so-called activation function g that adds a non-linearity and therefore allows even rather simple networks to capture non-trivial data.

Typical choices for activation functions are for example the sigmoid function that we already know

$$g_{\text{sigmoid}}(x) = \frac{1}{1 + e^{-x}}, \tag{4.129}$$

which corresponds to the Fermi-Dirac distribution function and maps input values $x \in \mathbb{R}$ to the interval $[0, 1]$, the Tanh function

$$g_{\text{Tanh}}(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \tag{4.130}$$

that has a similar shape, although values range anywhere in the interval $[-1, 1]$. Very popular is also a rather simple variant, the so-called rectified linear unit "ReLU"

$$g_{\text{ReLU}}(x) = \max\{0, x\}, \tag{4.131}$$

which simply cuts off any negative part of x. In classification problems, the final hidden layer usually employs a different activation function, the so-called soft-max function, that is defined as

$$g_{\text{soft}-\max}(\mathbf{x}) = \frac{e^{x_i}\mathbf{e}_i}{\sum_{j=1}^{n} e^{x_j}}. \tag{4.132}$$

Clearly, $g_{\text{soft}-\max}$ operates on the data from all neurons of the layer at the same time, which differentiates it from the other activation functions. However, given that

$$\sum_i [g_{\text{soft}-\max}(\mathbf{x})]_i = \sum_i \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} = 1, \tag{4.133}$$

we guarantee the property that for layer N denoting the output layer $|\mathbf{x}^N|_1 = 1$, i.e., the components of the output vector can indeed be interpreted as probabilities.

Now, we have seen that neural networks provide a convenient way to construct extremely complex models, which makes them a powerful tool for many industrial applications, where they provide a solution to problems that involve complicated and usually unknown relationships between input and output vectors. Once the parameters of the model are optimized the gained knowledge about these relationships is encoded in the weights w_{ij}^l. With increasing complexity of the network this information becomes harder and harder to understand, leading to a conundrum. We have then simply mapped the complexity of one data set, the initial data (X, Y), to another data set W^l that is not necessarily easier to understand. The latter becomes obvious if we try to write down an explicit expression for the model. Even with a single hidden layer we already have

$$\mathbf{x}^2 = \sum_{i=0}^{n} \sum_{j=0}^{n_1} \sum_{k=1}^{m} g(w_{ij} x_i w_{jk}) \mathbf{e}_k, \tag{4.134}$$

and for two hidden layers

$$\mathbf{x}^3 = \sum_{i=0}^{n} \sum_{j=0}^{n_1} \sum_{k=0}^{n_2} \sum_{l=1}^{m} g\left(w_{kl} g(w_{ij} x_i w_{jk})\right) \mathbf{e}_l. \tag{4.135}$$

It is clear that, accounting for the complicated form of the activation function g, the general expression Eq. 4.135 is not very intuitive. Assuming we are faced with a problem where the intent is not simply to be able to compute Y from X, but instead we are interested in the specific relationship between X and Y, such a complicated model is quickly becoming intractable. This issue that is common to very complicated models is referred to as a lack of "interpretability" and will be discussed further in Chapter 7.

Chapter 5

Topology + Non-local Correlations

In this chapter we summarize a project where we studied the nature of non-local correlations with applications to topological systems. The aim is to study correlation effects and how they interact with the theory of topological classification of insulating phases while abstracting out any unnecessary details of the specific models. To this end we chose the ionic Hubbard model on the square lattice as a test bed that allows us to draw conclusions for a large variety of different two-dimensional lattices, since correlations on the square lattice are especially strong.

In the following we explain the motivation for our work and then introduce the main quantity of interest, the self-energy dispersion amplitude, in Sec. 5.2. We briefly discuss selected important properties of the ionic Hubbard model in Sec. 5.3 before then descending into our detailed analysis of the self-energy dispersion amplitude in this model in Sec. 5.4. We focus primarily on the square lattice, however, in the end we also show results for the hexagonal lattice for comparison.

Parts of the results discussed in this chapter were published as Ref. [135]:

Thomas Mertz, Karim Zantout and Roser Valentí
Self-Energy Dispersion in the Hubbard Model
Phys. Rev. B **98**, 235105 (2018)

TPSC calculations were performed by Karim Zantout.

5.1 Motivation

Since the publication of the seminal 2012 paper by Wang and Zhang, where the concept of the topological Hamiltonian was introduced [64], a lot of research has been conducted, applying this methodology to a variety of models to study the effect of electronic interactions on the topological phase diagrams. We remember that the topological Hamiltonian (for more details see Sec. 2.5.2) is defined as

$$h_t(\mathbf{k}) = H_0(\mathbf{k}) + \Sigma(\omega = 0, \mathbf{k}), \tag{5.1}$$

where $H_0(\mathbf{k})$ is the non-interacting Bloch Hamiltonian and $\Sigma(\omega, \mathbf{k})$ the single particle self-energy defined via the Dyson equation (Eq. 3.134).

For many years the dynamical mean-field theory (DMFT) has been the major workhorse in condensed matter physics, fueled by its success in describing the Mott-Hubbard transition in the

Hubbard model [124,136]. Naturally, also many investigations into correlated topological phases made use of DMFT [137–142]. However, as shown in Sec. 3.5, the DMFT self-energy is local by construction and as a consequence the topological Hamiltonian in the DMFT approximation reads

$$h_t(\mathbf{k}) = H_0(\mathbf{k}) + \Sigma_{\mathrm{DMFT}}(\omega = 0). \tag{5.2}$$

As we can see, the self-energy is reduced to a local quantity and therefore the momentum-dependence results entirely from the non-interacting part $H_0(\mathbf{k})$. Topological invariants for the usual quantum Hall effects are, however, measures that describe, e.g., the winding of eigenstates of $H(\mathbf{k})$ in momentum space for the case of the Chern number which brings into question the ansatz Eq. 5.2, where part of this important piece of information (the momentum dependence of the self-energy) is neglected.

We could, in principle, choose a particular topological model, perform calculations with DMFT and another method that retains a momentum dependence and by comparison of the two results judge the quality of the DMFT approximation in this context. This, however, is very specific to a particular model. Hence, we choose to first develop an appropriate abstraction that allows us to make observations that apply to a broader range of models.

The starting point must of course be the generic Hubbard model that describes electrons subject to the hopping amplitudes t_{ij} and Hubbard interaction U through

$$H = \sum_{ij} t_{ij} c_i^\dagger c_j + U \sum_i c_{i\uparrow}^\dagger c_{i\uparrow} c_{i\downarrow}^\dagger c_{i\downarrow}. \tag{5.3}$$

Some of the most influential models that can be found in the literature are the Hofstadter [21], Haldane [143], Kane-Mele [63] and Bernevig-Hughes-Zhang [144] models. Albeit being defined on different lattices, all of these models have in common a local potential term. This potential divides the lattice into A and B sublattices that are connected to each other via the hopping terms. The generic Hubbard model containing this potential is called the *ionic Hubbard model*

$$H = \sum_{i \neq j} t_{ij} c_i^\dagger c_j + \Delta \sum_i \mathrm{sgn}(i) c_i^\dagger c_i + U \sum_i c_{i\uparrow}^\dagger c_{i\uparrow} c_{i\downarrow}^\dagger c_{i\downarrow}, \tag{5.4}$$

where $\mathrm{sgn}(i) = \pm 1$ if i belongs to the A/B sublattice. The term *ionic* indicates a possible realization of such a model, where one imagines a crystal composed of ions, where a free electron is attracted to positively charged ions (corresponding to negative Coulomb energy) while being repelled from negatively charged ions (positive Coulomb energy). In reality, a crystal composed of atoms of different chemical elements always realizes an ionic Hubbard model. To see this let us imagine a crystal containing two atom sorts located on A and B sublattices. Since the Coulomb energies E_A, E_B differ we can write

$$\sum_{i\in A} E_A c_i^\dagger c_i + \sum_{i\in B} E_B c_i^\dagger c_i = \sum_{i\in A} \frac{E_A - E_B}{2} c_i^\dagger c_i + \sum_{i\in B} \frac{E_B - E_A}{2} c_i^\dagger c_i + \frac{E_A + E_B}{2} \sum_i c_i^\dagger c_i \tag{5.5}$$

$$=: \Delta \left[\sum_{i\in A} c_i^\dagger c_i - \sum_{i\in B} c_i^\dagger c_i \right] + \frac{E_A + E_B}{2} \sum_i c_i^\dagger c_i. \tag{5.6}$$

The first term is apparently just the sought-after ionic potential, where we arbitrarily defined $\Delta = \frac{E_A - E_B}{2}$ as half the energetic distance between the two sites. Since it is up to us to define which sublattices the labels A and B correspond to the sign of Δ is not well-defined. Indeed, we can immediately see that changing $\Delta \to -\Delta$ effectively interchanges A and B and therefore we only shift the lattice. As a consequence the physics don't change and it is enough to restrict the discussion to positive Δ for convenience. The second term is proportional to the total density

and therefore just a constant shift in the energy scale, which can be neglected by redefining the location of the zero of energy. Hence, the physical properties of H transform as the identity under

$$H \to H - \frac{E_A + E_B}{2}, \tag{5.7}$$

so that we can use only the first term of Eq. 5.6. The generic model now serves as a template for all the previously mentioned topological models, where we make use of the fact that the square lattice features very strong non-local correlations and can therefore be considered as an upper limit [145].

5.2 Self-Energy Dispersion Amplitude

In order to judge the reasonableness of a topological Hamiltonian in the form of Eq. 5.2 with a constant self-energy we first try to answer how accurate the assumption that DMFT makes, i.e., the locality of the self-energy, is. A priori, it is entirely possible that there is a broad class of models where the self-energy is, in fact, local and therefore DMFT would provide a very good description and consequently the momentum-dependence of the topological Hamiltonian would be governed entirely by the non-interacting Bloch Hamiltonian in these cases.

We propose now a systematic way of describing the role of the momentum-dependent corrections on top of DMFT through the definition of the *self-energy dispersion amplitude*

$$d_a(\omega) = \max_{k,k'} \parallel \Sigma(\omega, k) - \Sigma(\omega, k') \parallel_\infty, \tag{5.8}$$

where $\parallel \cdot \parallel_\infty$ denotes the matrix norm with respect to orbital/site degrees of freedom $\parallel A \parallel_\infty = \max_{ij} |A_{ij}|$. We illustrate the dispersion amplitude in Fig. 5.1, by plotting an arbitrary function f that represents a matrix element $\Sigma_{ab}(\omega, k)$ of the self-energy at a particular frequency ω. Clearly, f can be visualized as a surface that is parameterized by momentum $\mathbf{k} = (k_x, k_y)$ and which has a particular thickness. This thickness is what we define as the "amount" of momentum dependence. This definition makes sense in a way that if we imagine a completely

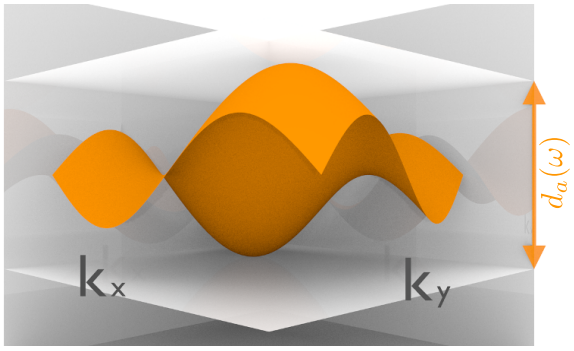

Figure 5.1: Illustration of the self-energy dispersion amplitude. A single matrix element $f(k) = |\Sigma_{ab}(\omega, k)|$ is plotted schematically for an arbitrary frequency ω as a function over the first Brillouin zone. The self-energy dispersion amplitude is just the thickness of the surface defined by f. Due to the maximum norm the value of d_a for matrix-valued self-energies is given by maximal thickness among all functions describing the individual matrix elements. [Figure from Ref. [135]]

flat function $f(k) = \text{const.}$ the dispersion amplitude would naturally vanish. This is expected, since a constant function does not depend on k and therefore has no dispersion. On the contrary, the further any two values $f(k), f(k')$ lie apart the more dispersive f appears to be according to Eq. 5.8.

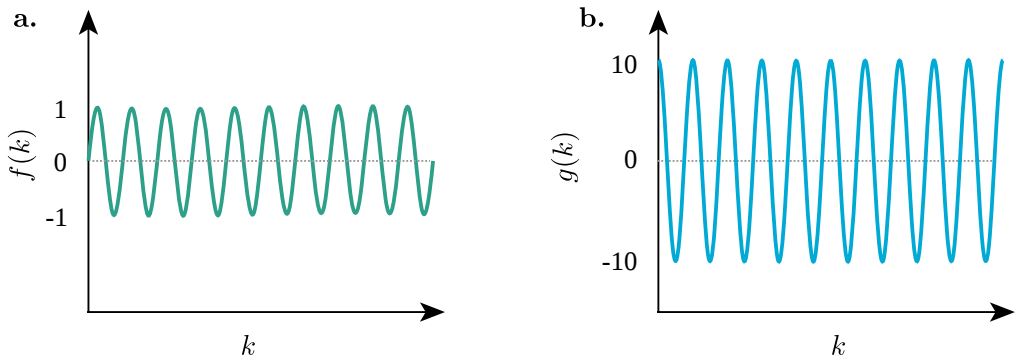

Figure 5.2: Dispersion of the function **a.** $f(k) = \sin(10k)$, $k \in [0, 2\pi]$ and **b.** $g(k) = f'(k) = 10\cos(10k)$. The thickness of f is apparently 2, however, its derivative shows a much stronger momentum-dependence, ranging from -10 to 10. This indicates that the slope of a function is not a suitable measure of the strength of its momentum dependence. In this case the difference is one order of magnitude.

We note that, while appearing rather trivial, the definition of d_a in Eq. 5.8 is a rather unique choice for a dispersion measure. It appears for example to completely neglect the slope of f, which in many cases shapes our intuition about how a function changes. However, large $d_a(\omega)$ necessarily implies that $\exists k \in 1.\,\mathrm{BZ}$ with $|f'(k)| = \frac{d_a(\omega)}{2\pi}$ by virtue of the mean value theorem. Trying to formulate a measure based on f' on the other hand would likely fail, since one can construct functions with $f' \gg 1$ while $d_a(\omega) \ll 1$, which is a consequence of the fact that the statement of the mean value theorem cannot be reversed. One such example would be $f(k) = \epsilon \sin(mk)$ with $\epsilon \ll 1$ and $m\epsilon \gg 1$. Since $f'(k) = \epsilon m \cos(mk)$, a measure based on f' alone would indicate a strong dispersion although f is essentially constant. Our definition of d_a on the other hand is consistent with the intuition that small variations are negligible. This is illustrated in Fig. 5.2, where we use the example $m = 10, \epsilon = 1$, i.e., $f(k) = \sin(10k)$.

Given that d_a is an absolute measure lacking a clear reference as to what values constitute a large dispersion we define also the relative dispersion amplitude

$$d_r(\omega) = \begin{cases} \frac{d_a(\omega)}{N_k^{-1} \| \sum_k \Sigma(\omega, k) \|} & \text{if } \sum_k \Sigma(\omega, k) \neq 0 \\ d_a(\omega) & \text{else.} \end{cases} \tag{5.9}$$

d_r, in contrast to d_a defines an inherent unit in the sense that $d_r = 1$ surely indicates a strong momentum dependence, while $d_r = 0.01$ means the self-energy is rather dispersion-less. We illustrate in Fig. 5.3 the difference between d_r and d_a by plotting two functions f and g, which bear the same momentum dependence, but are shifted along the y axis with respect to each other, i.e., $f(k) = g(k) + \text{const.}$. In Fig. 5.3**a**, to the eye both functions appear to have the same amount of dispersion and indeed we have $d_a = 0.1$ for both. We then change the scale of the axis such that g appears at the same distance to 0 as f, which is shown in Fig. 5.3**b**. Apparently, the same function appears much more dispersive and we could argue that the momentum dependence is actually more important for g than it is for f, since the ratio between the strength of the momentum-dependent and constant contributions is much larger.

This simple example gives rise to the following interpretation of d_a and d_r: the dispersion amplitude d_a is a measure of the amount of momentum dependence, while d_r measures the importance of the momentum dependence and is scale-independent.

By construction, d_r measures the dispersion of Σ relative to its mean. Thus, if Σ is very large only variations on the same scale would be considered important. This reasoning makes sense because a small relative dispersion amplitude indicates that the main contribution of Σ to the topological Hamiltonian would be its average value, while variations are comparatively

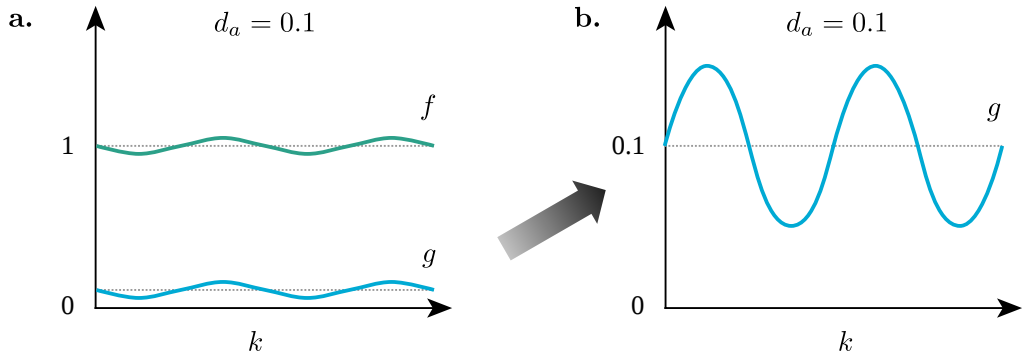

Figure 5.3: Illustration of the relative self-energy dispersion amplitude. We assume a dispersion amplitude $d_a = 0.1$ in arbitrary energy units. **a.** Two functions that have the same dependence on k, but different means appear to have the same momentum-dependence. This is reflected in the equal values of $d_a = 0.1$ in both cases. **b.** Zooming in on the smaller function g such that its mean is at the same level as that of f leads to a calibration of scales between the two functions. g now appears to be more strongly dispersive than f, since the deviations from its mean are much larger at this scale.

small. On the other hand, $d_r \sim 1$ means that the momentum dependence cannot be neglected and is expected to impact, e.g., the calculation of topological invariants.

We proceed by showing quite formally that the error of the DMFT approximation can be expressed through the self-energy dispersion amplitude as

$$\varepsilon(\omega) = \| \Sigma_{\text{exact}}(\omega, k) - \Sigma_{\text{DMFT}}(\omega) \| \leq d_a(\omega) + r(\omega), \tag{5.10}$$

where $r(\omega) \geq 0$ is the error of the local self-energy and is related to $d_a(\omega)$.

Proof. We start by expanding the exact self-energy $\Sigma_{\text{exact}}(\omega, k)$ around Σ_{DMFT}. The particular form of the expansion is irrelevant at this point and it suffices to note that one can write

$$\Sigma_{\text{exact}}(\omega, k) = \Sigma_{\text{DMFT}}(\omega) + S(\omega, k), \tag{5.11}$$

where S contains the sum of all Feynman diagrams that are not accounted for by DMFT, i.e., all non-local diagrams. We divide S into two parts S_0 and S_1, where $\sum_k S_0 = 0$. Therefore,

$$\Sigma_{\text{exact}}(\omega, k) = \Sigma_{\text{DMFT}}(\omega) + S_0(\omega, k) + S_1(\omega, k), \tag{5.12}$$

and the local self-energy is

$$\frac{1}{N_k} \sum_k \Sigma_{\text{exact}}(\omega, k) = \Sigma_{\text{DMFT}}(\omega) + \frac{1}{N_k} \sum_k S_1(\omega, k). \tag{5.13}$$

Apparently, S_1 contains an additional correction of the local self-energy on top of the DMFT result that is a consequence of contributions of non-local diagrams. By defining $R(\omega) = \frac{1}{N_k} \sum_k S_1(\omega, k)$ we can rewrite S_1 as $S_1(\omega, k) = R(\omega) + [S_1(\omega, k) - R(\omega)]$, where the term in brackets has a vanishing momentum average by construction and can therefore be accumulated in $S_0(\omega, k)$. In total, we have

$$\Sigma_{\text{exact}}(\omega, k) = \Sigma_{\text{DMFT}}(\omega) + S_0'(\omega, k) + R(\omega), \tag{5.14}$$

with $S_0'(\omega, k) = S_0(\omega, k) + S_1(\omega, k) - R(\omega)$. For the DMFT error we then obtain

$$\| \Sigma_{\text{exact}}(\omega, k) - \Sigma_{\text{DMFT}}(\omega) \| = \| S_0'(\omega, k) + R(\omega) \| \tag{5.15}$$

$$\leq \| S_0'(\omega, k) \| + \| R(\omega) \| . \tag{5.16}$$

Since $S_0'(\omega, k)$ contains the entire momentum dependence, it is clear that $S_0'(\omega, k) \leq d_a(\omega)$ and with the definition $r(\omega) = \| R(\omega) \|$ we obtain the desired decomposition of the DMFT error that is given in Eq. 5.10.

What is left to show is that $r(\omega)$ is related to $d_a(\omega)$. By definition $r(\omega) = \| R(\omega) \| \geq 0$ and $R(\omega)$ is defined as the k-average over S_1, which is the sum over all diagrams with non-vanishing k-average. Assuming that $d_a = 0$, S_1 could only contain constant terms that do not depend on k. By construction, these contributions are, however, contained in the DMFT self-energy and therefore necessarily $R(\omega) = 0$. □

Apparently, the choice of d_a as a measure of the amount of momentum dependence conveniently lends itself as an upper bound for the DMFT error, which makes it a very powerful tool for the discussion of the applicability of the local approximation. The residual term $r(\omega)$ can take finite values only if $d_a(\omega) > 0$, however, in general we cannot state how large $r(\omega)$ can be for given $d_a(\omega) > 0$. As a rule of thumb we can argue, though, that small d_a indicates a rather local system and therefore the DMFT solution is expected to be a good approximation. Hence, $r(\omega)$ is expected to be small for small $d_a(\omega)$. On the contrary, if d_a is large, r may also be large, however, since we are not so much interested in a strict upper bound rather than an indicator of the quality of the DMFT, d_a alone is sufficient for our purposes.

5.3 Ionic Hubbard Model

In the following we will discuss the physics of the ionic Hubbard model with a focus on the DMFT error based on our description in terms of the self-energy dispersion amplitude. The ionic Hubbard model provides an interesting testbed for this discussion, since the ionic potential Δ provides a proverbial knob to tune the localization of the system as we will see. For this discussion we use a variant of the two-dimensional ionic Hubbard model. While traditionally the sublattices A, B are defined such that the potential is staggered along the two axes [146], we here use a model where the sublattice index refers to an x-coordinate only, i.e., we have a striped lattice and the potential is staggered only along the x-axis as it is illustrated in Fig. 5.4. The localization in this model proves to be much less severe, which makes the discussion more interesting. The one-dimensional formulation of the model has already been investigated extensively [147–151], albeit not in the present context. Investigations into the charge density-wave phases, magnetic order and metal-insulator transition in the two-dimensional model can be found, e.g., in Refs. [146, 152–158].
The Hamiltonian we use in the following is given by

$$H = -t \sum_{\langle i,j \rangle} c_i^\dagger c_j - \Delta \sum_{i \in A} n_i + \Delta \sum_{i \in B} n_i + U \sum_i n_{i\uparrow} n_{i\downarrow}, \tag{5.17}$$

where $c_i^{(\dagger)}$ are the fermionic annihilation (creation) operators on site i and an additional spin index is implied. $n_{i\sigma} = c_{i\sigma}^\dagger c_{i\sigma}$ is the spin density operator on site i. In the following we always have $\Delta \geq 0$. It then follows that the energy on the A sublattice is $-\Delta$ and $+\Delta$ on the B sublattice. An electron thus reduces its energy by occupying the A sublattice, so that the ground state will naturally feature a higher occupation of the A sites. In particular, if we take the simple case of non-interacting electrons and $t = 0$, the A sublattice will be filled first and only the remaining electrons that do not fit occupy the B sites. The density is then given by $n_A = \min\{2n_f, 2\}$ and $n_B = 2n_f - n_A$, where n_f is the filling, i.e., the number of electrons per site. At finite temperature we have $n_{A/B} = 2f(\mp\Delta)$ (factor 2 because of the two spins), where f is the Fermi-Dirac distribution function. The chemical potential μ is then adjusted such that

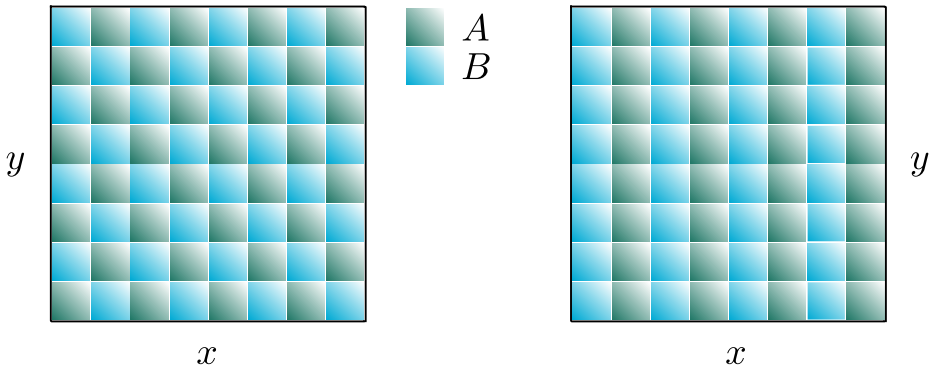

Figure 5.4: Illustration of the ionic potential. The A, B sublattices are shown in different colors. Left: traditional lattice with staggered potential in both directions. Right: striped lattice with staggered potential only in x-direction. Generally, the ionic potential leads to a localization of electrons. This is more severe in the case on the left, since the potential acts in both directions.

the correct filling is obtained

$$2n_f = n_A + n_B = 2[f(-\Delta) + f(\Delta)] = \frac{2}{e^{-\beta(\Delta+\mu)}+1} + \frac{2}{e^{\beta(\Delta-\mu)}+1}. \tag{5.18}$$

While an exact solution is not available, we plot the numerical solution in Fig. 5.5. At low temperatures we obtain the expected behavior, where the A sites are occupied first as soon as $\mu \geq -\Delta$. B sites are only occupied at $\mu \geq +\Delta$. At finite temperatures on the other hand the situation changes, which is why we are doing this exercise. The densities are severely broadened by the Fermi statistics and B sites are occupied much earlier.

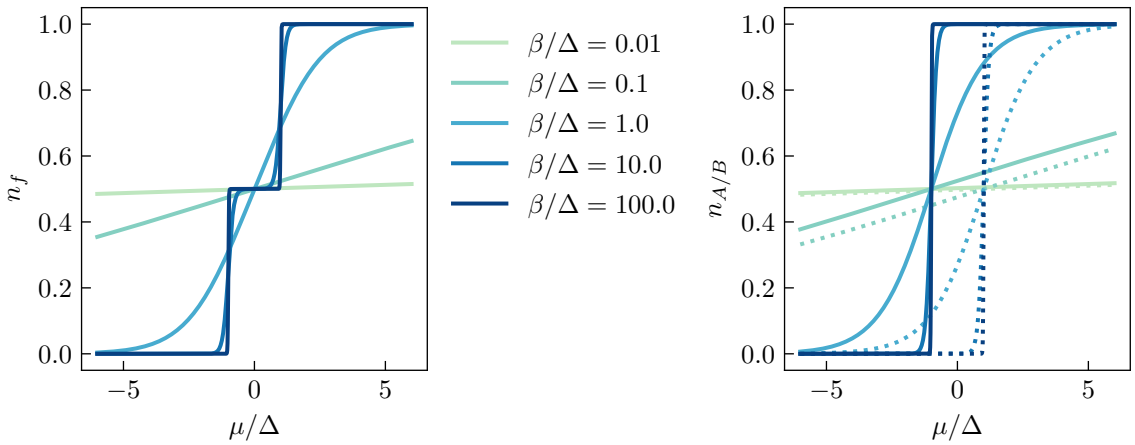

Figure 5.5: Density of the trivial ionic model at $t = U = 0$. Left: filling as a function of chemical potential, right: sublattice densities $n_{A/B}$ (n_B is dotted). Both are shown only for a single spin. At low temperatures, for $\mu \geq -\Delta$ the A sublattice is fully occupied while occupation of the B sublattice happens only at $\mu \geq \Delta$. At higher temperatures this is severely broadened.

We now add a kinetic energy to the system in terms of finite hopping amplitudes, i.e., $t > 0$, which adds dynamics to the system, since electrons can now move about the lattice. The energies are given by the eigenvalues of

$$H_0(\mathbf{k}) = \begin{pmatrix} -2t\cos(k_y) - \Delta & -t(1+e^{-ik_x}) \\ -t(1+e^{ik_x}) & -2t\cos(k_y) + \Delta \end{pmatrix}, \tag{5.19}$$

which can be expressed as

$$\varepsilon_{1,2}(\mathbf{k}) = -2t\cos(k_y) \pm \sqrt{2t^2\,(1+\cos(k_x)) + \Delta^2}. \tag{5.20}$$

The densities are now computed as $n_f = \sum_{\mathbf{k},n} f(\varepsilon_n(\mathbf{k}))$ and we obtain the results shown in Fig. 5.6 for $t/\Delta = 1$. Apparently, also t broadens the density and allows for occupations of the B sublattice at rather low energies. In fact, even at very low temperature $T = \beta^{-1}$ the occupation n_B is finite for the same values of μ as n_A, only the slope is smaller. Therefore the region from the trivial model where only the A sites were occupied does no longer exist. However, there exists a region, where the slope of n_B is significantly smaller than that of n_A. From the dispersion relation of Eq. 5.20 it follows that the lowest energy in the lower band is given by $\varepsilon_{1,\min}/\Delta = -2-\sqrt{5}$, while for the upper band we obtain $\varepsilon_{2,\min}/\Delta = -1$. These are exactly the lower and upper bounds of the region where n_A dominates, cf. the shaded region in Fig. 5.6. Therefore, occupying the states from the lower band increases predominantly n_A, and vice versa, which indicates that the site character of the states in the lower (upper) band is dominated by the A (B) sublattices.

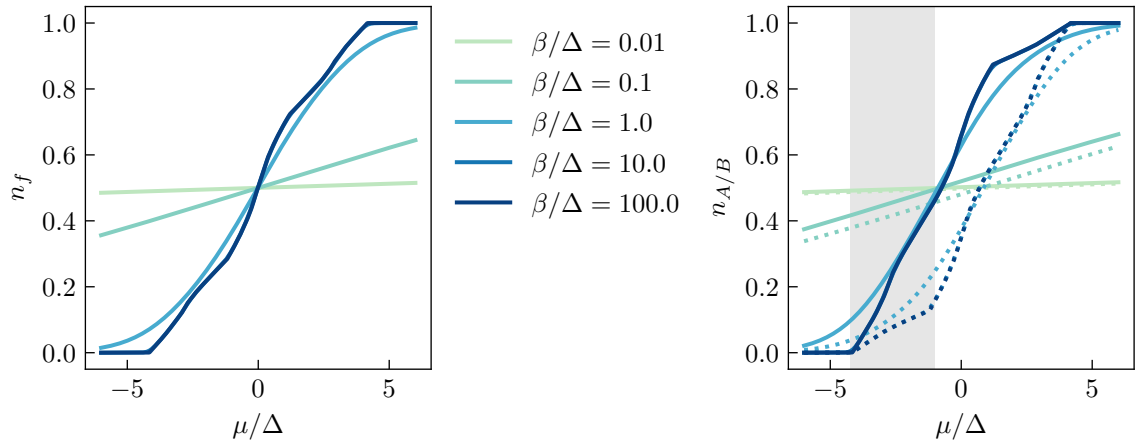

Figure 5.6: Density of the non-interacting ionic model at $t/\Delta = 1$ and $U = 0$. Left: filling n_f, right: sublattice densities $n_{A/B}$ (n_B is dotted) as a function of chemical potential μ. Both are shown only for a single spin. The region for $\mu/\Delta \in [-2-\sqrt{5}, -1]$, where only states from the lower band are occupied, is shaded gray. The much smaller sloper of n_B in this region indicates that the sublattice character of the lower band is predominantly of A type.

For the limits $t \to 0$ and $\Delta \to 0$ we obtain the static and homogeneous system, respectively. The static system has an energy gap of 2Δ, while in the $t > 0$ case we have to compare the largest energy of the lower band $\varepsilon_{1,\max} = 2t - \Delta$ and the smallest energy of the upper band $\varepsilon_{2,\min} = -2t + \Delta$ for general t/Δ. For $\Delta > 2t$ the band gap opens and it can be verified that also the site character of the states in the two bands becomes more and more distinct, i.e., electrons are localized to a particular sublattice.

So far we have seen that both temperature and nearest neighbor hopping lead to a more evenly distributed density and counteract the effect of the staggered ionic potential. We now add the repulsive onsite interaction U to the mix. First, we illustrate in Fig. 5.7 our expectations. Large amplitudes of the ionic potential lead to energetically favorable double occupation of A sites as we have seen in the non-interacting limit. The repulsive onsite interaction adds an energy penalty of U per doubly occupied site, i.e., depending on the ratio U/Δ, local interactions between electrons can lead to a more evenly distributed density. Regarding the locality of the self-energy, which we are primarily interested in, it is clear that while the non-interacting system is extremely localized for large Δ the repulsive interaction acts in the exact opposite direction.

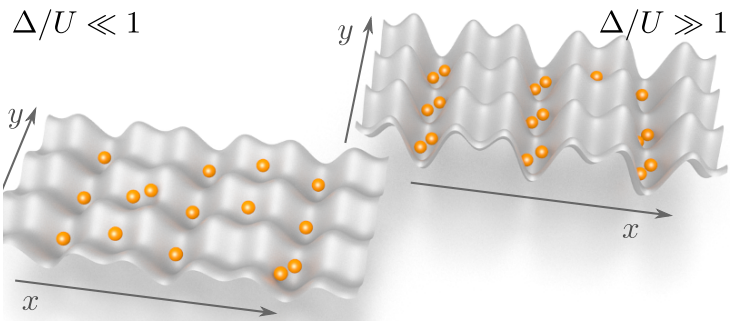

Figure 5.7: Illustration of the striped ionic lattice potential for two cases $\Delta/U \ll 1$, where the difference between the onsite energies of the A/B sublattices is rather small and $\Delta/U \ll 1$, where this difference is large. For $\Delta/U \ll 1$, we expect electrons to be spread out more or less evenly, since the repulsive interaction pushes electrons onto to higher-energetic B sites. For $\Delta/U \gg 1$, the deep potential wells force electrons to doubly occupy the A sites despite the energetic penalty due to the interaction. [Figure adapted from Ref. [135]]

At small U, far away from the Mott transition, this could lead to delocalization and therefore we expect also non-local contributions to the self-energy. The competition between Δ and U is for example revealed in the delayed antiferromagnetic transition [154] and signs of the delocalization as a result of electronic interactions have been reported in terms of an interaction-driven metallic phase [146].

For the following discussion we choose half filling, which is the most interesting case, since in the atomic limit only the A site would be occupied. At the same time the filling is large enough to force electrons into the B sublattice as soon as a kinetic energy is added. At $T = 0$ the sublattice densities of the non-interacting model are far apart at $n_{A\sigma} \approx 0.75$ and $n_{B\sigma} \approx 0.25$, cf. Fig. 5.6. We perform exact diagonalization calculations on finite clusters to obtain the ground states of the interacting system for comparison with the non-interacting data.

Exact diagonalization (ED) requires the choice of a finite size cluster, which can be expected to influence the results. Here, we use 4, 8 and 12 site clusters with periodic boundary conditions, which are shown in Fig. 5.8. The possible cluster sizes are somewhat limited by the periodicity requirement at the boundary, since this implies that only integer multiples of the size of the unit cell (here 2) are possible. We neglect also intermediate sizes 6 and 10 and end up with only integer multiples of 4.

For four sites there are in principle two choices, however, we neglect the clusters that extend only in one spatial dimension. For 8 sites there are then two choices (C8h, C8v), where "h", "v" denote horizontal and vertical alignment of the long axis, respectively. For clusters of size 12 there are three choices (C12h, C12v, C12m), where "m" denotes the middle, where neither extent is particularly large compared to the other. The next possible cluster size would be 16 sites for which the calculations become very costly due to the exponential scaling of the Hilbert space dimension $\dim(\mathcal{H}) \sim 4^L$. Since we work at fixed filling the actual dimension is lower than 4^L. In fact, we have

$$\dim(\mathcal{H}) = \binom{L}{\frac{L}{2}}\binom{L}{\frac{L}{2}} \sim \frac{4^L}{L}. \tag{5.21}$$

For $L = 16$ we would have $\dim(\mathcal{H}) \sim 2^{28} \approx 2.6 \times 10^8$. The ground state stored as a vector of complex double precision floating point numbers would require $\dim(\mathcal{H}) \times 16\text{Bytes} \approx 4\text{GB}$. The Hamiltonian matrix, however, even in sparse form would be much too large to store. Such calculations are only possible if the Hilbert space dimension is reduced through complicated symmetry-based mappings between states, or if the required matrix-vector products involving the Hamiltonian are computed on the fly without ever storing the Hamiltonian. Both of these

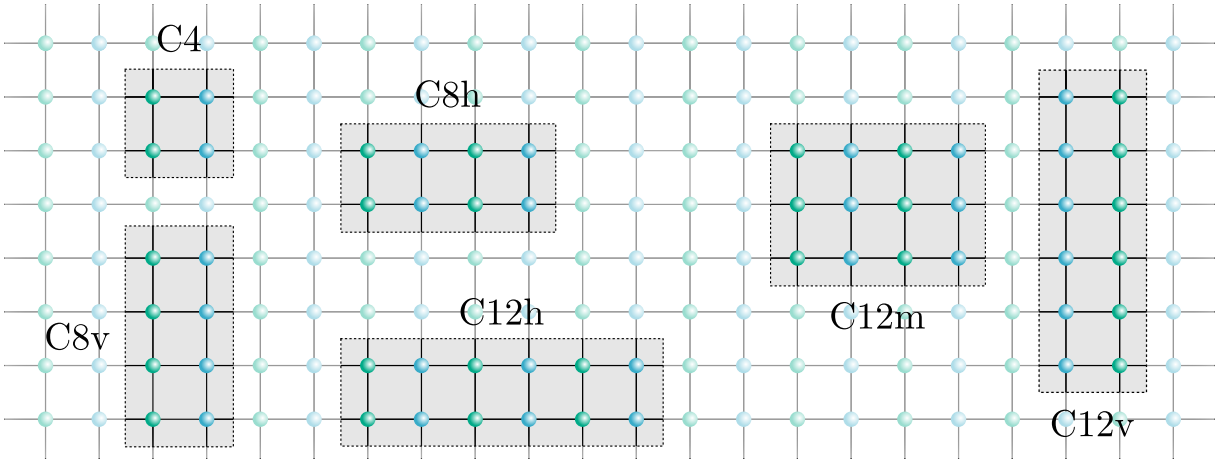

Figure 5.8: Clusters used in our ED calculations for the ionic Hubbard model. Neglecting one-dimensional clusters that extend only in one direction we restrict ourselves to cluster sizes that are integer multiples of 4. We performed calculations for all such clusters with sizes between 4 and 12 sites and define here the following naming scheme. Except for C4 all clusters have a long axis, which is indicated by Csx, where s is the cluster size and x=v for vertical orientations, i.e., along the stripes, or x=h for horizontal orientations, i.e., perpendicular to the stripes. All clusters use periodic boundary conditions.

options were not implemented for this project.

As we have already established, the ionic model features density waves, which are characterized by a nonzero staggered density

$$n_s := n_A - n_B. \tag{5.22}$$

Clearly, $n_s \geq 0$ for $\Delta \geq 0$, since the A sublattice will always have higher occupation numbers in the ground state. We plot the staggered density as a function of the onsite interaction U at $\Delta/t \approx 0.8$ for all clusters in Fig. 5.9. We note that for technical reasons we had used a slightly different Δ-grid for the smaller clusters, since a comparison at fixed Δ as in this plot was not originally planned. Performing all-new calculations just for this plot seemed a bit wasteful in terms of energy consumed by the HPC cluster.

In the non-interacting calculation we obtain $n_s \approx 0.6$ for $\Delta/t = 0.8$, which is rather close to some of the results of the finite size ED calculation. In general, however, the results obtained with different clusters strongly disagree. The two horizontally oriented clusters C8h and C12h predict much lower n_s at $U = 0$ and the same happens for the intermediate C12m cluster at larger U. We can understand this observation by considering that the dynamics happen primarily within stripes belonging to the same sublattice, while any process along the x-axis is literally fighting an uphill battle against the steep potential gradient. The horizontally oriented clusters restrict the size L_y and with that the dynamics much more and therefore finite size effects become more pronounced. Increasing the cluster size in x-direction is much less effective, since there electrons are rather localized and we only generate copies of the same physics. In general, best results are usually obtained by averaging over different clusters.

Clearly, all calculations predict that n_s decreases as a function of U, i.e., electron-electron interactions counteract the effect of the ionic potential and dampen the density waves as expected. Since we performed the calculations without a magnetic field, the Hamiltonian has SU(2) symmetry. Consequently, the same applies to the ground state, which in case of no degeneracy cannot spontaneously break symmetries of the Hamiltonian. In analogy to the non-striped version of the model we also expect an antiferromagnetic transition at finite U [154].

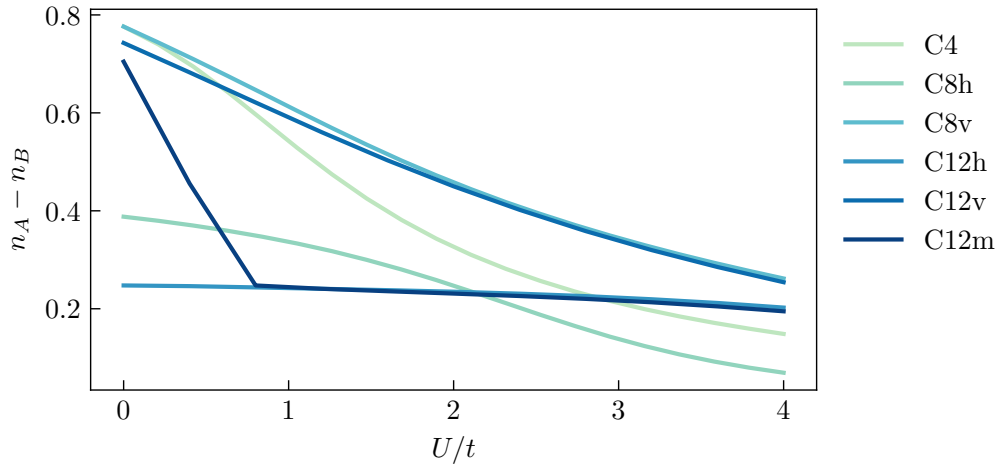

Figure 5.9: Staggered density $n_s = n_A - n_B$ as a function of U/t for $\Delta/t = \frac{16}{19} \approx 0.84$ for the 4- and 8-site clusters and $\Delta/t = 0.8$ for the 12-site clusters. Generally, the staggered density decreases monotonously as a function of U/t, since the repulsive interaction forces electrons to occupy also B sites. The result shows a clear dependence on the choice of the cluster with only C8v and C12v agreeing rather well. The clusters C8h and C12h perform significantly worse than the others, e.g., at $U = 0$, where the exact result (infinite system) is given by $n_s \approx 0.6$.

5.4 Self-Energy Dispersion in the Ionic Hubbard Model

We now turn to the main subject of interest, namely the momentum dependence of the self-energy in the ionic Hubbard model. We can immediately write down two limiting cases: i) $U \to \infty$ with $U/\Delta \gg 1$ and ii) $\Delta \to \infty$. For i) the system features magnetic order, which is of the antiferromagnetic kind due to the underlying square lattice. Independently of the magnetic order the double occupancy $\langle n_{i\uparrow} n_{i\downarrow} \rangle$ on both sublattices is suppressed by the strong repulsive interaction. The single particle Green's function and self-energy can be computed analytically in both cases. This is done in the following section.

5.4.1 Exact Limits

The limit $U \to \infty$ is equivalent to letting $t \to 0$. Here, we are interested, in particular, in the limit $U \gg \Delta \gg t$. Therefore,

$$H_{i)} = -\Delta \sum_{i \in A} n_i + \Delta \sum_{i \in B} n_i + U \sum_i n_{i\uparrow} n_{i\downarrow}, \tag{5.23}$$

which apparently is diagonal in the Wannier basis. The eigenstates therefore reduce to antisymmetrized products of local states. Focusing on a single unit cell we have states

$$\begin{aligned} |\psi_1\rangle &= c^\dagger_{A\uparrow} c^\dagger_{A\downarrow} |\text{vac}\rangle \\ |\psi_2\rangle &= c^\dagger_{A\uparrow} c^\dagger_{B\downarrow} |\text{vac}\rangle \\ |\psi_3\rangle &= c^\dagger_{B\uparrow} c^\dagger_{A\downarrow} |\text{vac}\rangle \\ |\psi_4\rangle &= c^\dagger_{B\uparrow} c^\dagger_{B\downarrow} |\text{vac}\rangle, \end{aligned} \tag{5.24}$$

with corresponding energies

$$\begin{aligned} E_1 &= U - 2\Delta \\ E_2 &= 0 \\ E_3 &= 0 \\ E_4 &= U + 2\Delta. \end{aligned} \tag{5.25}$$

The ground state $|\mathrm{gs}\rangle = \frac{1}{\sqrt{|c_2|^2+|c_3|^2}}(c_2|\psi_2\rangle + c_3|\psi_3\rangle)$ is degenerate and we choose arbitrarily $c_2 = 0$. According to the Lehman representation, cf. Eq. 3.128, the Green's function is then given by

$$G_{AA,\uparrow\uparrow}(\omega) = -i\sum_m \frac{|\langle m|c^\dagger_{A\uparrow}|\mathrm{gs}\rangle|^2}{\omega - E^{N+1}_m + E^N_0 + i\eta} + \frac{|\langle m|c_{A\uparrow}|\mathrm{gs}\rangle|^2}{\omega + E^{N-1}_m - E^N_0 + i\eta}, \tag{5.26}$$

where we have for states with an additional electron

$$\begin{aligned} |\psi_1\rangle^{N+1} &= c^\dagger_{A\uparrow}c^\dagger_{A\downarrow}c^\dagger_{B\downarrow}|\mathrm{vac}\rangle, \quad E^{N+1}_1 = U - \Delta \\ |\psi_2\rangle^{N+1} &= c^\dagger_{A\uparrow}c^\dagger_{B\uparrow}c^\dagger_{A\downarrow}|\mathrm{vac}\rangle, \quad E^{N+1}_2 = U - \Delta \\ |\psi_3\rangle^{N+1} &= c^\dagger_{A\uparrow}c^\dagger_{B\uparrow}c^\dagger_{B\downarrow}|\mathrm{vac}\rangle, \quad E^{N+1}_3 = U + \Delta \end{aligned} \tag{5.27}$$

and for those with one less

$$\begin{aligned} |\psi_1\rangle^{N+1} &= c^\dagger_{A\downarrow}|\mathrm{vac}\rangle, \quad E^{N-1}_1 = -\Delta \\ |\psi_2\rangle^{N+1} &= c^\dagger_{B\downarrow}|\mathrm{vac}\rangle, \quad E^{N-1}_2 = \Delta. \end{aligned} \tag{5.28}$$

Using this for the diagonal matrix element of the Green's function we arrive at

$$G_{AA,\uparrow\uparrow}(\omega) = -i\frac{1}{|c_2|^2+|c_3|^2}\left[\frac{|c_3|^2}{\omega + i\eta + \mu - (U-\Delta)} + \frac{|c_2|^2}{\omega + i\eta + \mu + \Delta}\right]. \tag{5.29}$$

At half filling the Hubbard model has particle-hole symmetry, i.e., $n \to (1-n)$ must be an identity operation. This implies that

$$\begin{aligned} Un_{i\uparrow}n_{i\downarrow} - \mu n_i &= U(1-n_{i\uparrow})(1-n_{i\downarrow}) - \mu(1-n_{i\uparrow}) - \mu(1-n_{i\downarrow}) && (5.30)\\ &= U - Un_i + Un_{i\uparrow}n_{i\downarrow} - 2\mu + \mu n_i && (5.31) \end{aligned}$$

$$\Rightarrow \qquad (2\mu - U)n_i + U - 2\mu = 0. \tag{5.32}$$

Since the first term involves an operator and the second does not, only one solution exists: $\mu = \frac{U}{2}$. We insert this into the Green's function from Eq. 5.29 and obtain

$$\begin{aligned} G_{AA,\uparrow\uparrow}(\omega) &= -i\frac{1}{|c_2|^2+|c_3|^2}\left[\frac{|c_3|^2}{\omega + i\eta + \frac{U}{2} - (U-\Delta)} + \frac{|c_2|^2}{\omega + i\eta + \frac{U}{2} + \Delta}\right] && (5.33)\\ &= -i\frac{1}{|c_2|^2+|c_3|^2}\left[\frac{|c_3|^2}{\omega + i\eta + \Delta - \frac{U}{2}} + \frac{|c_2|^2}{\omega + i\eta + \Delta + \frac{U}{2}}\right] && (5.34)\\ &= -i\frac{1}{|c_2|^2+|c_3|^2}\frac{(\omega + i\eta + \Delta)(|c_2|^2+|c_3|^2) + (|c_2|^2-|c_3|^2)\frac{U}{2}}{(\omega + i\eta + \Delta)^2 - \frac{U^2}{4}} && (5.35)\\ &= -i\left[\frac{\omega + i\eta + \Delta}{(\omega + i\eta + \Delta)^2 - \frac{U^2}{4}} + \frac{|c_2|^2-|c_3|^2}{|c_2|^2+|c_3|^2}\frac{\frac{U}{2}}{(\omega + i\eta + \Delta)^2 - \frac{U^2}{4}}\right]. && (5.36) \end{aligned}$$

In the S_z-symmetric ground state we then have

$$\begin{aligned}\langle \mathrm{gs}|S_{z,A/B}|\mathrm{gs}\rangle &= \frac{1}{|c_2|^2+|c_3|^2}(c_2^*\langle\psi_2| + c_3^*\langle\psi_3|)(\pm c_2|\psi_2\rangle \mp c_3|\psi_3\rangle) && (5.37)\\ &= \frac{1}{|c_2|^2+|c_3|^2}(\pm|c_2|^2 \mp |c_3|^2) && (5.38)\\ &\stackrel{!}{=} 0 && (5.39)\\ \Rightarrow \quad & |c_2| = |c_3|, && (5.40)\end{aligned}$$

and therefore

$$G_{AA,\uparrow\uparrow}(\omega) = -i\frac{\omega + i\eta + \Delta}{(\omega+i\eta+\Delta)^2 - \frac{U^2}{4}}. \tag{5.41}$$

The corresponding Matsubara Green's function is

$$G_{AA,\uparrow\uparrow}(i\omega) = \frac{i\omega + \Delta}{(i\omega+\Delta)^2 - \frac{U^2}{4}}. \tag{5.42}$$

In the same way we obtain

$$\begin{aligned}G_{BB,\uparrow\uparrow}(i\omega) &= \frac{i\omega - \Delta}{(i\omega-\Delta)^2 - \frac{U^2}{4}} && (5.43)\\ G_{AB,\uparrow\uparrow}(i\omega) &= G_{BA,\uparrow\uparrow}(i\omega) = 0. && (5.44)\end{aligned}$$

If we translate this single unit cell picture back into a lattice perspective we have

$$G_{AA}(i\omega, \mathbf{r}) = \delta_{\mathbf{r},0} G_{AA}(i\omega) \tag{5.45}$$

and the momentum space Green's function is

$$G_{AA}(i\omega, k) = \sum_i e^{-ik\cdot r_i} G_{AA}(i\omega, r_i) = G_{AA}(i\omega). \tag{5.46}$$

For the self-energy we therefore have

$$\Sigma_{\uparrow\uparrow}(\omega,k) = \begin{pmatrix} \frac{U^2}{4(i\omega+\Delta)} + \frac{U}{2} - 2\Delta + 2t\cos(k_y) & t(1+e^{-ik_x}) \\ t(1+e^{ik_x}) & \frac{U^2}{4(i\omega-\Delta)} + \frac{U}{2} + 2\Delta + 2t\cos(k_y) \end{pmatrix}, \tag{5.47}$$

where of course $t \ll U$ and therefore the leading term is just

$$\Sigma_{\uparrow\uparrow}(\omega,k) = \begin{pmatrix} \frac{U^2}{4(i\omega+\Delta)} & 0 \\ 0 & \frac{U^2}{4(i\omega-\Delta)} \end{pmatrix}, \tag{5.48}$$

which is indeed purely local, i.e., lacks any momentum dependence.

For the other limit ii) $\Delta \to \infty$ or $\Delta \gg U \gg t$, the ground state is given by $|\psi_1\rangle$ of Eq. 5.24, i.e., the A sublattice is doubly occupied and the B sublattice is completely empty. Since we assume U to be finite, the repulsion is not strong enough to occupy B sites, where the potential is much larger. At half filling this requires all A sites to be doubly occupied, i.e., $\langle n_{A\uparrow} n_{A\downarrow}\rangle = 1$, since the number of A and B sites is equal. For the Green's function we then have

$$\begin{aligned}G_{AA,\uparrow\uparrow}(i\omega) &= \frac{1}{i\omega + \frac{U}{2} - \Delta - (U-2\Delta)} \\ &= \frac{1}{i\omega + \Delta - \frac{U}{2}}.\end{aligned} \tag{5.49}$$

For the other matrix elements we have to take care of signs, since

$$c^\dagger_{B\uparrow}|\mathrm{gs}\rangle = -|\psi_2\rangle^{N+1}, \tag{5.50}$$

and we obtain

$$G_{BB,\uparrow\uparrow}(i\omega) = \frac{1}{i\omega + \frac{U}{2} + U - 2\Delta - (U-\delta)} \tag{5.51}$$

$$= \frac{1}{i\omega + \frac{U}{2} - \Delta}. \tag{5.52}$$

The off-diagonal terms vanish and we have for the self-energy

$$\Sigma(i\omega, k) = \begin{pmatrix} 2t\cos(k_y) + U - 2\Delta & t(1+e^{-ikx}) \\ t(1+e^{ikx}) & 2t\cos(k_y) + 2\Delta \end{pmatrix}, \tag{5.53}$$

for which the leading term in Δ is

$$\Sigma(i\omega, k) = \begin{pmatrix} U - 2\Delta & 0 \\ 0 & +2\Delta \end{pmatrix}. \tag{5.54}$$

Hence, both limits feature a completely local self-energy. The interesting question is now what happens in the intermediate region (region A shown in Fig. 5.10). It is clear that the staggered ionic potential Δ will dampen the momentum-dependence monotonically. For U, however, the situation is different. We know that, of course, $\Sigma_{U=0} = 0$. Therefore, the self-energy dispersion amplitude must first increase as a function of the onsite interaction and then decrease again towards the atomic limit. In order to make a prediction for the DMFT error in the ionic Hubbard model (and related models with an ionic potential) we need to investigate the self-energy dispersion as a function of both Δ and U, and map out a region where the self-energy is well-described by a local approximation (this is for sure true for $\Delta \gg U$) and also the opposite, i.e., where the self-energy features a very prominent momentum-dependence, which leads to a large error in DMFT.

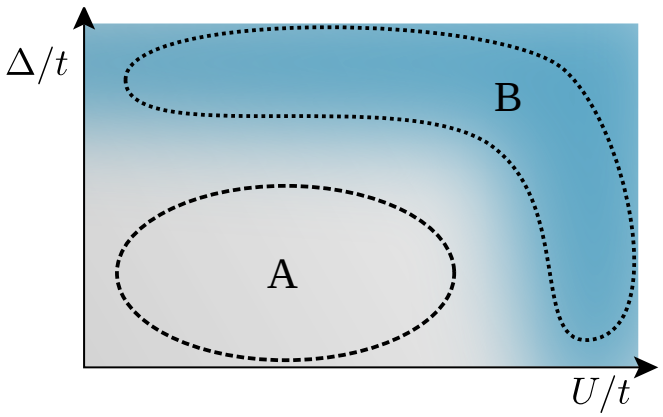

Figure 5.10: Expected behavior of the self-energy dispersion amplitude d_a as a function of Δ and U. Through the discussion of the limiting cases we already know that the momentum-dependence vanishes towards large values of both Δ/t and U/t. This is illustrated by the blue region B, where both localizing terms dampen the momentum dependence. On the other hand, we expect an intermediate region A at $t \sim U, \Delta$, where the momentum-dependence is estimated to be important.

5.4.2 Numerical Results

We begin by reviewing the $\Delta = 0$ case, which we assume to be most difficult for DMFT and which serves as a benchmark of our methods. The ionic Hubbard model reduces to the simple one band square lattice, where a large quantity of reference material is available, see, e.g., the review Ref. [159]. The choice for the square lattice is motivated by the fact that correlations are very strong and antiferromagnetic order is presently projected to occur at any $U > 0$ at $T = 0$ and small U for small finite T, which DMFT incidentally fails to capture. In contrast to momentum-resolved methods, DMFT draws the critical interaction strength at $U_c/t \approx 10$ at low T. Due to our error estimate from Eq. 5.10 we conclude that the self-energy in the 2D square lattice must be rather strong to account for the stark deviation of the local result from that of other methods that take the self-energy dispersion into account. Therefore, the square lattice can be considered a limiting case, which serves as an upper bound for the problematic region A (see Fig. 5.10) and at the same time as a lower bound for the region of validity of DMFT (region B). For other lattices, such as triangular, honeycomb etc., the momentum dependence is expected to be weaker and therefore DMFT will provide more accurate results. The issues are further alleviated in higher spatial dimensions, where the DMFT phase diagram resembles that of, e.g., the dynamical vertex approximation (DΓA) much more closely. One-dimensional systems on the other hand are seldom studied within DMFT, which can be seen as a low-order expansion in the inverse dimension and therefore improves with increasing d. For such systems more specialized methods like density matrix renormalization group (DMRG) and other matrix product state (MPS) variants are available, which themselves are more difficult to apply to two-dimensional systems and more or less do not work practically in three dimensions.

We choose here one momentum-resolved method, the two-particle self-consistent (TPSC) method [160–162], which we use as a reference for the momentum-dependent self-energy. The TPSC data was kindly provided by Karim Zantout. In the following calculations we fix the inverse temperature to $\beta t = 10$, which is low enough to observe strong momentum dependence but not so low as to lead to numerical instabilities in DMFT (note the exponential scaling in β of the employed CT-QMC algorithm that is used as the impurity solver [55, 163, 164]).

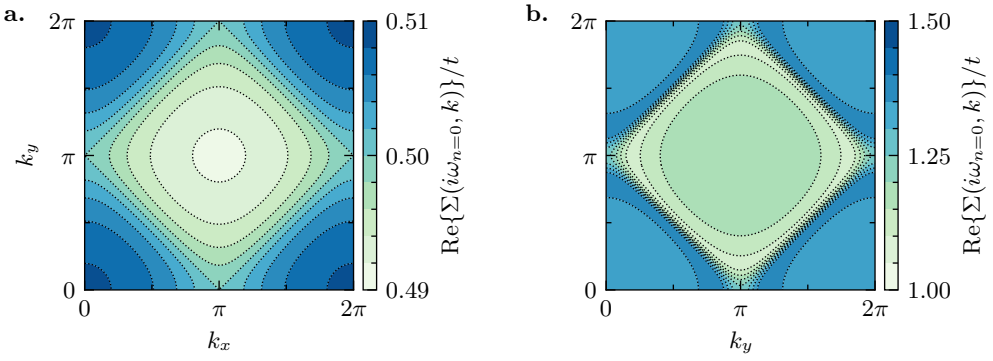

Figure 5.11: Momentum-resolved real part of the TPSC self-energy $\mathrm{Re}\{\Sigma(i\omega_{n=0}, k)\}/t$ at $i\omega_0 = \frac{\pi}{\beta}$ for the square lattice at **a.** $U/t = 1$ and **b.** $U/t = 2.5$ ($\Delta = 0$). The value of $\mathrm{Re}\Sigma$ is centered around $\Sigma_{\mathrm{HF}} = U/2$. The momentum dependence is weak at $U/t = 1$ and rather strong at $U/t = 2.5$, where we measure values of $d_a(i\omega_0)/t = 0.02$ and 0.5, respectively. The scale-independent dispersion amplitude d_r is 0.04 and 0.4, respectively, indicating a much lower relevance of the momentum dependence at low U. [Figure adapted from Ref. [135]]

In Fig. 5.11 we plot the real part of the TPSC self-energy $\mathrm{Re}\{\Sigma(i\omega_{n=0}, k)\}$ at two values of the onsite interaction strength $U/t = 1$ and 2.5. We note that at half filling the static Hartree-Fock mean field value for the self-energy is $\Sigma_{\mathrm{HF}} = \frac{U}{2}$ and the momentum dependence is an

added correction to this value. At low U/t the scale is rather narrow and ranges only from 0.49 to 0.51, i.e., the maximal deviation from the Hartree-Fock value is only 0.01. The self-energy dispersion amplitude (Eq. 5.8) is in this case $d_a(i\omega_0)/t = 0.02$, and we classify the importance of the momentum-dependence via Eq. 5.9 as $d_r(i\omega_0)/t = 0.04$. At intermediate $U/t = 2.5$ the Hartree-Fock value is $\Sigma_{\rm HF} = 1.25$ and again, the real part of the self-energy is centered around this value. Here, the scale is expanded and we observe values ranging from $\approx$1 to $\approx$1.5, which amounts to a maximal deviation of 0.25 from the average and yields $d_a(i\omega_0)/t = 0.5$ and $d_r(i\omega_0)/t = 0.4$. According to our measure, the self-energy dispersion is therefore 10 times more important for intermediate U. The plots here serve as a visual representation of our momentum dependence measures.

For further validation we also compare the spectral function at $\omega = \mu = \frac{U}{2}$ with that obtained by CPT. The non-interacting spectral function is given by

$$A_0(\omega, k) = -\frac{1}{\pi}\mathrm{Im}\left\{\frac{1}{\omega + i\eta + 2t(\cos(k_x) + \cos(k_y))}\right\} \tag{5.55}$$

$$= -\frac{1}{\pi}\mathrm{Im}\left\{\frac{\omega - i\eta + 2t(\cos(k_x) + \cos(k_y))}{[\omega + 2t(\cos(k_x) + \cos(k_y))]^2 + \eta^2}\right\} \tag{5.56}$$

$$= \frac{1}{\pi}\frac{\eta}{[\omega + 2t(\cos(k_x) + \cos(k_y))]^2 + \eta^2}. \tag{5.57}$$

We computed the spectral function for the interacting system with CPT and compare with the TPSC spectral function in Fig. 5.12. Since TPSC works with the Matsubara Green's function we typically have to rely on analytic continuation methods to obtain real frequency data. In this case, however, we are only interested in the zero-frequency value. Given that

$$G(i\omega_n, k) = \frac{1}{i\omega_n + \mu - H_0(k) - \Sigma(i\omega_n, k)}, \tag{5.58}$$

and upon comparison with the retarded Green's function

$$G^R(\omega, k) = \frac{1}{\omega + i\eta - H_0(k) - \Sigma(\omega, k)}, \tag{5.59}$$

we find that for $\eta = \omega_0$ we can approximate

$$G^R(\mu, k) \approx G(i\omega_0, k), \tag{5.60}$$

and therefore

$$A_{\rm TPSC}(\mu, k) \approx -\frac{1}{\pi}\mathrm{Im}\{G(i\omega_0, k)\}. \tag{5.61}$$

In Fig. 5.12**a** we show the non-interacting spectral function $A_0(\mu, k)$ from Eq. 5.57, in Fig. 5.12**b** the TPSC spectral function evaluated according to Eq. 5.61 and in Fig. 5.12**c** the corresponding CPT result for an 8-site cluster at $U/t = 1$. Upon comparison we find that CPT predicts a rather strong momentum-dependence of the self-energy as indicated in the momentum-dependent renormalization of the spectral weight. At $(\pi, 0)$ and $(0, \pi)$ the spectral function is clearly dampened with respect to the non-interacting value. The renormalization in TPSC is less pronounced in the plot (but visible in the data), which is a consequence of the large Lorentzian broadening η. The CPT data shown is computed with $\eta = 0.01$, the plot at $\eta = \omega_0$ looks almost identical to the TPSC data. Since TPSC has a momentum-dependent self-energy that we have already looked at, cf. Fig. 5.11, we conclude that either a good analytic continuation or extrapolation to $\omega = 0$ rather than the approximation of Eq. 5.61 or lower temperature would be necessary to observe

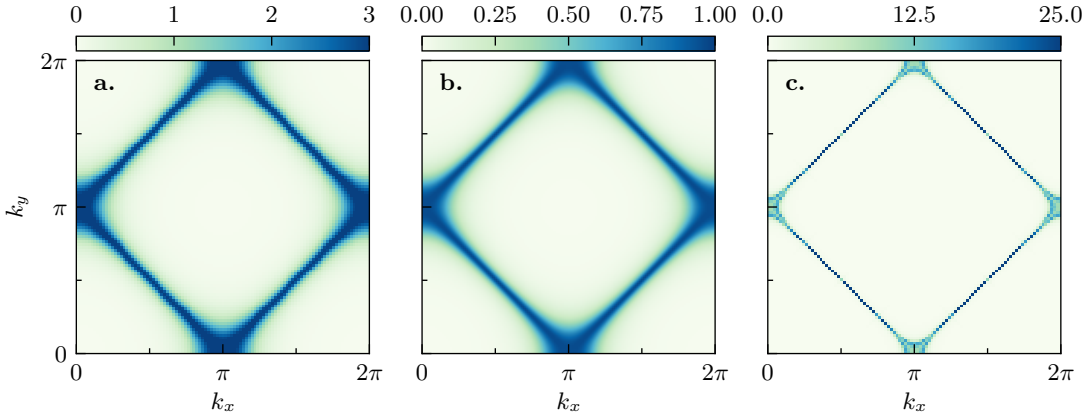

Figure 5.12: Spectral function $A(\omega = \mu, k)$ of the square lattice in units of $[t^{-1}]$. **a.** non-interacting limit (Eq. 5.57) with $\eta = \pi/\beta \approx 0.314$, **b.** TPSC result approximated with Eq. 5.61 and **c.** CPT result averaged over two 8-site clusters with $\eta = 0.01$. **b.**,**c.** at $U/t = 1$. A_0 and A_{TPSC} are qualitatively very similar, but at the value of η chosen the TPSC result does not reveal the momentum-dependence. The CPT result, shows a clear momentum-dependence in terms of damping of $A_{\mathrm{CPT}}(\mu, k)$ around $k = (0, \pi)$ and $k = (\pi, 0)$.

the same momentum-dependence in the spectral function. In any case, the CPT result at $T = 0$ provides us with another indication that the self-energy must be momentum-dependent, even at small U/t.

We note that even though the CPT spectral function looks very promising and compares well with TPSC (given the same η), the method is by construction not useful for the calculation of the self-energy dispersion amplitude $d_a(\omega)$. Due to the inherent translational symmetry breaking a reliable momentum-dependent self-energy cannot be obtained from CPT, see Sec. 3.4.

The TPSC method on the other hand has been benchmarked extensively for the square lattice [161], recently also for multi-orbital models [162], and our result also agrees well with the dual fermion result shown in Refs. [165, 166].

Having established the baseline at $\Delta = 0$ we compute the self-energy with TPSC for various values of Δ/t and U/t to measure the amount and importance of the momentum-dependence. According to our definitions for the absolute and relative self-energy dispersion amplitude of Eq. 5.8 and Eq. 5.9 there is still a residual frequency-dependence left that adds ambiguity. Since the self-energy—like the Green's function—decays like $1/\omega$ for large frequencies we expect the maximal dispersion to appear at low frequencies. To find the optimal value of ω suitable for an upper bound, i.e., where $d_a(\omega) = \max.$, we compute $d_a(\omega)$ for different frequencies. The result is shown in Fig. 5.13 for $U/t = 0.6$. Focusing first on $\Delta/t = 0$, apparently, the maximum is indeed at the lowest Matsubara frequency $i\omega_0$ and we observe the power law decay with increasing n. Two values for $d_a(i\omega_n)$ are shown for each Δ/t: full lines and dotted lines, where we set either off-diagonal or diagonal matrix elements to zero to reveal which matrix elements contain the strongest momentum-dependence. For the square lattice d_a is much stronger on the diagonal. However, increasing Δ/t starts to dampen the dispersion of the self-energy, at least in the diagonal matrix elements. While off-diagonal matrix elements at first appear rather stable w.r.t. increasing the strength of the ionic potential, the dispersion strength finally drops off at $\Delta/t > 1$. For the limit $\Delta \gg U, t$ we can therefore project the non-dispersive limit that we predicted in section Sec. 5.4.1.

The distribution over diagonal and off-diagonal matrix elements is not particularly important for the remainder of the discussion, since in the following we will always compute $d_a(\omega)$ using

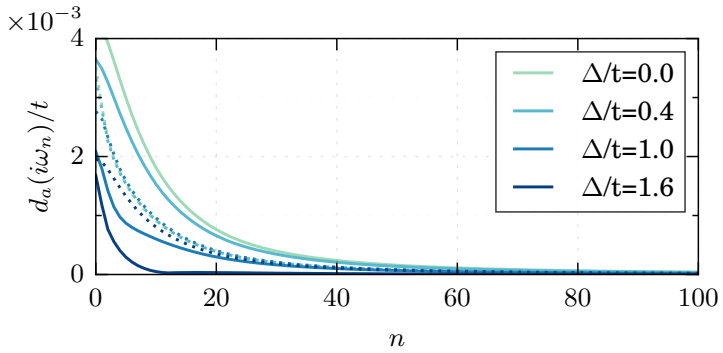

Figure 5.13: Absolute self-energy dispersion amplitude $d_a(i\omega_n)$ as a function of Matsubara frequencies for a range of Δ/t at $U/t = 0.6$. Filled lines correspond to diagonal matrix elements and dotted lines are off-diagonal matrix elements. All matrix elements decay as a function of frequency. Off-diagonal matrix elements are only weakly dependent on Δ, so that at large Δ the dispersion of off-diagonal matrix elements dominates that of the diagonal. The maximum is found at $i\omega_0 = \pi/\beta$ for all cases. [Figure from Ref. [135]]

the full self-energy matrix. However, having established that $d_a(i\omega_n)$ has a global maximum at $i\omega_0$ we can simplify the upper limit of the DMFT error from Eq. 5.10 in the following by defining the upper bound

$$\varepsilon(\omega) = \| \Sigma_{\text{exact}}(\omega, k) - \Sigma_{\text{DMFT}}(\omega) \| \leq d_a(\omega = 0) + r(\omega = 0), \tag{5.62}$$

where we assume that the residual local error $r(\omega)$ follows the same frequency-dependence. Incidentally, the zero frequency value of the self-energy, which obtains the strongest corrections by non-local methods, is also the one that appears in the classification of topological phases in terms of the topological Hamiltonian.

We proceed by computing the momentum-dependent self-energy also for a range of values for the interaction parameter U. Here, we note that even though TPSC is non-perturbative we have to restrict the discussion to the regime of low to intermediate interactions, where the TPSC approximation is valid. Due to the fact that the region of applicability of the topological Hamiltonian is also limited to the low- to intermediate-U regime this does not lower the significance of our study.

The absolute amount of momentum-dependence in terms of the self-energy dispersion amplitude is shown in Fig. 5.14. We clearly observe an increase in d_a as a function of U/t. At the same time the momentum-dependence gets weakened by the ionic potential Δ, i.e., the gradient of d_a points away from large U and towards large Δ as we expected due to our investigation of the limiting cases. By inspecting the data we conclude that values of $U/\Delta > 2$ are necessary for the self-energy to remain momentum-dependent despite the localizing ionic potential. Note that the scale in the plot has been chosen such that also large values are well-represented, therefore the seemingly vanishing dispersion for $U/t < 1.5$ is deceptive. We have identified the lack of a reference value as a weakness of the d_a measure already and have also provided an alternative, the relative dispersion amplitude, that measures the importance of the self-energy dispersion and has the property of a fixed scale ($d_r \in [0, 1]$).

In order to understand the transition from strongly momentum-dependent to predominantly local as a function of the ionic potential Δ we develop an analysis based on the energy scales involved. As we have established earlier, the competition between the parameters U and Δ leads to the transition as soon as Δ overwhelms U, since U introduces non-local Feynman diagrams, while Δ pushes the energies of the two sublattices apart in order to decouple them from each other. This is valid at weak to intermediate interaction strengths, since in the strong coupling

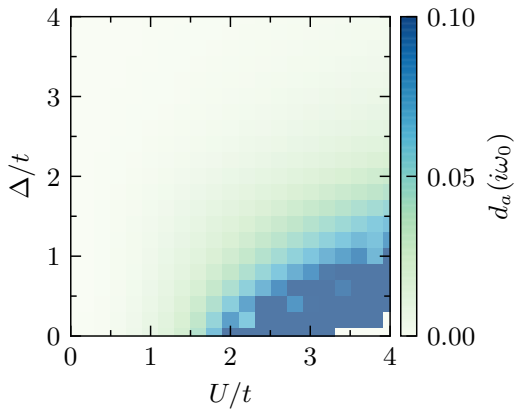

Figure 5.14: Absolute self-energy dispersion amplitude $d_a(i\omega_0)$ at the smallest Matsubara frequency for the ionic Hubbard model on the square lattice. d_a increases as a function of U, while large values are obtained for $U/t \geq 2$. The ionic potential Δ on the other hand dampens d_a, therefore we find only relatively small values at large $\Delta > U/2$. Since the scale of d_a is not well-defined, small values do not necessarily mean that the momentum-dependence is unimportant.

regime we have seen that U alone drives the system into a local limit. As a first approximation we compare the energy cost required by each potential to realize the state favored by the other. The ionic potential favors double occupancy of A sites, i.e., a state

$$|\text{ionic}\rangle = c^\dagger_{A\uparrow} c^\dagger_{A\downarrow} |\text{vac}\rangle, \tag{5.63}$$

while the repulsive Hubbard interaction favors occupation of different sites

$$|\text{Hubbard}\rangle = c^\dagger_{A\uparrow} c^\dagger_{B\downarrow} |\text{vac}\rangle. \tag{5.64}$$

The energy cost to move from one to the other is given by $E_{\text{ionic}} - E_{\text{Hubbard}} = -2\Delta + U$, where 2Δ amounts to the energy needed to spread the electrons evenly among the two sublattices and U is the cost of doubly occupying a site. For the point of equilibrium, where the transition happens, we therefore expect the ratio

$$\frac{\Delta_c}{U_c} = \frac{1}{2}. \tag{5.65}$$

This crude but simple argument provides a physical understanding of the observed data, where we have also noted that roughly $\Delta > U/2$ is required for the dispersion to be suppressed by Δ.

We can follow the same idea in a more precise form to improve the estimate for the transition line. To this end we compare the full energies corresponding to the ionic potential and Hubbard interaction, respectively. Since $E_{\text{ionic}} \leq 0$ for physical states that have a B-site occupation of at most $n_B = n_A$ and $E_{\text{Hubbard}} \geq 0$ we expect the critical values to satisfy $-E_{\text{ionic}} = E_{\text{Hubbard}}$ and with

$$E_{\text{ionic}} = \langle H_{\text{ionic}} \rangle = -\Delta \langle n_A \rangle + \Delta \langle n_B \rangle, \tag{5.66}$$

and

$$E_{\text{Hubbard}} = \langle H_{\text{Hubbard}} \rangle = U \langle n_{A\uparrow} n_{A\downarrow} \rangle + U \langle n_{B\uparrow} n_{B\downarrow} \rangle, \tag{5.67}$$

we obtain

$$-\Delta_c (\langle n_B \rangle - \langle n_A \rangle) = U_c (\langle n_{A\uparrow} n_{A\downarrow} \rangle + \langle n_{B\uparrow} n_{B\downarrow} \rangle) \tag{5.68}$$

$$\Rightarrow \qquad \frac{\Delta_c}{U_c} = \frac{\langle n_{A\uparrow} n_{A\downarrow} \rangle + \langle n_{B\uparrow} n_{B\downarrow} \rangle}{\langle n_A \rangle - \langle n_B \rangle} \tag{5.69}$$

$$=: \frac{D_A + D_B}{n_A - n_B}. \tag{5.70}$$

We can verify immediately that Δ_c/U_c is positive, since $\langle n_A\rangle > \langle n_B\rangle$ for finite Δ. Further, we find that $0 \leq \langle n_A\rangle, \langle n_B\rangle \leq 2$ and $0 \leq D_A, D_B \leq 1$ and with $n_A + n_B = 2$ at half filling we have $0 \leq n_A - n_B \leq 2$. Hence,

$$\frac{\Delta_c}{U_c} \leq 1, \tag{5.71}$$

which immediately implies that throughout at least half of the Δ-U phase diagram the self-energy is only relatively weakly momentum-dependent. Of course, the simple ansatz Eq. 5.65 is compatible with Eq. 5.71 and within the Hartree approximation we can simplify Eq. 5.70 to

$$\left[\frac{\Delta_c}{U_c}\right]_{\text{Hartree}} = \frac{\langle n_{A\uparrow}\rangle\langle n_{A\downarrow}\rangle + \langle n_{B\uparrow}\rangle\langle n_{B\downarrow}\rangle}{\langle n_A\rangle - \langle n_B\rangle} \tag{5.72}$$

$$= \frac{\langle n_{A\uparrow}\rangle^2 + \langle n_{B\uparrow}\rangle^2}{\langle n_A\rangle - \langle n_B\rangle} \tag{5.73}$$

$$= \frac{1}{4}\frac{\langle n_A\rangle^2 + \langle n_B\rangle^2}{\langle n_A\rangle - \langle n_B\rangle} \tag{5.74}$$

$$= \frac{1}{4}\frac{n^2 - 2nn_B + 2n_B^2}{n_A - n_B} \tag{5.75}$$

$$= \frac{1}{4}\frac{4 - 4n_B + 2n_B^2}{2 - 2n_B} \tag{5.76}$$

$$= \frac{1}{2}\left(\frac{2 - 2n_B + n_B^2}{2 - 2n_B}\right) \tag{5.77}$$

$$= \frac{1}{2}\left(1 + \frac{n_B^2}{1 - 2n_B}\right), \tag{5.78}$$

where we used $n := n_A + n_B = 2$. To lowest order in n_B the result Eq. 5.78 is equal to $\frac{1}{2}$, i.e., the approximation that we arrived at earlier.

For the atomic limit the ground state is known exactly. At $T = 0$ we can extract the double occupancy

$$D_A = \Theta(2\Delta - U), \tag{5.79}$$

i.e., the system exhibits a first-order phase transition from a paramagnet to an antiferromagnet at $U = 2\Delta$. This corresponds exactly to our estimate of Eq. 5.65, i.e., the line separating weak and strong momentum-dependence at finite t becomes a first order phase transition in the atomic limit $t/U = 0$. At finite temperature T the transition is broadened and we obtain from Eq. 5.70 $U_c/\Delta_c = 2\tanh(2\beta\Delta_c) + 2$.

We now compute the critical line for finite hopping t/U numerically. Since the right hand side of Eq. 5.70 is basically a function of U and Δ we cannot simply provide an analytic expression. Therefore, we compute D_A, D_B, n_A, n_B on a grid of U, Δ values and then look for solutions to Eq. 5.70. In practice we compute

$$h(\Delta, U) = U(D_A + D_B) + \Delta(n_B - n_A), \tag{5.80}$$

and obtain the critical line by plotting the contour at $h = 0$, i.e., the roots of h. The result is shown in Fig. 5.15, where we show data obtained with TPSC, DMFT and ED. In Fig. 5.15**a** we compare with the diagram of the absolute dispersion amplitude and find that while the region with large momentum-dependence is contained below the critical line, it is not immediately clear that there is a relation between the two. In Fig. 5.15**b** we plot instead the relative dispersion amplitude $d_r(i\omega_0)$, cf. Eq. 5.9, which encodes the importance of the momentum-dependence. The critical line obtained from our energy analysis fits very nicely at low U/t and overshoots the dispersive region a bit at large U/t. Apparently, the region where the self-energy is most

important can be described by the two energy scales of the potentials U and Δ. The separation of the critical line from the strongly momentum-dependent regime at large U/t could have several reasons, one of which is clearly also the accuracy of the numerical methods including TPSC. On the other hand our analysis does not take into account the kinetic energy, which means that the critical line has to approach $\Delta_c/U_c = 1/2$ for large U/t.

We note that while in principle $d_r \in [0, 1]$, we nevertheless chose an arbitrary scale for the plot here, which facilitates the good fit between our critical line and the self-energy data. Although this seems random at first, we can extract an additional piece of information. Namely, "important" self-energy as defined by the energy criterion (ionic potential energy larger smaller than Coulomb repulsion) means that the momentum-dependence is at least 10% of the local value. Conversely, points in the region above the critical line will have a self-energy dispersion of much less than that ($d_r \sim \mathcal{O}(1)$), which is a good indication that the DMFT error is small. A remarkable result of this analysis is that we have constructed an implicit measure of the non-

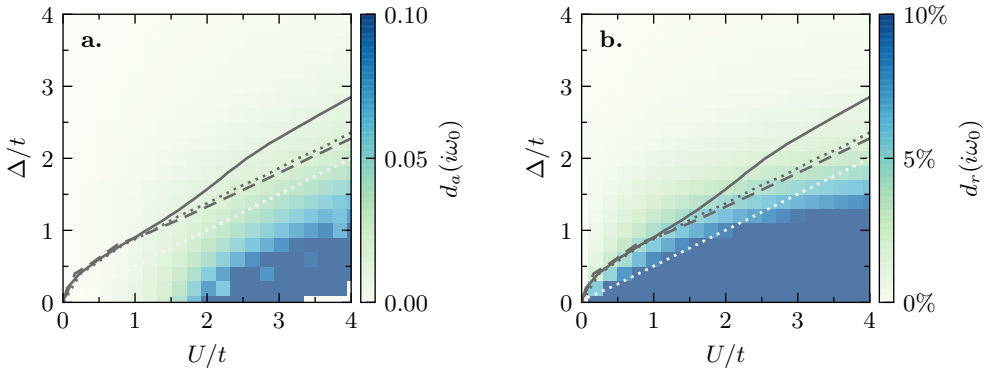

Figure 5.15: a. Absolute and **b.** relative self-energy dispersion amplitude for the square ionic Hubbard model. Lines represent the critical line Eq. 5.65 (white, dotted) and Eq. 5.70 (gray lines). The latter was computed with TPSC (solid), DMFT (dotted) and ED (dashed). The critical line provides a suitable bound in both cases, however, the fit is better for the relative dispersion amplitude, especially at small U. [Figure adapted from Ref. [135]]

local corrections to the self-energy based on purely local quantities. Hence, even a local method such as DMFT can produce an estimate for the critical line. Considering the close relation of the self-energy dispersion amplitude and the DMFT error, cf. Eq. 5.10, the critical ratio Eq. 5.70 can be used as an internal error check for DMFT that provides an estimation of the expected quality of the local approximation. Within our calculations we have found remarkable agreement of the DMFT prediction with both TPSC and ED, where only TPSC deviates towards larger values of U, which could also be a consequence of TPSC's loss of precision at stronger interaction strength.

Finally, we find that even the simple estimate of Eq. 5.65, which does not require any calculations, provides a reasonable enough approximation of the critical line and therefore could be used to judge the applicability of the DMFT approximation.

5.4.3 Comparison with DMFT

So far, we have studied the upper bound for the DMFT error solely by means of the self-energy dispersion amplitude that we calculated within TPSC. In the following we want to investigate the type of error that is to be expected for a selection of quantities and verify our findings based on a direct comparison of observables between TPSC and DMFT.

For our comparison we choose an inherently non-local quantity, here the Green's function

itself and for the local observables the sublattice densities n_A, n_B and the double occupancies D_A, D_B. For the Green's function we note that since

$$A(\omega, k) = -\frac{1}{\pi}\mathrm{Im}\left\{\mathrm{tr}\left[G^R(\omega, k)\right]\right\} \tag{5.81}$$

$$= -\frac{1}{\pi}\sum_{i\in\{A,B\}}\mathrm{Im}\left\{G^R_{ii}(\omega, k)\right\}, \tag{5.82}$$

it makes sense to compare the diagonal matrix elements of the Green's function, where the A sublattice is naturally more interesting. We show the comparison of G_{AA} for a selection of high-symmetry k-points computed with TPSC and DMFT in Fig. 5.16**a** for $U/t = 2$. At $\Delta = 0$, where the momentum-dependence in the bare square lattice is projected to be rather important, we observe a significant deviation at the $\Gamma = (0, 0)$ and $Y = (0, \pi)$ points. From the previous comparison of the spectral function, see Fig. 5.12, we would expect a deviation at X and Y due to the reduced spectral weight at these points. The two lines quickly approach each other so that we cannot find a notable error beyond $\Delta/t \approx 1$, which incidentally is also the point where the momentum-dependence of the self-energy vanishes. At intermediate Δ/t there is only a discrepancy at $(\pi/2, \pi)$, which is in part caused by the rough data grid used especially for the DMFT calculations. Due to the large value of the Green's function, the relative error is much smaller than what we observed at $\Delta = 0$. Towards larger Δ/t both methods agree very well. In this case, the imaginary part of the Green's function approaches 0, which is an indication of the metal to band insulator transition driven by the ionic potential, which opens a spectral gap at around $\Delta/t = 3$.

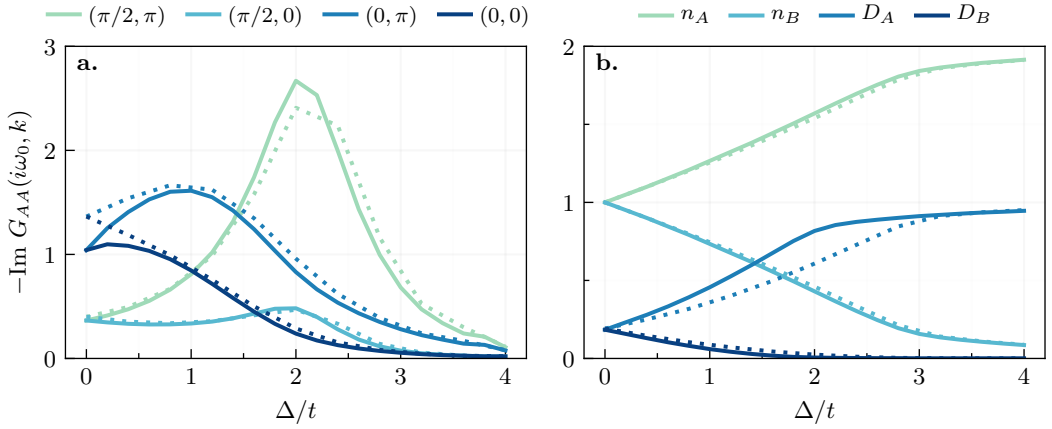

Figure 5.16: **a.** Imaginary part of the Green's function at various k-points computed with TPSC (solid lines) and DMFT (dotted lines). At $\Delta = 0$ there is a noticeable deviation at $\Gamma = (0, 0)$ and $Y = (0, \pi)$. Towards larger Δ/t both methods agree well. **b.** Local observables n_A, n_B (densities on A and B sublattices) and double occupations D_A, D_B, again, for TPSC (solid lines) and DMFT (dotted lines). We observe only a discrepancy in D_A at intermediate Δ. [Figure adapted from Ref. [135]]

In Fig. 5.16**b** we show the densities and double occupancies of the two sublattices. Remarkably, the densities obtained by the two methods are almost exactly the same on both sublattices. We observe clearly the splitting into higher and lower occupied sites and the convergence towards the atomic limit with $n_A = 2$ and $n_B = 0$. The double occupancies, too, converge towards the atomic limit values of $D_A = 1$ and $D_B = 0$, however, at intermediate values of Δ/t we note a quantitative disagreement between TPSC and DMFT, which persists in the region beyond $\Delta/t \approx 1$, where no notable momentum-dependence has been found. In fact, we relate the error

in the double occupancy to the ansatz used for the calculation of certain sum rules in TPSC, that define $\langle nn \rangle$.

5.4.4 Triangular Lattice

The same analysis can of course be performed for other lattices, here we briefly discuss the triangular lattice, which is especially interesting due to the geometric frustration that does not allow for the simple antiferromagnetic order that appears on the square lattice. At the same time we expect the increased connectivity between sublattices to strengthen the role of the kinetic energy. The lattice is shown in Fig. 5.17, where we use the lattice vectors

$$\mathbf{a}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \mathbf{a}_2 = \begin{pmatrix} \cos(\pi/3) \\ \sin(\pi/3) \end{pmatrix}. \tag{5.83}$$

The corresponding reciprocal lattice vectors are given by

$$\mathbf{b}_1 = \begin{pmatrix} \sin(\pi/3) \\ -\cos(\pi/3) \end{pmatrix}, \qquad \mathbf{b}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{5.84}$$

We introduce the ionic potential in a similar way as for the square lattice along the $\mathbf{a}_1$ direction, see Fig. 5.17. Through this choice we essentially only rotated the $\mathbf{a}_2$ lattice vector by 30° compared to the square lattice, which increases the connectivity by 2.

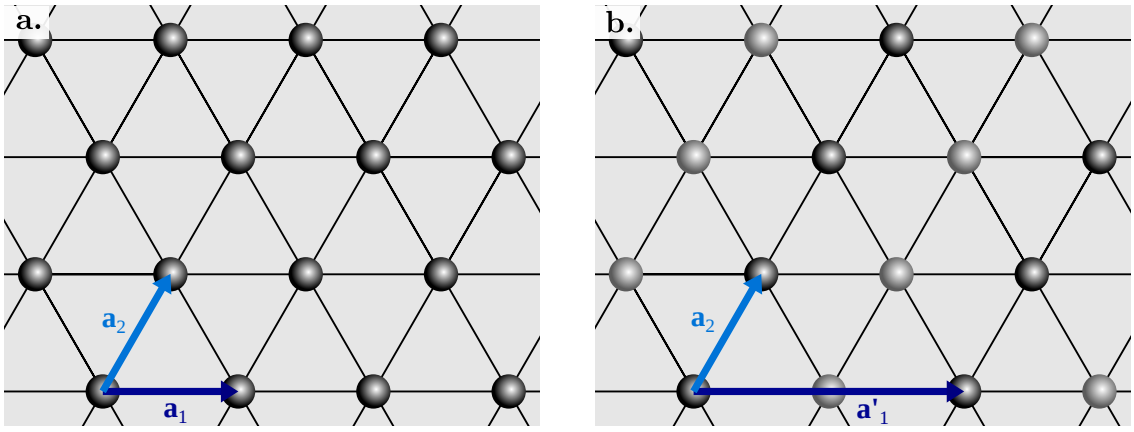

Figure 5.17: **a.** Triangular lattice and lattice vectors. **b.** Modified lattice with ionic potential along the $\mathbf{a}_1$ direction and therefore an enlarged unit cell ($\mathbf{a}'_1 = 2\mathbf{a}_1$). Sites with $-\Delta$ (A sublattice) are dark, sites with $+\Delta$ (B sublattice) are light.

The Hamiltonian of the ionic Hubbard model can then be written on the triangular lattice using Eq. 5.4 and taking into account the six nearest neighbors. In contrast to the square lattice each A site has now two neighbors in the B sublattice and vice versa. The Bloch Hamiltonian is given by

$$H(\mathbf{k}) = \begin{pmatrix} -2t\cos(k_2) - \Delta & -t(1 + e^{-ik_1} + e^{-ik_2}) \\ -t(1 + e^{ik_1} + e^{ik_2}) & -2t\cos(k_2) + \Delta \end{pmatrix}, \tag{5.85}$$

so that the limits are essentially the same as for the square lattice. The spectral function, however, is completely different as demonstrated in Fig. 5.18. We plot here for the purpose of illustration a cut-off function

$$\tilde{A}(\omega, k) = \max\{1, A(\omega, k)\}, \tag{5.86}$$

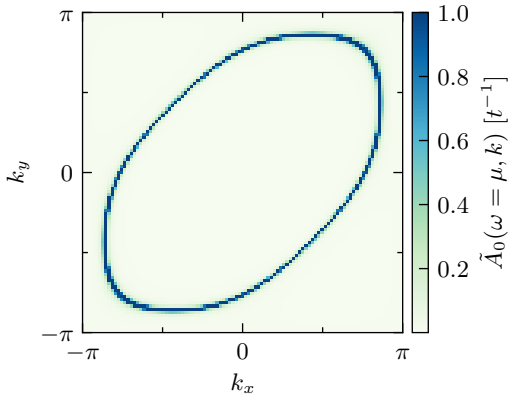

Figure 5.18: Rescaled spectral function $\tilde{A}_0(\omega, k)$ (in units of t^{-1}) of the non-interacting triangular lattice at $\Delta = 0$. In contrast to the square lattice the perfect nesting is lifted, which leads to a reduced momentum-dependence also in the self-energy.

to avoid large peaks due to the δ-singularity at specific momenta. In analogy to Eq. 5.57 we obtain

$$A_0(\omega, k) = \frac{1}{\pi} \frac{\eta}{[\omega - \varepsilon_k]^2 + \eta^2}, \tag{5.87}$$

which apparently has δ-peaks at all k for which $\varepsilon_k = \mu$, since

$$A_0(\omega = \mu, k) = \lim_{\eta \to 0} \frac{1}{\pi} \frac{\eta}{[\mu - \varepsilon_k]^2 + \eta^2} \tag{5.88}$$

$$= \delta(\varepsilon_k - \mu). \tag{5.89}$$

The generally similar structure of $A(\omega, k)$ is therefore not surprising given that all non-interacting models with a Fermi surface look like this. However, in contrast to the square lattice the box-shape, which is often called "perfect nesting" is gone. Considering the definition of the susceptibility at zero frequency

$$\chi^0(0, q) = \sum_n \int G(i\omega_n, k) G(i\omega_n, k + q) \, \mathrm{d}k, \tag{5.90}$$

we see that large contributions are apparently obtained for momenta, k, k', at which the single particle Green's function has similar large values. For the box-shape in the square lattice we can immediately tell that for any momentum k for which the spectral function has a finite value we can define a $k' = k + (\pm\pi, \pm\pi)$, where the spectral function is again very large. Therefore, the susceptibility of the square lattice has maxima at, e.g., $q = (\pi, \pi)$. As a consequence of this strong momentum-dependence we expect also the self-energy to be very dispersive.

For the spectral function of the triangular lattice, on the other hand, that is shown in Fig. 5.18, we find that it is hardly possible to define a small set of possible translation vectors q, since no single vector seems to work for a larger set of k-points. Therefore, we immediately conclude that the self-energy will feature a smaller momentum-dependence than that of the square lattice, which confirms our choice of the square lattice as the prototypical maximally dispersive case, which lends itself as a practical upper bound regarding the momentum-dependence of the self-energy.

Nevertheless, we want to back up our previous claim and compute the momentum-dependence also for the triangular lattice and compare the critical line from our predictions with the actual boundary of the dispersive region. In Fig. 5.19**a** we show the relative self-energy dispersion

amplitude (importance measure) and the critical line (Eq. 5.70) obtained with TPSC, ED and DMFT. Apparently, the momentum-dependence of the self-energy is indeed very weak and the

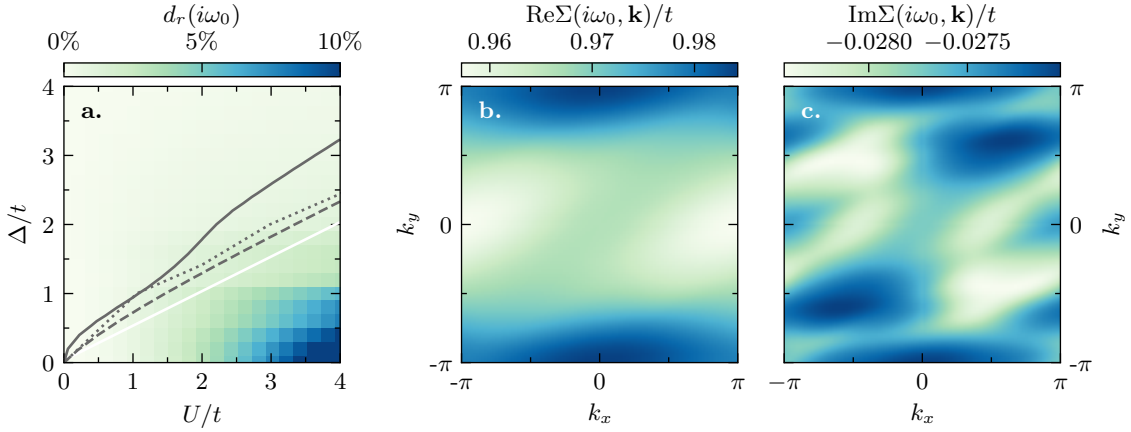

Figure 5.19: a. Relative self-energy dispersion amplitude $d_r(i\omega_0)$ for the ionic Hubbard model on the triangular lattice. The dispersion is strong only at large U and small Δ and the dispersive region lies well below the critical line of Eq. 5.70 obtained with TPSC (solid), DMFT (dotted) and ED (dashed). The solid white line corresponds to the $t/U = 0$ limit. **b.** and **c.**: Real and imaginary part of the TPSC self-energy for the diagonal matrix element on the A sublattice at $U/t = 2$ and $\Delta/t = 0$ for the lowest Matsubara frequency $i\omega_0$. The dispersion is indeed very weak.

dispersive region is shifted towards higher U/t compared to the square lattice. The citical lines are now way above the transition and do not seem to capture the dispersion of the self-energy very well. We can still use theses as an upper bound, though, albeit in a very rough approximation. Looking back at the definition of the critical lines in Eq. 5.70 we notice that the point $U = \Delta = 0$ trivially satisfies the equation, which explains why all lines pass through this point. This was a good description for the square lattice, where the self-energy immediately picks up a momentum-dependence, at arbitrarily small $U/t > 0$. Here, on the other hand, we observe that below $U/t \approx 1$ the self-energy is virtually dispersionless and only assumes values worth mentioning at about $U/t = 2$. An optimal transition line should therefore cut the U-axis at $U/t \approx 2$, which is a feature that our previous energy analysis cannot provide.

In Fig. 5.19**b**,**c** we show the corresponding real and imaginary part of the self-energy at $U/t = 2$ and $\Delta/t = 0$. Only the diagonal matrix element is shown ($\Sigma_{AA} = \Sigma_{BB}$) and we have verified that the momentum-dependence of Σ_{AB} is much weaker. The real part is more dispersive than the imaginary part, which would vanish at precisely $\omega = 0$. The dispersion amplitude is $d_a(i\omega_0) \approx 0.025$, which corresponds to a scale independent relative amplitude of $d_r(i\omega_0) \approx 2.5\%$—a rather low value.

Since the competition of the potential energies cannot fully capture the importance of the momentum-dependence in this case, we compute also the remaining energy component, the kinetic energy, from our ED ground state. We show the result in Fig. 5.20, where we observe in subfigure **a** that the kinetic energy generally increases as a function of both Δ and U. With regards to the negative values we first note that we define here

$$E_{\mathrm{kin}} = -t \sum_{\langle i,j \rangle} \langle c_i^\dagger c_j \rangle, \tag{5.91}$$

where the expectation value is evaluated in the ground state. Hence, the kinetic energy contains also the potential energy of the electrons in the periodic lattice potential, which results in a negative value. For free electrons the kinetic energy would of course be positive.

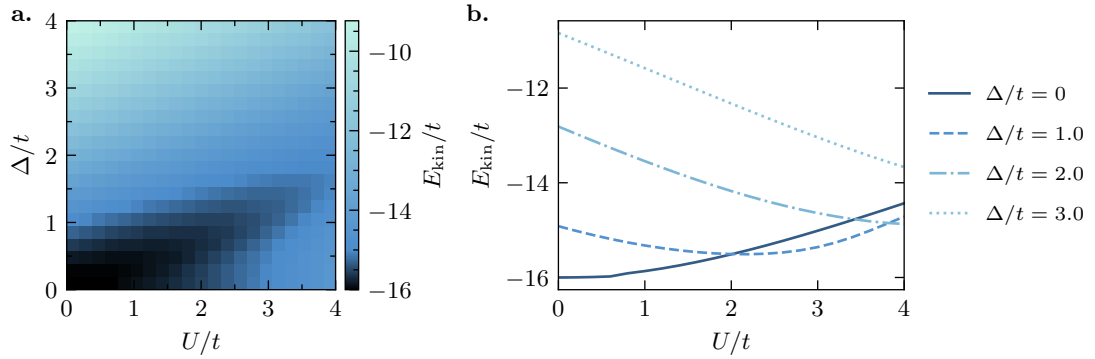

Figure 5.20: a. Kinetic energy of the ED ground state as a function of U/t and Δ/t. **b.** Cuts through **a.** at different Δ/t. The kinetic energy increases with U at $\Delta = 0$, but decreases for small U and finite Δ/t. The region where E_{kin} is much larger compared to the value at $U = 0$ roughly corresponds to the dispersive region.

Given that the ionic potential localizes the electrons to a particular sublattice one would expect the kinetic energy to decrease with increasing localization, however, due to the striped nature of the potential the movement in one lattice direction is still possible. Eventually, at $\Delta/t \to \infty$ the entire movement will freeze out and $E_{\mathrm{kin}} = 0$. In that limit the lattice is irrelevant since all sites are fully occupied.

Regarding the electron-electron interactions we understand that its effect is to separate spins from each other leading to a localization of its own and therefore the kinetic energy will also slowly approach 0 for $U \to \infty$. However, inspecting in particular the cuts at constant Δ/t in Fig. 5.20**b** we find that the kinetic energy gains weight, i.e., becomes more negative at finite Δ/t as a function of U/t. This is again a consequence of the competition between the two potentials and interactions somewhat reversing the localization effect of the ionic potential. The turnaround point for this behavior, i.e., the point where the kinetic energy starts to approach 0 again as a funtion of U/t, which is the minimum in $E_{\mathrm{kin}}(U)$ lies around $U/t \approx 2$ for $\Delta/t = 1$ and $U/t \approx 4$ for $\Delta/t = 2$, which fits rather nicely to the dispersive region. Another point of interest is the largest value of U, where $E_{\mathrm{kin}}(U) - E_{\mathrm{kin}}(U = 0) = 0$, i.e., where U truly overpowers the ionic potential. In order to bring this out visually we plot this difference in Fig. 5.21 with a different color scale.

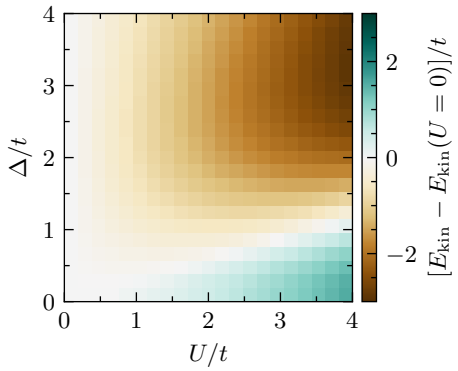

Figure 5.21: Difference of the kinetic energy w.r.t. the non-interacting value. The color scale is chosen such that wherever $E_{\mathrm{kin}} = E_{\mathrm{kin}}(U = 0)$ the color is white. Two distinct regions with opposite signs are revealed. Positive sign means the kinetic energy is larger than the respective reference value in the non-interacting system, which corresponds to the large-U limit, where E_{kin} approaches 0.

Clearly, this graphical representation shows two different regions, one with a negative sign, the other with positive sign. The large-U limit is apparently that where the kinetic energy approaches 0 as a function of U, i.e., where the difference plotted is positive. Inspecting the boundaries of the large-U limit extracted in this way we find good agreement with the dispersive region shown in Fig. 5.19, which backs up our previous claim that the self-energy dispersion is strong only when the electron-electron interaction dominates the ionic potential.

We note that the kinetic energy does not lend itself as a general indicator for the momentum-dependence of the self-energy, since the same behavior is also observed for the square lattice, where the direct comparison of potential energies provides a much more suitable upper bound.

5.4.5 Comment on the Numerical Implementation

The data presented in this chapter was computed using TPSC, ED and DMFT. The TPSC implementation was provided and operated by Karim Zantout. For ED (and CPT) and DMFT we used our own implementation. The impurity solver used by the DMFT code is the hybridization expansion continuous-time Quantum Monte Carlo code (CT-HYB) provided in the open source ALPS package [123, 164, 167, 168]. Unfortunately, the ALPS project has since been abandoned and is no longer maintained. Therefore, attempts to quickly rebuild the program to obtain additional data points in reasonable time have unfortunately been unsuccessful. This should be understood as a warning that relying too much on the maintenance of third party software can be problematic.

5.5 Discussion

We conclude this chapter by briefly summarizing the main results. We have introduced the concept of self-energy dispersion and defined two measures capturing the "amount" and "importance" of the momentum-dependence, respectively. We showed that the absulute self-energy dispersion amplitude is directly related to the error of the DMFT self-energy and motivated that an investigation of this quantity can provide insight into the validity of the DMFT approximation.

The ionic Hubbard model, a bare bones example for many topological models, shows an interesting behavior in that the self-energy is only dispersive at $U > \Delta/2$. Using an analysis of the energy scales involved we have derived an approximation for the critical line that provides an upper bound for the region where DMFT cannot safely be applied. This line depends only on local quantities and can be calculated within DMFT itself, providing an internal accuracy check. Comparison with TPSC and ED (both taking into account non-locality) yields good agreement. As a result, we find that DMFT can be applied with confidence in the majority of the parameter space.

The square lattice can be regarded as an upper bound for the momentum-dependence in two spatial dimensions and we have shown for the triangular lattice that the momentum-dependence is much weaker, which even increases the region of validity of the DMFT approximation. Typical topological models, such as the Haldane or Kane-Mele models, are formulated on Honeycomb lattices, where the momentum-dependence is again much weaker than for the square lattice. An application of the combination of topological Hamiltonian and DMFT should therefore work well in a broad regime of parameters.

Chapter 6

Statistical Analysis of the Chern Number

This chapter revolves around a continuation of our previous work investigating the effect of the self-energy dispersion on the topological classification that we discussed in Chapter 5. Having established the existence of a large region in the parameter space of the usual topological models where the momentum-dependence of the self-energy is irrelevant, we now want to investigate the kinds of errors we can expect when we are dealing with a measurable self-energy dispersion. Here, we focus entirely on the application of the topological Hamiltonian, i.e., we study the possible misclassification in terms of the Chern number as a result of a local approximation.

To this end we developed a statistical analysis that is completely unbiased and therefore provides insight on very general grounds. We expect the results to hold for a multitude of different models.

The analysis is performed at the example of the Haldane model, which we review extensively in Sec. 6.2. Interaction effects can naturally only be discussed for an interacting model, which we obtain by adding a local Hubbard interaction to the non-interacting Hamiltonian. The corresponding variant of the Hubbard model—the Haldane-Hubbard model—is discussed briefly in Sec. 6.3, where we review the current state of the art. The statistical approach is then introduced in Sec. 6.4, where we investigate local and non-local contributions separately. The regime of non-local self-energies, where our method truly unfolds its full potential, is discussed in Sec. 6.5.

Parts of the results discussed in this chapter were published as Ref. [169]:

Thomas Mertz, Karim Zantout and Roser Valentí
Statistical analysis of the Chern number in the interacting Haldane-Hubbard model
Phys. Rev. B **100**, 125111 (2019)

6.1 Motivation

We started this project with a calculation of the phase diagram for the Haldane Hubbard model with TPSC, in order to determine via a comparison with the corresponding DMFT result if the explicit momentum-dependence leads to a measurable deviation of transition lines between

topological phases. Unfortunately, the different regions of applicability of the two methods did not allow us to penetrate the most interesting region in parameter space, where the energy scale of the Coulomb repulsion between electrons starts to dominate the interaction with ionic sites. Although the internal checks of the TPSC routine were not alerting us of imminent problems, experience motivated us to take the TPSC results in the regime of stronger interactions with a grain of salt.

Instead, we tried to formulate analytical requirements for the momentum-dependent self-energy to change the topological classification with respect to the non-interacting result. This idea was mainly based on the requirement that if the bare Hamiltonian and the self-energy commute, i.e.,

$$[H_0(k), \Sigma(k)] = 0 \quad \forall k, \tag{6.1}$$

the topological index necessarily remains the same with or without Σ. This, however, turned out to be a rather weak requirement and a finite commutator is not indicative of an interaction-driven topological phase. In fact, we find that the requirement of Eq. 6.1 is fulfilled only for the most trivial cases. In this context we also evaluated the potential of diagrammatic expansions of the self-energy, however, due to the non-linearity of the Chern number and the generally complicated response to changes in the Hamiltonian, these attempts were abandoned.

Given these obstacles we were faced with a choice: either pick another approximate method that yields a self-energy and allows us to produce another version of the phase diagram, or take an entirely different route—acknowledging that we have no means to procure the exact solution to the many-body problem. The latter choice led us to the development of the statistical method that we describe later in this chapter, which is based on an entirely different premise than what is currently applied in the field. Instead of computing a particular solution to the many-body problem we instead evaluate *all possible solutions* in a statistical fashion that allows us to make very general statements about the effect of many-body interactions for topological systems. The specific form of the topological Hamiltonian is very useful in this context as it allows us to reduce the entire problem to a single-particle picture, where this procedure is much more straight-forward. In the interest of full transparency we have to mention at this point that the power of this method comes at a price—since a solution for a particular system is not determined, all statements hold only generally and we can only make probabilistic statements for individual cases.

6.2 The Haldane Model

In this section we review the Haldane model, which serves as a testbed for the following discussion. Initially proposed by Duncan Haldane [143] as a "Model for a Quantum Hall Effect without Landau Levels" this model extends the tight-binding model on the honeycomb lattice to include complex next-nearest neighbor hopping terms that break time-reversal symmetry without requiring a finite net magnetic flux through the lattice. Long being thought of as extremely unlikely to appear in nature and therefore being primarily of theoretical interest, in 2014 researchers at ETH Zurich were able to manufacture a quantum system of ultracold fermions that is indeed described by the Haldane model [170].

The honeycomb lattice is naturally bipartite with a unit cell of two sites. It is composed of two triangular lattices that are shifted against each other such that the sites of the B lattice lie in the center of mass of the triangles formed by the sites of sublattice A. The resulting lattice is illustrated in Fig. 6.1, where we color one sublattice dark, the other light. In Fig. 6.1**a** we highlight one unit cell that corresponds exactly to the unit cell of one of the triangular lattices.

The corresponding lattice vectors are drawn as well and their (normalized) coordinates are

$$\mathbf{a}_1 = \begin{pmatrix} \cos(\pi/3) \\ \sin(\pi/3) \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 \\ \sqrt{3} \end{pmatrix}, \qquad \mathbf{a}_2 = \begin{pmatrix} -\cos(\pi/3) \\ \sin(\pi/3) \end{pmatrix} = \frac{1}{2}\begin{pmatrix} -1 \\ \sqrt{3} \end{pmatrix}. \tag{6.2}$$

In Fig. 6.1**b** we draw also the unit cells that contain one of the nearest neighbors and their co-

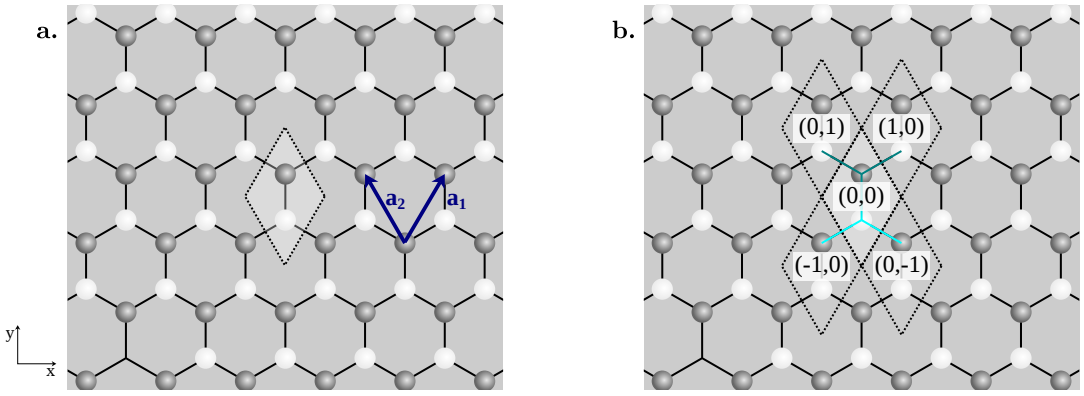

Figure 6.1: Illustration of the honeycomb lattice and the choice of basis. **a.** One unit cell and the corresponding lattice vectors are shown. **b.** We highlight also the neighboring unit cells that contain nearest neighbors of the two sites in the (0,0) cell and indicate their coordinates in terms of the lattice vectors. Links to nearest neighbors are colored, where teal is used for dark to light and turquoise for light to dark.

ordinates in terms of the lattice vectors. For the reciprocal lattice vectors we compute according to Eq. 3.29

$$\mathbf{a}_2 \times \mathbf{a}_3 = \frac{1}{2}\begin{pmatrix} \sqrt{3} \\ 1 \\ 0 \end{pmatrix}, \qquad \mathbf{a}_3 \times \mathbf{a}_1 = \frac{1}{2}\begin{pmatrix} -\sqrt{3} \\ 1 \\ 0 \end{pmatrix}, \tag{6.3}$$

where $\mathbf{a}_3 = (0,0,1)^T$ and we consider $\mathbf{a}_{1,2}$ in three-dimensional space while performing the vector product. Then,

$$\mathbf{b}_1 = \begin{pmatrix} 1 \\ \frac{1}{\sqrt{3}} \end{pmatrix}, \quad \mathbf{b}_2 = \begin{pmatrix} -1 \\ \frac{1}{\sqrt{3}} \end{pmatrix} \tag{6.4}$$

in units of 2π. In Fig. 6.2**a** we plot the reciprocal lattice vectors from Eq. 6.4. The Wigner-Seitz construction of the Brillouin zone is shown explicitly in Fig. 6.2**b**, where orthogonal lines are drawn through the midpoints of the connection lines between integer linear combinations of the reciprocal lattice vectors. In Fig. 6.2**c** we show the Brillouin zone and the high symmetry points.

It is clear that Γ corresponds to the origin of the reciprocal lattice, i.e., $\Gamma = 0\mathbf{b}_1 + 0\mathbf{b}_2$. From the construction in Fig. 6.2**b** it is also immediately clear that $M = \mathbf{b}_1/2$. The explicit derivation of the coordinates of the K points is not so obvious, however. We therefore perform an exact derivation here. Apparently, due to the Wigner-Seitz construction the K and K' points lie on the lines that halve the vector $\mathbf{b}_1 + \mathbf{b}_2$ and $\mathbf{b}_1$ in case of K, $\mathbf{b}_2$ in case of K'. We can decompose the vector K into two components

$$K = \frac{1}{2}\mathbf{b}_1 + c_1\mathbf{t}_1, \tag{6.5}$$

where $\mathbf{t}_1$ is orthogonal to $\mathbf{b}_1$ and therefore this parameterization crosses $\mathbf{b}_1$ at the M point. Clearly, we then have

$$\mathbf{b}_1 \cdot K = \frac{1}{2} \parallel \mathbf{b}_1 \parallel^2 . \tag{6.6}$$

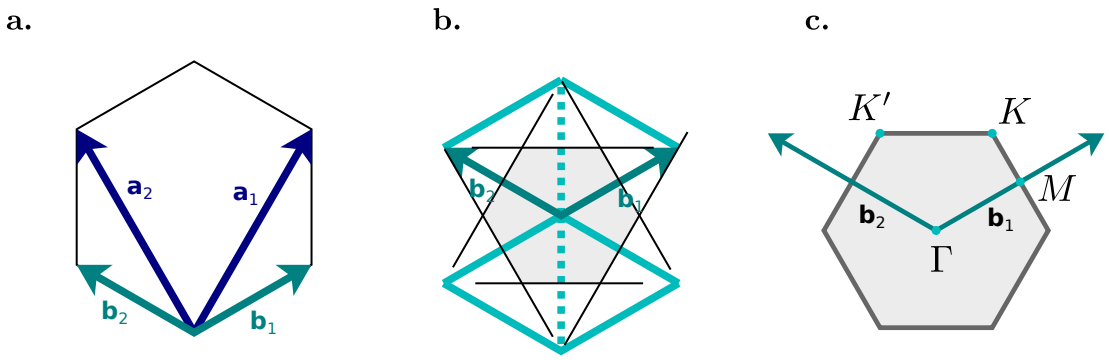

Figure 6.2: **a.** Reciprocal lattice vectors $\mathbf{b}_{1,2}$ of the triangular lattice in relation to the lattice vectors $\mathbf{a}_{1,2}$. **b.** Wigner-Seitz construction of the first Brillouin zone, orthogonal lines are drawn through the midpoints of the Γ–Γ lines. Turquoise lines are copies or linear combinations of $\mathbf{b}_{1,2}$. **c.** First Brillouin zone of the triangular lattice and high symmetry points.

We can do the same in terms of $\mathbf{b}_1 + \mathbf{b}_2$:

$$K = \frac{1}{2}(\mathbf{b}_1 + \mathbf{b}_2) + c_2\mathbf{t}_2, \tag{6.7}$$

$$\Rightarrow \quad (\mathbf{b}_1 + \mathbf{b}_2) \cdot K = \frac{1}{2} \parallel \mathbf{b}_1 + \mathbf{b}_2 \parallel^2, \tag{6.8}$$

where $\mathbf{t}_2$ is an orthogonal vector that intersects $\mathbf{b}_1 + \mathbf{b}_2$ at its midpoint. The same thought applies for K' and we obtain the equations

$$\begin{aligned} \mathbf{b}_2 \cdot K' &= \frac{1}{2} \parallel \mathbf{b}_2 \parallel^2, \\ (\mathbf{b}_1 + \mathbf{b}_2) \cdot K' &= \frac{1}{2} \parallel \mathbf{b}_1 + \mathbf{b}_2 \parallel^2 . \end{aligned} \tag{6.9}$$

Eqs. 6.6, 6.8 and 6.9 represent two systems of linear equations. As in Ref. [171] the solution can be expressed as

$$K^{(\prime)} = (A^{(\prime)})^{-1} C^{(\prime)} \tag{6.10}$$

with

$$A = \begin{pmatrix} (\mathbf{b}_1)_x & (\mathbf{b}_1)_y \\ (\mathbf{b}_1 + \mathbf{b}_2)_x & (\mathbf{b}_1 + \mathbf{b}_2)_y \end{pmatrix}, \quad A' = \begin{pmatrix} (\mathbf{b}_2)_x & (\mathbf{b}_2)_y \\ (\mathbf{b}_1 + \mathbf{b}_2)_x & (\mathbf{b}_1 + \mathbf{b}_2)_y \end{pmatrix}, \tag{6.11}$$

and

$$C = \frac{1}{2}\begin{pmatrix} \parallel \mathbf{b}_1 \parallel^2 \\ \parallel \mathbf{b}_1 + \mathbf{b}_2 \parallel^2 \end{pmatrix}, \quad C' = \frac{1}{2}\begin{pmatrix} \parallel \mathbf{b}_2 \parallel^2 \\ \parallel \mathbf{b}_1 + \mathbf{b}_2 \parallel^2 \end{pmatrix}. \tag{6.12}$$

This construction works even if there is no perfect 60° angle between the lattice vectors. Since we are more interested in expressing K in terms of the reciprocal lattice vectors and not in an orthogonal basis, we review Eq. 6.6 and Eq. 6.8 and find similarly

$$\mathbf{b}_1 \cdot K = c_1\mathbf{b}_1 \cdot \mathbf{b}_1 + c_2\mathbf{b_1} \cdot \mathbf{b}_2 = \frac{1}{2} \parallel \mathbf{b}_1 \parallel^2, \tag{6.13}$$

$$(\mathbf{b}_1 + \mathbf{b}_2) \cdot K = c_1(\mathbf{b}_1 \cdot \mathbf{b}_1 + \mathbf{b}_2 \cdot \mathbf{b}_1) + c_2(\mathbf{b}_1 \cdot \mathbf{b}_2 + \mathbf{b}_2 \cdot \mathbf{b}_2) = \frac{1}{2} \parallel \mathbf{b}_1 + \mathbf{b}_2 \parallel^2, \tag{6.14}$$

where $c_{1,2}$ are the desired expansion coefficients. A solution is given by the inversion of the matrix

$$B = \begin{pmatrix} \parallel \mathbf{b}_1 \parallel^2 & \mathbf{b}_1 \cdot \mathbf{b}_2 \\ \parallel \mathbf{b}_1 \parallel^2 + \mathbf{b}_1 \cdot \mathbf{b}_2 & \mathbf{b}_1 \cdot \mathbf{b}_2 + \parallel \mathbf{b}_2 \parallel^2 \end{pmatrix} = \frac{2}{3}\begin{pmatrix} 2 & -1 \\ 1 & 1 \end{pmatrix}, \tag{6.15}$$

as

$$K = B^{-1}C = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ -1 & 2 \end{pmatrix}\frac{2}{3}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 2 \\ 1 \end{pmatrix}. \tag{6.16}$$

For K' we simply interchange $1 \leftrightarrow 2$ in Eq. 6.13 and define

$$B' = \begin{pmatrix} \mathbf{b}_1 \cdot \mathbf{b}_2 & \| \mathbf{b}_2 \|^2 \\ \| \mathbf{b}_1 \|^2 + \mathbf{b}_1 \cdot \mathbf{b}_2 & \mathbf{b}_1 \cdot \mathbf{b}_2 + \| \mathbf{b}_2 \|^2 \end{pmatrix} = \frac{2}{3}\begin{pmatrix} -1 & 2 \\ 1 & 1 \end{pmatrix}, \tag{6.17}$$

so that

$$K' = (B')^{-1}C' = \frac{1}{2}\begin{pmatrix} -1 & 2 \\ 1 & 1 \end{pmatrix}\frac{2}{3}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 1 \\ 2 \end{pmatrix}. \tag{6.18}$$

We summarize this by noting again that the points $M = (1/2, 0)^T, K = (2/3, 1/3)^T, K' = (1/3, 2/3)^T$ are given in terms of the reciprocal lattice vectors and not in terms of the canonical Cartesian basis. This is much more convenient since we will always work with $\mathbf{k} = (k_1, k_2)^T = k_1\mathbf{b}_1 + k_2\mathbf{b}_2$ with $k_1, k_2 \in [0, 1)$ in units of 2π.

Having established the lattice and reciprocal lattice we can now expand the generic tight-binding model

$$H = t\sum_{i,j} c_i^\dagger c_j. \tag{6.19}$$

Due to

$$|i\rangle = \sum_k |k\rangle\langle k|i\rangle = \frac{1}{\sqrt{N}}\sum_k e^{-ikx_i}|k\rangle, \tag{6.20}$$

we can express the creation operator as

$$c_i^\dagger|\text{vac}\rangle = |i\rangle = \frac{1}{\sqrt{N}}\sum_k e^{-ikx_i}|k\rangle = \frac{1}{\sqrt{N}}\sum_k e^{-ikx_i}c_k^\dagger|\text{vac}\rangle, \tag{6.21}$$

which yields the identities

$$\begin{aligned} c_{i,a}^\dagger &= \frac{1}{\sqrt{N}}\sum_k e^{-ikx_i}c_{k,a}^\dagger, \\ c_{i,a} &= \frac{1}{\sqrt{N}}\sum_k e^{ikx_i}c_{k,a}, \end{aligned} \tag{6.22}$$

where we added an additional sublattice index $a \in \{A, B\}$. Therefore, by defining $\delta x_{ij} = x_i - x_j \in \{(0,0), (0,1), (1,0), (-1,0), (0,-1)\}$ we have

$$H = t\sum_{\langle i,j\rangle}\frac{1}{N}\sum_{k,k'} e^{-ikx_i}e^{ik'x_j}c_{k,a}^\dagger c_{k',b} \tag{6.23}$$

$$= t\sum_{\langle i,j\rangle}\frac{1}{N}\sum_{k,k'} e^{-ik(x_j+\delta x_{ij})}e^{ik'x_j}c_{k,a}^\dagger c_{k',b} \tag{6.24}$$

$$= t\sum_{k,k'}\sum_{\delta x}\frac{1}{N}\sum_j e^{i(k'-k)x_j}e^{-ik\delta x}c_{k,a}^\dagger c_{k',b} \tag{6.25}$$

$$= t\sum_{k,k'}\sum_{\delta x}\delta_{k,k'}e^{-ik\delta x}c_{k,a}^\dagger c_{k',b} \tag{6.26}$$

$$= t\sum_k\sum_{\delta x} e^{-ik\delta x}c_{k,a}^\dagger c_{k,b}. \tag{6.27}$$

Executing the sum we obtain

$$H = t\sum_k \left[c^\dagger_{k,A}c_{k,B} + c^\dagger_{k,B}c_{k,A} + e^{-ik_2}c^\dagger_{k,A}c_{k,B} + e^{-ik_1}c^\dagger_{k,A}c_{k,B} + e^{ik_1}c^\dagger_{k,B}c_{k,A} + e^{ik_2}c^\dagger_{k,B}c_{k,A}\right]$$
$$= t\sum_k \begin{pmatrix} c^\dagger_{k,A} & c^\dagger_{k,B}\end{pmatrix} \begin{pmatrix} 0 & 1+e^{-ik_2}+e^{-ik_1} \\ 1+e^{ik_2}+e^{ik_1} & 0\end{pmatrix} \begin{pmatrix} c_{k,A} & c_{k,B}\end{pmatrix}. \tag{6.28}$$

Here, we used the expansion $k = n_1\mathbf{b}_1 + n_2\mathbf{b}_2$ in terms of reciprocal lattice vectors and the relation $\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi\delta_{ij}$. With this definition, $k_i = 2\pi n_i \in [0, 2\pi)$. The Bloch Hamiltonian $H(k)$ can be read off here. Expressed in terms of Pauli matrices we have

$$H(k) = t(1+\cos(k_1)+\cos(k_2))\sigma_1 + t(\sin(k_1)+\sin(k_2))\sigma_2 \tag{6.29}$$
$$= \mathbf{h}\cdot\boldsymbol{\sigma}, \tag{6.30}$$

where $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)^T$ and $\mathbf{h}$ can be read off from the line above.

In order to construct the Haldane model we now add next-nearest neighbor terms following Haldane's original construction in Ref. [143]. The requirement is that the total flux per unit cell vanishes. In order to conserve the periodicity of the model we choose also the flux to be periodic and with $\Phi = \int_{\gamma=\partial U} \boldsymbol{A}(\boldsymbol{x})\cdot \mathrm{d}\gamma$, where ∂U denotes the boundary of the unit cell, this means that $\boldsymbol{A}(\boldsymbol{x})$ must also be periodic.

In Fig. 6.3**a** we show the lattice and demonstrate that the unit cell can be folded exactly into one hexagon. For the nearest neighbor hopping terms this means that any closed hopping path (that always contains full hexagons) encloses an integer number of unit cells. With the condition of zero net flux through a unit cell this means that the nearest neighbor terms can all be chosen real like in the Graphene model.

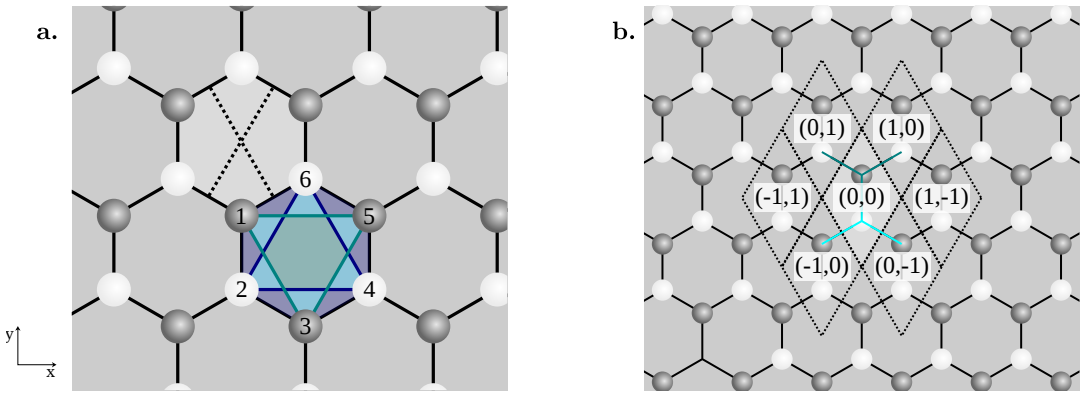

Figure 6.3: Construction of the Haldane model. **a.** We show that one hexagon contains exactly one unit cell (folded in) and three regions with different fluxes Φ_1, Φ_2, Φ_3. The sites are labeled to allow the definition of hopping paths. In **b.** we show the unit cells containing nearest neighbors and next-nearest neighbors and their coordinates. [Partly a visually more appealing reproduction of Fig. 1 from Ref. [143]]

Regarding the next-nearest neighbor hoppings we draw the links between A and B sites. We assume for now that A sites are light and B sites are dark. Then, the blue triangle in Fig. 6.3**a** marks the smallest closed path of next-nearest neighbor hopping among A sites. By breaking time-reversal symmetry locally, the path 2–6–4–2 must enclose some flux $2\pi\Phi_{2-6-4-2}/\Phi_0 = 3\phi_A$, where we distribute the phase evenly among the three links: $2\pi\phi_{ij}/\Phi_0 = \phi_A$. $\Phi_0 = \frac{h}{2e}$ is the flux quantum. Considering instead the path 2–6–1–2 we have two real and only one complex hopping, which indicates that $\phi_A = -2\pi(2\Phi_1 + \Phi_2)/\Phi_0$. The minus sign is a consequence of the different orientation of the boundary w.r.t. the previous path. The same can be found for the

hoppings between B sites and we obtain the equations

$$\phi_A = \phi_B = -2\pi(2\Phi_1 + \Phi_2)/\Phi_0 \tag{6.31}$$
$$3\phi_A = 2\pi(3\Phi_2 + \Phi_3)/\Phi_0 \tag{6.32}$$
$$0 = 6\Phi_1 + 6\Phi_2 + \Phi_3. \tag{6.33}$$

There exist an infinite number of solutions to this system of equations given by

$$\begin{aligned} \frac{\Phi_2}{\Phi_0} &= -2\frac{\Phi_1}{\Phi_0} - \frac{\phi_A}{2\pi}, \\ \frac{\Phi_3}{\Phi_0} &= 6\left(\frac{\Phi_1}{\Phi_0} + \frac{\phi_A}{2\pi}\right), \end{aligned} \tag{6.34}$$

where we can realize any phase ϕ_A for the next-nearest neighbor hoppings in an infinite number of ways by choosing Φ_1 arbitrarily. For example, we could choose $\Phi_1 = 0$ and obtain $\Phi_2/\Phi_0 = -\phi_A/(2\pi)$ and $\Phi_3/\Phi_0 = 6\phi_A/(2\pi)$.

Having established that any value of the phase $\phi = \phi_A = \phi_B$ can be realized by a specific flux pattern we write down the Haldane model in the general form

$$H = t_1 \sum_{\langle i,j\rangle} c_i^\dagger c_j + t_2 \sum_{\langle\langle i,j\rangle\rangle} e^{i\phi_{ij}} c_i^\dagger c_j + m \sum_i \mathrm{sgn}(i) c_i^\dagger c_i, \tag{6.35}$$

where $\mathrm{sgn}(i)$ is ± 1 for the $A(B)$ sublattice and $\phi_{ij} = \pm\phi$ for clockwise/counterclockwise hopping. The additional local potential m accounts for a possible difference in onsite energies between the two sublattices and is identical to the ionic potential discussed in Chapter 5. Here, we use the established nomenclature derived from Dirac Hamiltonians and refer to m as a (trivial) mass term (aka Semenoff mass [172]).

We note that in the preceding discussion we documented the construction used for this choice of lattice vectors. In our numerical implementation we strayed from Haldane's original formulation by defining the phase factor with positive sign for mathematically positive orientation of the closed path, i.e., anti-clockwise hopping. Since the lattice vectors were defined the other way around, however, this corresponds to a mirror under which the hopping direction changes. In a way this means that the definition of the winding strongly depends on the choice of the lattice vectors and therefore both have to be provided to uniquely define the model. Since the flux has been reversed in the mirrored version, everything derived here is consistent with the numerical results, although a different basis is being used.

Details on this "trivial" topic are usually hard to come by and since the formulation in Ref. [143] is rather cumbersome we compare also the details of the definition of lattice vectors. Haldane writes "let $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ be the displacements from a B site to its three nearest-neighbor A sites, defined so that $\hat{\mathbf{z}} \cdot (\mathbf{a}_1 \times \mathbf{a}_2)$ is positive" [143]. Apparently, the latter condition fixes only the numbering in anti-clockwise order and there are still three possibilities to define the vectors. We here choose $\mathbf{a}_1$ to correspond to the path 1–6, cf. Fig. 6.3**a**, which fixes $\mathbf{a}_2$ to 1–4 (by repeating the hexagon periodically) and $\mathbf{a}_3$ to 1–2. Further, the set of displacement vectors between next-nearest neighbors is defined by $\mathbf{b}_1 = \mathbf{a}_2 - \mathbf{a}_3 = (0,1)^T$ (4–6 or 3–1), $\mathbf{b}_2 = \mathbf{a}_3 - \mathbf{a}_1 = (-1,0)^T$ (6–2 or 5–3) and $\mathbf{b}_3 = \mathbf{a}_1 - \mathbf{a}_2 = (1,-1)^T$ (2–4 or 1–5). In our convention, using these $\mathbf{b}$ vectors, the A sites will obtain a phase $-\phi$ and B sites $+\phi$. The Bloch Hamiltonian can then be expressed as

$$H(k) = \begin{pmatrix} 2t_2 \sum_i \cos(-\mathbf{k}\cdot\mathbf{b}_i - \phi) + m & t_1 \sum_i e^{-i\mathbf{k}\cdot\mathbf{a}_i} \\ t_1 \sum_i e^{i\mathbf{k}\cdot\mathbf{a}_i} & 2t_2 \sum_i \cos(-\mathbf{k}\cdot\mathbf{b}_i + \phi) - m \end{pmatrix} \tag{6.36}$$

and with $\cos(a \pm b) = \cos(a)\cos(b) \mp \sin(a)\sin(b)$,

$$\begin{aligned} H(k) = 2t_2\cos(\phi)\sum_i \cos(\mathbf{k}\cdot\mathbf{b}_i)\mathrm{Id} + t_1\left[\sum_i \cos(\mathbf{k}\cdot\mathbf{a}_i)\sigma_1 + \sum_i \sin(\mathbf{k}\cdot\mathbf{a}_i)\sigma_2\right] \\ + \left[m - 2t_2\sin(\phi)\sum_i \sin(\mathbf{k}\cdot\mathbf{b}_i)\right]\sigma_3. \end{aligned} \tag{6.37}$$

This is identical to the Hamiltonian in Haldane's paper and in this form free of a specific choice of lattice vectors or coordinates.

By using the coordinates of neighboring unit cells as shown in Fig. 6.3**b** we can expand the sums and obtain

$$H(k) = \begin{pmatrix} H_{AA} & H_{AB} \\ H_{BA} & H_{BB} \end{pmatrix}, \tag{6.38}$$

where

$$\begin{aligned} H_{AA} &= 2t_2\left[\cos(-k_2 - \phi) + \cos(k_1 - \phi) + \cos(-k_1 + k_2 - \phi)\right] + m, \\ H_{AB} &= t_1\left[1 + e^{-ik_1} + e^{-ik_2}\right], \\ H_{BA} &= t_1\left[1 + e^{ik_1} + e^{ik_2}\right], \\ H_{BB} &= 2t_2\left[\cos(-k_2 + \phi) + \cos(k_1 + \phi) + \cos(-k_1 + k_2 + \phi)\right] - m, \end{aligned} \tag{6.39}$$

and k_1, k_2 are the coordinates in terms of reciprocal lattice vectors times 2π.

Eq. 6.39 is convenient for a numerical implementation, however, on paper a form in terms of Pauli matrices, like Eq. 6.37, is much simpler. We proceed by computing the eigenvalues of a generic $H(k) = a(k)\mathrm{Id} + \mathbf{h}(k)\cdot\boldsymbol{\sigma}$:

$$\det\left[\begin{pmatrix} a(k) - \lambda + h_3(k) & h_1(k) - ih_2(k) \\ h_1(k) + ih_2(k) & a(k) - \lambda - h_3(k) \end{pmatrix}\right] \tag{6.40}$$

$$= (a(k) - \lambda)^2 - h_3(k)^2 - (h_1(k)^2 + h_2(k)^2) \stackrel{!}{=} 0 \tag{6.41}$$

$$\Rightarrow \lambda_\pm = a(k) \pm \sqrt{\mathbf{h}(k)^2}. \tag{6.42}$$

Apparently, these have a rather simple form for the non-topological honeycomb lattice of Eq. 6.30

$$\varepsilon_{1,2}(k) = \lambda_\pm = \pm t\sqrt{(1 + \cos(k_1) + \cos(k_2))^2 + (\sin(k_1) + \sin(k_2))^2}. \tag{6.43}$$

The two bands touch if $h_1(k) = h_2(k) = 0$, which is only satisfied for $k = K, K'$ as we can see from

$$\begin{aligned} 1 + \cos(k_1) + \cos(k_2) &\stackrel{!}{=} 0 \\ \sin(k_1) + \sin(k_2) &\stackrel{!}{=} 0. \end{aligned} \tag{6.44}$$

The trivial solution to the second equation $k_1 = -k_2$ leads to $\cos(k_1) = \frac{1}{2}$ and therefore $k_1 = \frac{2\pi}{3}$. Then, $k_2 = -\frac{2\pi}{3}$ or modulo 2π: $k_2 = \frac{4\pi}{3}$, which corresponds to K'. The second solution is obtained by choosing $\cos(k_2) = \frac{1}{2}$ and is given by K. At these two points there is clearly no band gap, i.e., Hamiltonians of this form always describe (semi-) metals. A band gap can only be opened by the presence of a finite term $h_3(k)$. In particular, $h_3(K) = m$ opens a band gap of width $2m$. For the Haldane model we observe in Eq. 6.37 that both the mass term m and the next-nearest neighbor hoppings add to $h_3(k)$, i.e., both of these can open a band gap. It turns

out that only $m = t_2 = 0$ and specific combinations of the two leave the band gap closed. This can be seen by requiring that either $h_3(K) = 0$

$$\begin{aligned} 0 &\stackrel{!}{=} m - 2t_2 \sin(\phi) \sum_i \sin(\mathbf{K} \cdot \mathbf{b}_i) && (6.45)\\ &= m - 2t_2 \sin(\phi) \left[\sin(2\pi/3) + \sin(-4\pi/3) + \sin(2\pi/3)\right] && (6.46)\\ &= m - t_2 \sin(\phi) 3\sqrt{3} && (6.47) \end{aligned}$$

or $h_3(K') = 0$

$$\begin{aligned} 0 &\stackrel{!}{=} m - 2t_2 \sin(\phi) \sum_i \sin(\mathbf{K}' \cdot \mathbf{b}_i) && (6.48)\\ &= m - 2t_2 \sin(\phi) \left[\sin(4\pi/3) + \sin(-2\pi/3) + \sin(-2\pi/3)\right] && (6.49)\\ &= m + t_2 \sin(\phi) 3\sqrt{3}. && (6.50) \end{aligned}$$

Hence, the t_2–m phase diagram will show two metallic lines defined by

$$\begin{aligned} m &= 3\sqrt{3} \sin(\phi) t_2, \\ m &= -3\sqrt{3} \sin(\phi) t_2, \end{aligned} \qquad (6.51)$$

that intersect at $m = t_2 = 0$. From this we know straight away that since topological transitions must be accompanied by a gap closing, the Haldane model of Eq. 6.35 can only feature at most four topological phases separated by the lines defined through Eq. 6.51, and the slope will be controlled by the phase ϕ. Apparently, if $\phi = n\pi$ with $n \in \mathbb{N}^0$ there will only be two insulating phases separated by a metallic line at $m = 0$. The inverse case, i.e., a separation line defined by $t_2 = 0$, cannot happen since $\sin(\phi) \in [-1, 1]$ and therefore the maximal slope is $3\sqrt{3}$ obtained at $\phi = (n + 1/2)\pi$. In the ϕ–m diagram, as shown in Haldane's paper [143], Eq. 6.51 has a different interpretation and we have instead two overlapping sin functions, where we can again define four distinct regions.

Since the Haldane model lacks time-reversal and particle-hole symmetry the equivalence classes are revealed by a $\mathbb{Z}$ topological index in 2 dimensions given by the Chern number. Following this analysis it is now enough to compute the Chern number only for one specific value of ϕ to obtain the complete classification for arbitrary values.

We compute the Chern number numerically by diagonalizing the Hamiltonian and evaluating the integral over the Berry curvature. The result as a function of m, t_2, ϕ is shown in Fig. 6.4. In Fig. 6.4**a** we look at next-nearest neighbor hopping amplitude t_2 vs. Semenoff mass m. The solid lines are the analytical result from Eq. 6.51 for the phase transition. Apparently, although topologically four distinct phases are possible, only three different Chern numbers appear. This is clear, though, since at $t_2 = 0$ the model is trivial regardless of the sign of m. Transitions happen from $C = \pm 1$ to $C = 0$ anywhere along the lines and from $C = 0$ to $C = 0$ and $C = -1$ to $C = 1$ at $t_2 = m = 0$. Positive t_2 produces $C = 1$ in a certain range of m and negative t_2 produces $C = -1$. The entire diagram is symmetric w.r.t. $m = 0$ since we do not differentiate the sublattices apart from the different sign in the sublattice energy. Replacing $m \to -m$ therefore only exchanges $A \leftrightarrow B$.

In Fig. 6.4**b** we show the Chern number as a function of the flux ϕ and m. Again, the lines are obtained from Eq. 6.51. The $C = 1$ phase appears only for $\phi \in (0, \pi)$ and $C = -1$ for $\phi \in (\pi, 2\pi)$. At $\phi = \pi$ the next-nearest neighbor hopping becomes real again, which leads to a vanishing of the h_3 matrix element that closes the gap. The largest range of m values that produce a topologically non-trivial phase can be found at $\phi = \pm\pi/2$ for $C = \pm 1$, respectively. It is this value that we will later restrict to as a proxy for all other combinations of parameters.

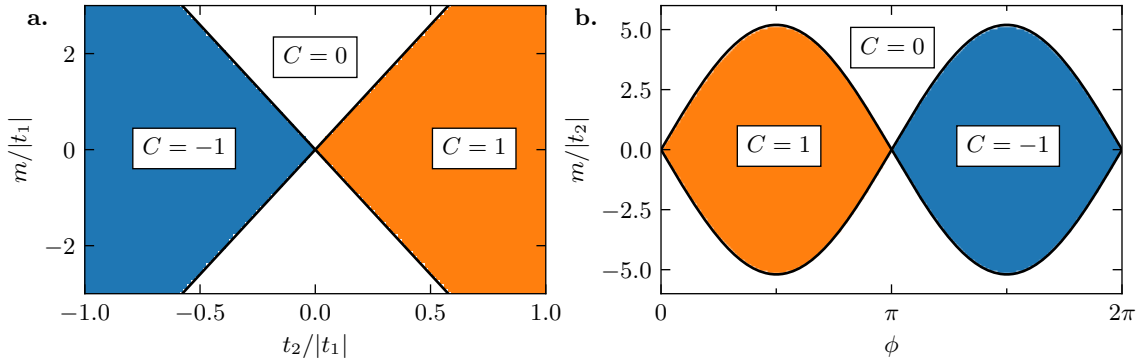

Figure 6.4: Phase diagram of the Haldane model. **a.** Next-nearest neighbor hopping amplitude t_2 vs. Semenoff mass m. **b.** Flux ϕ vs. m. There are three topologically distinct phases with different Chern numbers (color code and insets). The analytical expression of Eq. 6.51 for the gap closure is shown as black lines. [Reproduction of Fig. 4 from Ref. [173] and Fig. 2 from Ref. [143]]

6.3 Haldane-Hubbard Model

Before diving into the actual systematic analysis we briefly summarize the status quo regarding the phase diagram of the Haldane-Hubbard model. Essentially, we study here

$$H = \sum_k (c_{kA}^\dagger, c_{kB}^\dagger) h(k) (c_{kA}, c_{kB})^T + U \sum_i n_{i\uparrow} n_{i\downarrow}, \tag{6.52}$$

where $h(k)$ is the Bloch Hamiltonian of the Haldane model, cf. Eq. 6.39. U is the strength of the local Hubbard interaction that we assume here to be repulsive, i.e., $U > 0$ and that penalizes the double occupation of sites. We notice straight away that with the Semenoff mass m and this repulsive onsite interaction we are facing a variant of the ionic Hubbard model that we studied in Chapter 5, here on a honeycomb lattice and with complex next-nearest neighbor hoppings. This analogy offers the immediate insight that the momentum-dependence of the self-energy will be unimportant for large m and U and specifically for $m > U/2$. Therefore, also the Chern number is expected to be impervious to the local approximation in this broad region of phase space.

Noting the previous realization that the flux ϕ essentially broadens the width of the topological region up to a value of $\phi = \pi/2$ we restrict here to a specific parameter set given by $t = t_1, t_2/t = 0.2$ and $\phi = \pi/2$ at half filling, for which plenty of data can be found in the literature. As seen before, other values of ϕ will only shrink the topological region or invert the sign of the Chern number. The same applies for t_2. Both are effects that are unimportant for the type of universal conclusion that we strive to obtain.

The currently accepted version of the phase diagram is shown in Fig. 6.5, where results from DMFT, static mean-field theory (MF), ED [140], dynamical cluster approximation (DCA) [174] and Bold Diagrammatic Monte Carlo (BDMC) [175] are shown in addition to our own phase transition line obtained with self-energy data from TPSC. All data were produced at very low temperatures: ED calculations at $T = 0$, DMFT at $T = 0$ [140] and $T/t = 0.1$ by ourselves, BDMC at $T/t = 0.1$ and TPSC at $T/t = 0.1$. Our DMFT data was found to agree very well with that from Ref. [140].

Four different phases are found, namely the two phases known from the non-interacting model: a Chern insulator (CI) with Chern number $C = 2$ (doubled due to the two spins) and the trivial band insulator (BI). With finite electron-electron interactions an additional symmetry-broken topological insulator (SBTI) with Chern number $C = 1$ appears, where clearly the SU(2)

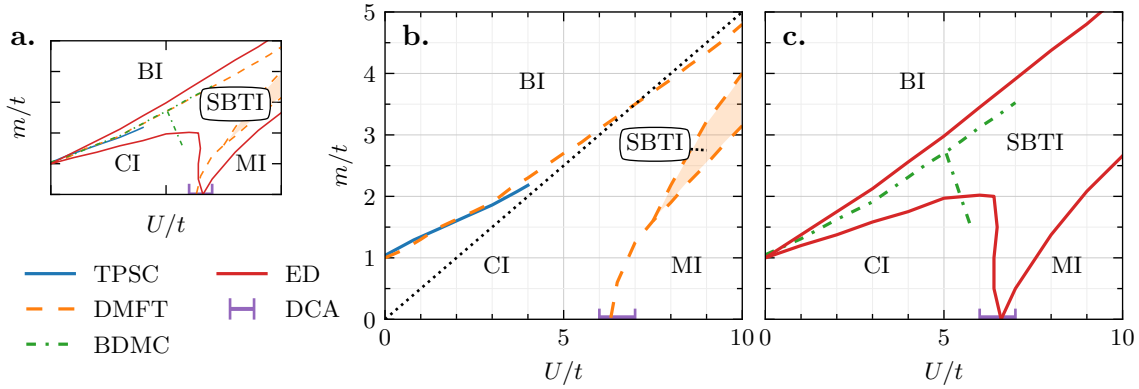

Figure 6.5: Phase diagram of the half filled Haldane-Hubbard model at $t_1 = t, t_2/t = 0.2, \phi = \pi/2$. **a.** Complete phase diagram showing all methods. For improved legibility we show larger versions with subsets of methods in **b.** and **c.**. The methods agree well regarding the CI–BI and CI–MI transitions, only the transition from CI to SBTI is inconsistent between ED, DMFT and BDMC. The dotted black line corresponds to $m/U = 1/2$. Our analysis in Chapter 5 predicts that DMFT is very reliable at least above this line. For DCA we only know that the location of the CI–MI transition for $m = 0$ is at $U \approx 6 - 7$. [Figure adapted from Ref. [169] using data from [140, 174, 175]]

symmetry of the model is spontaneously broken. At large U/t a Mott insulator (MI) appears. Apparently, all methods agree rather well as seen in the overview in Fig. 6.5**a**. For clarity we separate the different methods into subfigures **b** and **c**. The prediction for the continuation of the topological CI–BI transition, that is located at about $m/t \approx 1.04$ for $U = 0$, into the correlated regime agrees remarkably well between all methods for which this data was available. In particular, TPSC, DMFT and BDMC obtain an almost identical line. As for the transition from the Chern insulator to the Mott insulator we observe qualitative agreement. Here, also DCA data was available up to $t_2/t = 0.15$ [174], which we extrapolated to $t_2/t = 0.2$ and marked the corresponding point in the plot with an error bar.

The location of the transition from CI to SBTI on the other hand is highly contradictory. In DMFT this phase appears only at very large U/t, where the critical U depends strongly on the value of m. In contrast, in ED this phase is predicted even for small values of U around the transition from CI to BI. BDMC predicts yet another result that lies somewhat in between the other two, where the SBTI phase exists only beyond $U/t = 5$ and the critical value is only weakly dependent on m. In addition, the slope of the transition line has the opposite sign as for DMFT.

Clearly, we cannot put the same level of trust in all methods and we know that DMFT contains the error of the local approximation, ED suffers from finite size effects and TPSC is only reliable for intermediate U. BDMC is in principle an exact method, however, the result converges only as a function of the sample size which can be slowed down severely by the sign problem. In addition, this method uses a self-consistency cycle, i.e., the solution is obtained as the fixed point of some sequence, which requires convergence on another level—in addition to the statistical error—that is not guaranteed. Therefore, while we expect the BDMC result to be most reliable we cannot say with absolute certainty that this is an exact result.

In order to rule out differences based solely on different simulation protocols we performed our own DMFT calculations using the protocol explained in Ref. [175]. This means we first perform a DMFT calculation with explicitly broken S_z symmetry, i.e., $t_{1,2,\uparrow} \to t_{1,2,\uparrow}/\delta$ and $t_{1,2,\downarrow} \to t_{1,2,\downarrow}\delta$. For the actual calculation we used a value of $\delta/t = 1.1$. With this explicit symmetry-breaking the $C = 1$ phase exists even at $U = 0$. We then initialize a calculation for the symmetric model using the solution that lacks S_z symmetry. Since the self-consistent

problem defined by the DMFT equations is solved iteratively, the initial value can have a great influence on the outcome. The idea here is that a symmetry-broken initial value should be more likely to produce a $C = 1$ phase than the standard symmetric choice. However, even when using this scheme we only confirmed the result by Vanhala et al. [140], indicating that the difference to the BDMC result must be a consequence of the approximation and not different implementations of the self-consistent scheme.

Looking at the sources of errors, the differences between the methods seem to stem mostly from different handling of non-local contributions. DMFT completely lacks these, ED captures only short-ranged contributions, while BDMC in principle contains also long-ranged correlation effects. Since the model is basically just a variant of the ionic Hubbard model we can apply the findings of Chapter 5. In principle, the DMFT error should vanish in the non-dispersive regime, i.e., above roughly $m = U/2$, which is our best guess lacking any further data for the double occupancy and densities. The CI $\rightarrow$ BI transition lies above or in the region of $m \approx U/2$, which explains the good agreement between all methods regardless of their different handling of non-local effects: the self-energy is local. The phase under debate is located below this line, though, which raises doubts about the quality of the DMFT approximation.

The idea is now to investigate how non-local correlation effects affect the topological classification and how far (if at all) we can trust the local approximation, given that the Chern number is a fundamentally non-local measure.

6.4 Statistical Method

In order to investigate the effect of different local and non-local contributions to the self-energy on the topological classification via the topological Hamiltonian

$$h_t(k) = H_0(k) + \Sigma(\omega = 0, k), \tag{6.53}$$

we perform an analysis of the behavior of the Chern number as a function of different perturbations. To this end we make explicit the two terms

$$\Sigma(\omega, k) = \Sigma_{\mathrm{loc}}(\omega) + \Sigma_{\mathrm{non-loc}}(\omega, k), \tag{6.54}$$

where we assume that the momentum average of $\Sigma_{\mathrm{non-loc}}(\omega, k)$ vanishes. We have already shown in Chapter 5 and Ref. [135] that this decomposition is unique and well-defined. In this form, all corrections to the local self-energy that are a consequence of non-local contributions to the self-energy are already absorbed in Σ_{loc}.

6.4.1 Local Self-Energy

We begin the discussion by taking a closer look at the first term in Eq. 6.54, i.e., the local self-energy. $\Sigma_{\mathrm{loc}}(\omega)$ is just a constant matrix that does not depend on momentum at all. In the context of topology, only the zero-frequency value matters, so that a description in terms of a 2×2 matrix is complete. Note that we restrict to paramagnetic self-energies here, where $\Sigma_{\uparrow\uparrow} = \Sigma_{\downarrow\downarrow}$ and $\Sigma_{\uparrow\downarrow} = \Sigma_{\downarrow\uparrow} = 0$. Therefore, we can use a formulation in terms of a spinless model. We will discuss the magnetic case later. To lowest order in U we can express the self-energy as

$$\Sigma_{\mathrm{loc}}(\omega = 0) = \Sigma^{\mathrm{MF}} = \frac{U}{2} n, \tag{6.55}$$

where MF refers to the static mean-field or Hartree approximation. An insightful way to obtain this is not via a diagrammatic expansion but rather an expansion in fluctuations on the operator level. Decomposing the density operator as

$$n = \langle n \rangle + \delta n, \tag{6.56}$$

where δn is an operator defined through Eq. 6.56, the Hubbard interaction operator can be written in the form

$$H_U = U \sum_i n_{i\uparrow} n_{i\downarrow} = U \sum_i [\langle n_{i\uparrow}\rangle + \delta n_{i\uparrow}][\langle n_{i\downarrow}\rangle + \delta n_{i\downarrow}] \tag{6.57}$$

$$= U \sum_i [\langle n_{i\uparrow}\rangle\langle n_{i\downarrow}\rangle + \langle n_{i\downarrow}\rangle \delta n_{i\uparrow} + \langle n_{i\uparrow}\rangle \delta n_{i\downarrow} + \delta n_{i\uparrow}\delta n_{i\downarrow}]\,. \tag{6.58}$$

To first order in the fluctuations δn this is

$$H_U = U \sum_i \left[-\langle n_{i\uparrow}\rangle\langle n_{i\downarrow}\rangle + \langle n_{i\downarrow}\rangle n_{i\uparrow} + \langle n_{i\uparrow}\rangle n_{i\downarrow} + \mathcal{O}\big((\delta n)^2\big)\right], \tag{6.59}$$

where the first term is just a constant energy shift and the second and third terms are now single particle operators. Dropping the second order term we can therefore write the Hubbard Hamiltonian as

$$H = H_0 + U \sum_i \big(\langle n_{i\downarrow}\rangle n_{i\uparrow} + \langle n_{i\uparrow}\rangle n_{i\downarrow}\big). \tag{6.60}$$

If we assume that we are in a paramagnetic phase we can set $\langle n_{i\uparrow/\downarrow}\rangle = \frac{1}{2}\langle n_i\rangle$ and with that finally obtain

$$H = H_0 + \frac{U}{2} \sum_i \langle n_i\rangle n_i. \tag{6.61}$$

Given that the Green's function for this single-particle operator is easily obtained as

$$G_{ii}(\omega, k) = \frac{1}{\omega + \mu - H(k) - \frac{U}{2} n_i}, \tag{6.62}$$

we can indeed read off the expression from Eq. 6.55. Incidentally, for a bipartite lattice as the one at hand, the local density is a function of the sublattice index A, B and therefore alternates between the two sublattices. Without loss of generality we can write

$$\Sigma^{\mathrm{MF}} = \frac{U}{2} \Delta n\, \sigma_3 + \text{const.}, \tag{6.63}$$

where $\Delta n = (n_A - n_B)/2$ is half the difference between sublattice densities and σ_3 is the third Pauli matrix. It is straight-forward to show that this expression is indeed correct

$$\Sigma^{\mathrm{MF}} = \frac{U}{2} \begin{pmatrix} n_A & 0 \\ 0 & n_B \end{pmatrix} = \frac{U}{2} \begin{pmatrix} \frac{n_A+n_B}{2} + \frac{n_A-n_B}{2} & 0 \\ 0 & \frac{n_A+n_B}{2} - \frac{n_A-n_B}{2} \end{pmatrix} \tag{6.64}$$

$$= \frac{U}{2} \left[\frac{n_A + n_B}{2} \mathrm{Id} + \frac{n_A - n_B}{2} \sigma_3 \right]. \tag{6.65}$$

The constant term is simply an energy shift in the topological Hamiltonian that is absorbed in the chemical potential, i.e., has no effect on the Chern number. The second term proportional to σ_3, however, has the same form as the mass term and therefore correlates directly with a tunable parameter of the non-interacting Hamiltonian. We can immediately predict that in the Hartree approximation, but also possibly with a renormalized amplitude in the general case, the Hubbard interaction leads to a *reduction* of the mass term in the topological Hamiltonian and therefore to a shift of the topological phase transition towards larger values of m with increasing U.

From this simple analysis we can already understand the rough appearance of the phase diagram. Clearly, due to the choice of $m \geq 0$ we have $n_B \geq n_A$ and therefore the term in

Eq. 6.65 proportional to σ_3 is negative. Hence, $m \to m + U\frac{n_A - n_B}{4} \leq m$. In addition, we know from our previous study of the generic ionic Hubbard model that increasing U at finite m reduces the difference of local densities to first order according to $|n_A - n_B| \sim 1/U$. Therefore, in the topological Hamiltonian, constant values of the effective m are shifted upwards along the axis of the actual m, following a straight line with slope $\sim \frac{n_B - n_A}{4}$ as a function of U. As a consequence the phase transition, originally at $m/t \approx 1$, is shifted towards higher values on a straight line for $U > 0$. This agrees very well with the observations made by studying the numerical results of Fig. 6.5. We expect all topological models that are variants of the ionic Hubbard model to share this feature (not necessarily the near constant slope, but the upwards direction of the phase transition).

The derivation above is valid in the Hartree approximation only. However, the general case can be treated in a similar fashion by noting that due to a corresponding symmetry of H the self-energy satisfies

$$\Sigma_{AA}(\omega = 0, k) = -\Sigma_{BB}(\omega = 0, k) \tag{6.66}$$

up to a constant term. Any constant does not affect the eigenvalues of the topological Hamiltonian and can therefore be neglected as it facilitates just a shift in the chemical potential. For $m = 0$ we have a sublattice symmetry (inversion symmetry) and therefore $\Sigma_{AA} = \Sigma_{BB}$. Any finite value $m \neq 0$, however, breaks this symmetry and therefore we can write in analogy to the Hartree derivation

$$\Sigma_{AA/BB} = \frac{1}{2}\left[\Sigma_{AA} + \Sigma_{BB} \pm (\Sigma_{AA} - \Sigma_{BB})\right], \tag{6.67}$$

where the first term is just the unimportant constant. For the local self-energy as a whole we define

$$\Sigma_{\mathrm{loc}}(\omega = 0) = a\sigma_1 + b\sigma_2 + \delta\Sigma\sigma_3 = \begin{pmatrix} \delta\Sigma & a - ib \\ a + ib & -\delta\Sigma \end{pmatrix}, \tag{6.68}$$

where $a, b \in \mathbb{R}$ and

$$\delta\Sigma = \frac{\Sigma_{AA}(\omega = 0) - \Sigma_{BB}(\omega = 0)}{2} \leq 0. \tag{6.69}$$

Eq. 6.68 can be regarded as two terms (diagonal and off-diagonal) that can modify the topological phase diagram as a function of U. As motivated in terms of the Hartree result the diagonal part of the self-energy (with and without corrections that go beyond first order) is directly proportional to σ_3, which means that it results in a simple shift of the mass term

$$m \mapsto m + \delta\Sigma, \tag{6.70}$$

where m becomes smaller as a result of $\delta\Sigma \leq 0$. This behavior can be found all over the literature for the Haldane model [140, 175] but also for the Bernevig-Hughes-Zhang model [138, 139] or the time-reversal symmetric Hofstadter-Hubbard model [137, 141]. Incidentally, the latter two have a vanishing Chern number due to the presence of time-reversal symmetry and the topological properties are characterized by a $\mathbb{Z}_2$ invariant. Since this argument only makes use of a mapping between the topological Hamiltonian and specific non-interacting Hamiltonians it does not depend at all on the specifics of the topological invariant to be computed.

The remainder of the local self-energy are off-diagonal terms proportional to σ_1 and σ_2. For simplicity we discuss both at the same time. Comparing with Eq. 6.37 we find that the situation is not as clear, since the Haldane Hamiltonian does not contain a parameter that corresponds to a constant off-diagonal term. Within our convention we neglect phases corresponding to displacements within the unit cell in the Fourier transform. Therefore, a constant term $t_1\sigma_1$ does appear, however, t_1 appears also in other non-constant terms and therefore cannot be tuned independently. Therefore, we introduce the parameters a, b from Eq. 6.68 as additional parameters to the model and vary their values to investigate the response of the Chern number.

For example, the case $a = -t_1, b = 0$ takes out one nearest-neighbor hopping and therefore corresponds to one-dimensional zigzag chains that are coupled via the next-nearest neighbor hopping. With $a \gg 0$ on the other hand, the case of coupled dimers is realized. We illustrate these two cases in Fig. 6.6, where we also show a diagram of the phases obtained. The trivial phase is robust with respect to a to a degree where it does not change at all. The non-trivial topological phase with positive Chern number is relatively stable for $a/t \in [-1, 1)$ and vanishes for any values beyond. Interestingly, at $a/t < -1$ a non-trivial phase with negative Chern number appears, which has also been found for a similar limit in the Hofstadter model [176]. We note that any finite value of a breaks the residual 3-fold rotational symmetry that the model retains with finite m. The topological phase itself is not bound to this symmetry as we can see from the stability of C over a rather large phase space region.

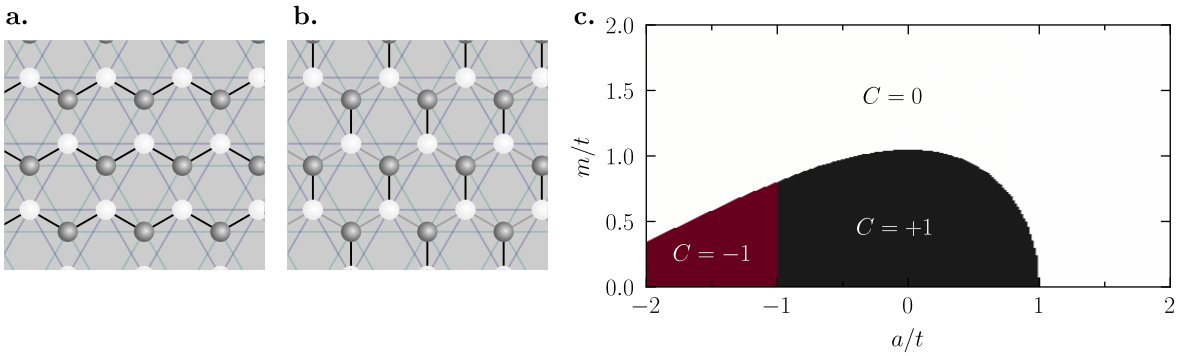

Figure 6.6: Two limits achieved by tuning the parameter a from Eq. 6.68 ($b = 0$). **a.** Coupled zigzag chains ($a = -t$) and **b.** coupled dimers ($a \gg t$). Next-nearest neighbor hoppings are present in both cases. **c.** Phase diagram as a function of a and m. The limit **a.** lies on the phase transition and is therefore metallic for small m/t. The dimer limit is deep in the trivial phase region. negative coupling between two sites changes the sign of the Chern number within a small region.

The same value of the topological index for negative a, cf. Fig. 6.6**c**, as in the $\phi < 0$ case indicates that the two phases are the same and therefore a smooth connection between the two Hamiltonians exists. We note that in order to reconnect to the Haldane model we need to reintroduce the symmetry between nearest neighbor hoppings t_1 and change the phase of the next-nearest neighbor hoppings. Both paths cross a metallic line. What sounds like a contradiction is actually resolved by introducing additional parameters. While it is true that the connection cannot be made within the limited set of parameters offered by the Haldane model, the topological index guarantees that this is possible in general. The same argument applies to the phase diagram of Fig. 6.4, where the connection between the trivial phase at $m > 0$ to the trivial phase at $m < 0$ is obstructed by a metallic phase at $m = 0$. An additional parameter is needed to lift the degeneracy at $m = 0$. This can always be achieved by adding large constant parameters (no momentum-dependence) to the Hamiltonian, since given that the energy spectrum is defined as $\epsilon_k = a(k) \pm |\mathbf{h}(k)|$, as shown in Eq. 6.42, a degeneracy occurs only if $|\mathbf{h}(k) = 0|$ for some k. We can always add large enough parameters to $\mathbf{h}(k)$ to remove these roots, which lifts the degeneracy.

At this point we emphasize that we did not take into account the physicality of the values of a when computing the diagram of Fig. 6.6. Upon comparison with the self-energy provided by TPSC the value corresponding to a, i.e., the off-diagonal, is always positive, which indicates that a phase transition through negative a is highly unlikely to happen as a result of onsite electron-electron interactions. Note that DMFT cannot provide an off-diagonal term, since only a single site problem is solved analytically. In order to obtain more than the diagonal matrix elements of Σ one would have to use cellular DMFT with a two-site cluster.

In general, the off-diagonal term of the local self-energy is a complex number. To accommodate for this we compute the phase diagram for this general case, allowing for different values also in the imaginary part of the perturbation, i.e., $\Sigma^{AB}_{\text{loc}} = a + ib$. The result is shown in Fig. 6.7[1]. We observe a symmetry w.r.t. the sign of b, while the sign of a has a big impact on the topological phase. As seen before, the negative Chern number phase appears only for negative a. Now, we find that this phase is also susceptible to an imaginary part b, since a transition to first the usual $C = 1$ phase and then the trivial phase appears at $m = 0$. At finite m the two non-trivial phases start to separate so that a direct transition is only happening at $b = 0$ for $a/t = -1$. This point remains the location of a phase transition until $m/t \approx 0.8$, where the $C = 1$ phase starts to shrink to zero. Above the value $m/t \approx 1.04$, where the phase transition of the Haldane model lies, no topological phase is found irrespective of the value of a, b.

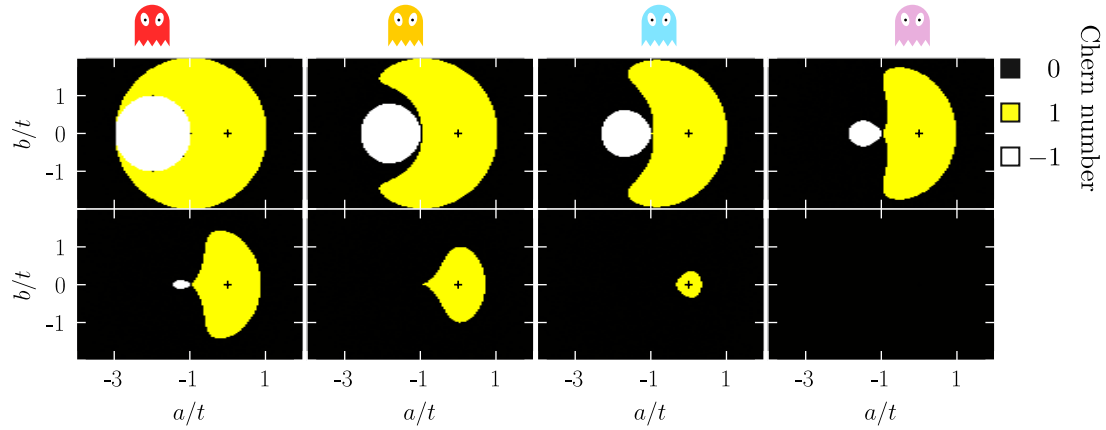

Figure 6.7: Phase diagram as a function of a, b for $m/t = 0, 0.1, 0.2, 0.4$ (top row, left to right, and $m/t = 0.6, 0.8, 1, 1.2$ (bottom row, left to right). The origin is marked with "+". We find the Haldane phase with $C = 1$ (yellow), the trivial phase $C = 0$ (black) and a $C = -1$ phase (white) at negative a. The topological phases are more stable towards negative real part a and are symmetric w.r.t. b. With increasing m the topological phases begin to vanish. The larger $C = +1$ phase region remains approximately until the phase transition of the Haldane model at $m/t \approx 1$.

In order to obtain a more general understanding we now employ a statistical approach where instead of using fixed grids over multiple parameters we sample these parameters from a random distribution. In the limit of large numbers of samples the entire phase space is covered and therefore the same result is obtained. The benefits of such methods are twofold: (i) it is computationally more efficient since, in general, a smaller number of points needs to be computed to cover a large space, (ii) the statistical mean provides a reduced view on the data that is much easier to evaluate and interpret. This corresponds to integrated quantities for grid-based methods.

In the following we hold m fixed and sample over a and b from a uniform distribution. In order to quantify the effect of the off-diagonal contributions we define the amplitude

$$z = \left|\Sigma^{AB}_{\text{loc}}\right| = |a + ib|. \tag{6.71}$$

For the distribution we fix not a, b to specific intervals but rather the phase, so that z is a controlled parameter. In exponential notation we have $\Sigma^{AB}_{\text{loc}} = ze^{i\alpha}$, where α is drawn from a uniform distribution $U([0, 2\pi))$. This does not correspond to a uniform distribution of a, b, since the number of samples is constant as a function of z. Given that the length of the line segment

[1] We use the similarity of some of this data to Pac-Man to honor the classic video game that inspired so many to develop an interest in computers and all kinds of technology.

covered by the sampling procedure is the circumference of the circle with radius z, i.e., $2\pi z$, the density of samples is given by $\rho(z) = n_{\text{samples}}/(2\pi z)$. Therefore, the probability density falls off as $1/z$. This is done for technical convenience, the result itself does not depend on the exact distribution as long as it is uniform in z, since the evaluation of the sample mean is carried out not over the entire phase space but per z value.

During the sampling procedure we collect a large amount of data that needs to be reduced to some interpretable quantity. As a first indicator we use the average Chern number

$$\langle C \rangle_{\Sigma_{\text{loc}}} = \frac{1}{n_{\text{samples}}} \sum_{m=1}^{n_{\text{samples}}} C(h_t^m), \tag{6.72}$$

that is just defined as the arithmetic mean of the Chern number values obtained from the topological Hamiltonian $h_t^m = H_0 + \Sigma_{\text{loc}}^{AB,m}$ for various samples of Σ_{loc}^{AB}.

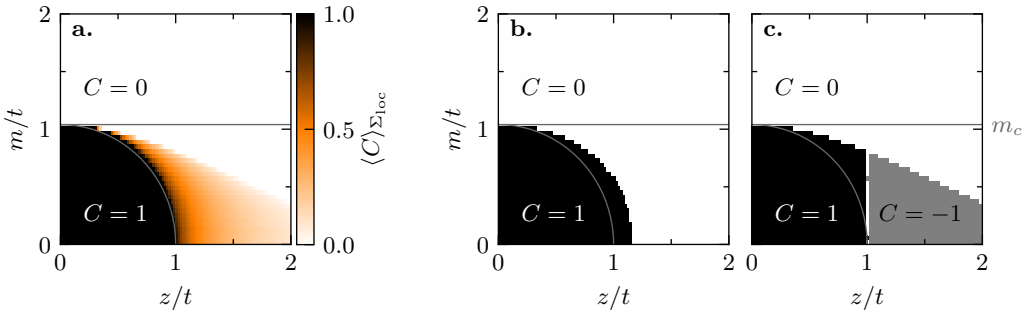

Figure 6.8: a. Average Chern number computed by sampling over the local self-energy Σ_{loc}^{AB}. In **b.** and **c.** we show for comparison snapshots where the phase $\alpha = \arctan(b/a)$ is fixed to $\alpha = 0.25\pi$ and $\alpha = \pi$, respectively. There is a core region that is almost circular with radius t, where the Chern number is stable w.r.t. any local perturbations. This region is marked with a gray line. The second (horizontal) gray line marks the transition at $z = 0$. Below $m/t = 1.04$ and to the right of the stable region we observe an unstable regime (shaded orange), where nonzero Chern numbers may appear depending on the value of α. The snapshots in **b.**,**c.** illustrate the origin of the shaded region. [Subfigure **a.** adapted from Ref. [169]]

The data obtained is shown in Fig. 6.8, where we observe a core region with $(z^2+m^2)/t^2 \lesssim 1$, where the average Chern number is equal to 1. Clearly, this value can only occur if *all* samples satisfy $C_m = 1$ individually, which means that the Chern number of the non-trivial phase is stable w.r.t. off-diagonal perturbations within a radius of 1 around the Haldane model. The topologically trivial phase of the Haldane model ($z = 0$) is found above $m_c \approx 1.04t$ (here indicated by a straight line), where $\langle C \rangle = 0$ for all values of z. This can be understood from our earlier analytical consideration of the energy eigenvalues (Eq. 6.42). Since $m > m_c$ already lifts the degeneracy of the eigenvalues at the K, K' points, the spectrum will be gapped for any choice of values for a, b or $z = \sqrt{a^2 + b^2}$. Therefore, a topological phase transition above m_c is impossible. Below, however, moving outside of the stable region we find a shaded region with $0 < \langle C \rangle < 1$, where clearly different values of C appear in the sample set. In comparison to the Haldane model a topological transition is at least possible here. Going beyond this shaded region towards larger z we find the trivial phase also below m_c. Here, also $\langle C \rangle = 0$, which seems to suggests that a topological transition at this m, z is certain. Upon comparison with snapshots of the Chern number at fixed phases α (i.e., no sampling necessary) shown in Fig. 6.8**b,c**, the origin of the stable and shaded regions becomes clear as we recognize the data shown in Fig. 6.7.

At $z = 0$ the average Chern number is a good choice also as a statistical indicator, since the Chern number for the usual Haldane model at $\phi = \pi/2$ is two-valued $C \in \{0, 1\}$. Moreover, $C = C^2$ and by identifying the random variable X with the Chern number we conclude

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = \mathrm{E}[X] - \mathrm{E}[X]^2. \tag{6.73}$$

Hence, the variance is defined entirely by the mean, or equivalently, the mean contains also information about the variance, which makes it a suitable choice for a stochastic error measure. On the other hand we know that C can also assume a negative value, which not only invalidates Eq. 6.73 but allows for an accidental averaging of different values to zero.

Although this does not happen in this case, where the averaged signal can still be differentiated from zero, we propose a different statistical measure in the form of a "probability of change":

$$P(C \neq C_{\mathrm{ref}}) = \langle \min\{1, |C - C_{\mathrm{ref}}|\}\rangle. \tag{6.74}$$

By construction, $P(C \neq C_{\mathrm{ref}}) \in [0, 1]$ and $P = 0$ implies that all samples of C are equal to C_{ref}. Conversely, $P = 1$ can only occur if $C \neq C_{\mathrm{ref}}$ for all C in the sample set. Formally, this definition can be related to the distance between two probability distributions of random variables X, Y

$$d(X, Y) = \mathrm{E}_{X,Y}[|X - Y|], \tag{6.75}$$

where we only added a normalization constraint that ensures that $|X-Y| \leq 1$, which is important for an interpretation as a probability. In contrast to the average Chern number of Eq. 6.72, the probability of change takes into account changes for all samples, i.e., an averaging out of a subset of samples with opposite sign is no longer possible, thereby removing the risk of misinterpretation.

Given the large memory or storage requirements when keeping all Chern numbers that amount to $n_z \times n_m \times n_{\mathrm{samples}} \times 1\mathrm{Byte}$ (using 8-bit integers), which for typical grid dimensions used are $101 \times 51 \times 64000\mathrm{Bytes} \approx 315\mathrm{kiB}$, we show that for a computation of Eq. 6.74 only the count of samples with values $0, 1, -1$ are required, which results in a negligible file size. Taking advantage of the fact that the Chern number is an integer and can take only a small amount of values $C \in S \subset \mathbb{Z}$ we can express the expectation value in terms of counts $N_{1,s}, N_{2,s}$ as

$$\langle \min\{1, |C_1 - C_2|\}\rangle = \mathrm{E}\left[\min\{1, |C_1 - C_2|\}\right]_{C_1,C_2} \tag{6.76}$$

$$= \frac{1}{N_{1,s_1} N_{2,s_2}} \sum_{s_1,s_2 \in S} N_{1,s_1} N_{2,s_2} \min\{1, |s_1 - s_2|\}, \tag{6.77}$$

which is generally an $\mathcal{O}(1)$ operation and requires $\mathcal{O}(1)$ memory.

We compute the probability of change for the same data that we discussed in Fig. 6.8 and show the result in Fig. 6.9. Here, we use $C_{\mathrm{ref}} = C(z = 0)$, i.e., P represents the probability of a topological phase transition under the inclusion of an off-diagonal term $\Sigma^{AB}_{\mathrm{loc}}$ in the topological Hamiltonian. As we elaborated before, starting in the trivial phase of the Haldane model no $\Sigma^{AB}_{\mathrm{loc}}$ can close the band gap. Hence, the entire upper half of the graphic (everything above $m = m_c$) has $C = C_{\mathrm{ref}}$ and therefore $P = 0$. Additionally, we observe the stable region very clearly around $m = z = 0$. For the bounding gray line we use an ellipsis defined through $(m/m_c)^2 + z^2 = 1$, which fits the data rather well. Beyond this line, P increases rather rapidly and we find that values < 0.5 are only found close to the stable region. Using our more powerful statistical measure we now conclude unequivocally that there is a large region extending down from m_c at larger z, where the probability of change is $P = 1$. In this region a topological transition is guaranteed to happen. Upon comparison with the corresponding value of the Haldane model it is clear that this transition is from the non-trivial topological insulator to the trivial phase.

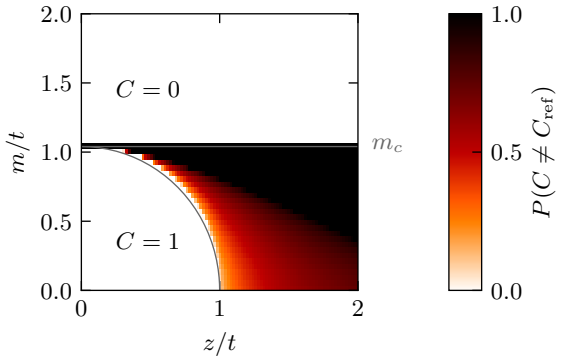

Figure 6.9: Probability of change of Eq. 6.74, where C_{ref} corresponds to the Haldane model at $z = 0$. Above m_c and around $m = 0, z = 0$ the non-trivial phase is stable w.r.t. Σ^{AB}_{loc}. In the region that lies between the gray lines a topological transition is rather likely with probabilities $P \sim 0.5 - 1$ except very close to the stable region. Extending down from m_c we even find a regime with $P = 1$.

We conclude the discussion of the local self-energy with the remark that while the diagonal term generally extends the topological phase towards larger values of m, the off-diagonal term has the opposite effect. At finite z/t the trivial $C = 0$ phase is extended towards smaller m irrespective of the phase of Σ^{AB}_{loc}. On the other hand, the non-trivial region is rather stable with respect to off-diagonal perturbations the farther away one is from the transition at $m = m_c$. The opposite effects of diagonal and off-diagonal terms can, in principle, cancel each other out, however, usually we expect $|\Sigma_{AA}| > |\Sigma_{AB}|$ and therefore the qualitative result obtained for the diagonal matrix elements is expected to remain, albeit weakened.

6.4.2 Magnetic Self-Energy

In the previous discussion we assumed that the self-energy is completely spin-independent and therefore block diagonal. A description in terms of a single spin was therefore appropriate. From this description, the effect of magnetism on the Chern number is not entirely clear, though. In a spinful description of an SU(2) symmetric model the Chern number can only assume even integer values, since it can always be decomposed into a sum of two components $C = C_{\sigma_1} + C_{\sigma_2}$, where $\sigma_{1,2}$ not necessarily correspond to S_z eigenvalues. In particular, for conserved S_z we have $C = C_\uparrow + C_\downarrow$, which is the case for the Haldane-Hubbard model. In the magnetically ordered phase, the SU(2) symmetry is broken spontaneously and hence the two contributions $C_\uparrow$, $C_\downarrow$ must not be the equal, which opens the possibility for odd total Chern numbers as found for the SBTI phase in Fig. 6.5.

Regarding the mean-field solution from Eq. 6.55, we can easily take into account the magnetization m (not to be confused with the mass term)

$$\Sigma^{\text{MF}}_\sigma = \frac{U}{2}(n - \sigma m) \tag{6.78}$$

for a single site, where $\sigma \in \{+1, -1\}$ for spin up and down, respectively. The density expectations in the mean-field Hamiltonian for two sites, cf. Eq. 6.60, are no longer equal and therefore

$$\Sigma^{\text{MF}}_i + \mu = U \begin{pmatrix} \langle n_{i\downarrow} \rangle & 0 \\ 0 & \langle n_{i\uparrow} \rangle \end{pmatrix} \tag{6.79}$$

on site i. Expanding the two sites we have

$$\Sigma^{\mathrm{MF}} + \mu = U \begin{pmatrix} n_{A\downarrow} & 0 & 0 & 0 \\ 0 & n_{B\downarrow} & 0 & 0 \\ 0 & 0 & n_{A\uparrow} & 0 \\ 0 & 0 & 0 & n_{B\uparrow} \end{pmatrix} \tag{6.80}$$

$$= \frac{U}{2} \begin{pmatrix} n_{A\downarrow} + n_{B\downarrow} & 0 & 0 & 0 \\ 0 & n_{A\downarrow} + n_{B\downarrow} & 0 & 0 \\ 0 & 0 & n_{A\uparrow} + n_{B\uparrow} & 0 \\ 0 & 0 & 0 & n_{A\uparrow} + n_{B\uparrow} \end{pmatrix} \tag{6.81}$$

$$+ \frac{U}{2} \begin{pmatrix} n_{A\downarrow} - n_{B\downarrow} & 0 & 0 & 0 \\ 0 & -(n_{A\downarrow} - n_{B\downarrow}) & 0 & 0 \\ 0 & 0 & n_{A\uparrow} - n_{B\uparrow} & 0 \\ 0 & 0 & 0 & -(n_{A\uparrow} - n_{B\uparrow}) \end{pmatrix}, \tag{6.82}$$

which for an antiferromagnetic phase with $n_\uparrow = n_\downarrow$ can be written conveniently as

$$\Sigma_\sigma^{MF} + \mu = \frac{U}{2} n + \frac{U}{2} (\Delta n \sigma_3 - \sigma \Delta m \sigma_3), \tag{6.83}$$

with $\Delta n = (n_A - n_B)/2$ and $\Delta m = (m_A - m_B)/2 = (n_{A\uparrow} - n_{A\downarrow} - n_{B\uparrow} + n_{B\downarrow})/2$. While the first term is the same as in the paramagnetic case we find an additional term proportional to the magnetization difference between the two sites, sometimes called "staggered magnetization" [154]. This second term is also proportional to σ_3, which means that it acts as a renormalization of the mass term in the topological Hamiltonian. However, in contrast to the paramagnetic case, this renormalization is now spin-dependent. In particular, the value of the mass term m (apologies for the clash in notation with the magnetization) differs by $2\Delta m$ between the two spins.

The general case can again be treated rather similarly and we obtain a relation similar to Eq. 6.67

$$\Sigma_\sigma = \frac{1}{2} \left[\bar{\Sigma} + (\Delta\Sigma - \sigma \Delta\Sigma_\sigma) \sigma_3 \right], \tag{6.84}$$

with $\bar{\Sigma} = \mathrm{tr}(\Sigma)$, $\Delta\Sigma = (\Sigma_{AA} - \Sigma_{BB})/2$ and $\Delta\Sigma_\sigma = (\Sigma_{AA\uparrow\uparrow} - \Sigma_{AA\downarrow\downarrow} - \Sigma_{BB\uparrow\uparrow} - \Sigma_{BB\downarrow\downarrow})/2$. The mass term renormalization is then given by

$$m \mapsto m + \frac{1}{2} (\Delta\Sigma - \sigma \Delta\Sigma_\sigma). \tag{6.85}$$

The beauty about this result is that without actually knowing the value of the self-energy we can in general predict the effect of the local self-energy on the phase diagram. It is clear that $\Delta\Sigma$ is negative, since the B site has a higher occupation at $m > 0$. In the regime where the topological Hamiltonian is valid, the second term is smaller than the first, since the density wave dominates over the onsite repulsion. Therefore, the previous result that the mass term is reduced is still generally valid with an additional spin-dependent term that leads to different m for different spins. This term is guaranteed to appear as a result of a spontaneously broken symmetry. The difference in m between spins is $\Delta\Sigma_\sigma$ and increases with the local magnetization. If the corresponding paramagnetic case is close to the phase transition, i.e., $m + \Delta\Sigma/2 \approx m_c$, the magnetic case will have one spin pushed below the transition, which means that $C_\sigma \neq C_{\bar{\sigma}}$ and therefore $C = C_\uparrow + C_\downarrow$ is odd. The antiferromagnetic ground state is degenerate and therefore $C_\uparrow$ and $C_\downarrow$ are ill-defined, or rather depend on the chosen ground state. In any case, within the magnetic phase of the Haldane-Hubbard model one spin is in the trivial phase ($C = 0$) and the other in the topological phase ($C = 1$) as has been confirmed in ED calculations, cf. Ref. [140].

6.5 Non-local Self-Energy

So far, we have taken into account only the local self-energy in Eq. 6.68. Now we focus on the non-local part. Due to the absence of an exact analytic expression we study a parameterized form and take as a first ansatz the TPSC self-energy. Upon comparison we had already confirmed that the TPSC prediction for the topological phase is the same as that of DMFT, even though the former is explicitly momentum-dependent. In order to investigate explicitly the term $\Sigma_{\mathrm{non-loc}}$ we also conducted a study taking into account only the momentum-averaged self-energy and found no change in the topological classification.

We go one step further here and investigate the susceptibility of the topological classification to the general form of the self-energy in TPSC and similar methods like RPA. In TPSC, there are two parameters U and U', that are usually determined self-consistently. Here, we use them as external parameters, the physical meaning of which is unimportant for the present argument. The parameterized ansatz function for a momentum-dependent self-energy is then given by

$$\Sigma(k) = ([V[U] + V[U']] * G^0)(k), \tag{6.86}$$

where $*$ denotes convolution and

$$V[U] = (1 - \chi^0 U)^{-1} \chi^0, \tag{6.87}$$

is the RPA expression [177] for the self-energy with the non-interacting susceptibility $\chi^0 = -G^0 * G^0$ [162]. Since χ^0 is a matrix in the sublattice basis we define U, U' as diagonal matrices with independent values for the A, B sublattices, which results in four free parameters for the self-energy. We now vary the values of U_A, U_B, U'_A, U'_B and compute the topological index with the topological Hamiltonian and H_0 close to the non-interacting phase transition. While restricting to moderate values for the parameters we do not find a topological phase transition that is induced by the explicit momentum-dependence of Eq. 6.86.

The ansatz in Eq. 6.86, albeit motivated by the previous experience with TPSC, is rather biased and does not allow for a general conclusion about the effects of the momentum-dependence. Hence, a more general approach is needed. In order to be able to investigate interacting systems in general, without even restricting to the Hubbard interaction, the only limitation being that a description in terms of the topological Hamiltonian is valid, we sample over the space of physical self-energy functions and compute the resulting distribution of the Chern number. Clearly, allowing arbitrary functions for Σ means that we cannot expect definite results that correspond to a specific phase. In fact, the aim is to learn very general qualitative information about how the momentum-dependence of the self-energy alters the topological classification (if it does so at all). A comparison with the Haldane-Hubbard phase diagram therefore does not make sense and we will not be able to say with certainty which version of the possible contours of the SBTI phase is most accurate. However, we will be able to understand more about where the differences originate and how the topological classification is changed.

6.5.1 General Formalism

For the Haldane-Hubbard model the zero frequency self-energy is a hermitian 2×2 matrix, where we again make use of the fact that the two spins are decoupled which enables us to investigate only one single spin. The complete solution is then again given by a sum of two Chern numbers of two spins where each has its own $\Sigma_{\mathrm{non-loc}}$. The general solution for a single spin is therefore enough to draw conclusions for the spinful model. As a general, maximally unbiased parameterization of the self-energy we define

$$\Sigma_{\mathrm{non-loc}} = \begin{pmatrix} f_0 + f_3 & f_1 - i f_2 \\ f_1 + i f_2 & f_0 - f_3 \end{pmatrix}, \tag{6.88}$$

where $f_0, f_1, f_2, f_3 : \mathbb{R}^2 \to \mathbb{R}$ are independent real-valued periodic functions of k. Clearly,

$$\Sigma_{\text{non-loc}} = \sum_i f_i \sigma_i \tag{6.89}$$

with $\sigma_0 = \text{Id}$. Using this decomposition in terms of hermitian matrices and the real-valuedness of f_i the parameterization is hermitian by construction. By inserting this into Eq. 6.54 with Eq. 6.68 we obtain the complete self-energy at zero frequency

$$\Sigma(\omega = 0, k) = f_0(k)\text{Id} + (a + f_1(k))\sigma_1 + (b + f_2(k))\sigma_2 + (\delta\Sigma + f_3(k))\sigma_3. \tag{6.90}$$

This decomposition is extremely general but well-defined with the requirement $\sum_k \Sigma_{\text{non-loc}}(k) = 0$ and therefore $\sum_k f_i(k) = 0$. We note that any constant term proportional to σ_0 would merely shift the chemical potential and is therefore neglected here. In order to satisfy our requirement of physical samples we take all f_i to be smooth functions, which is achieved by defining them in terms of a Fourier expansion

$$f_j(k) = \sum_{l_1,l_2,s} c_{s,l_1,l_2} \cos(l_1 k_1 + s l_2 k_2) + \sum_{l_1,l_2,s} c'_{s,l_1,l_2} \sin(l_1 k_1 + s l_2 k_2), \tag{6.91}$$

where $j \in \{0, 1, 2, 3\}$, $s \in \{-1, 1\}$, $l_1, l_2 \in \{0, \ldots, N_c\}$ and $c, c' \in \mathbb{R}$. In practice, we place a restriction on the parameters c, c' so that $c^{(\prime)}_{s,l_1,l_2} = c^{(\prime)}_{s,l_1,l_2}(1 - \delta_{l_1,0}\delta_{l_2,0})$. This ensures that no constant term $\propto \cos(0)$ is included. In addition, the sum over s is only performed if $l_1, l_2 \neq 0$, which guarantees that all coefficients of linearly independent functions are just single parameters c, c' (and not linear combinations thereof). N_c is the expansion order and defines the maximal frequency of modes included in the self-energy samples. Due to the use of trigonometric functions the periodicity in momentum-space in terms of reciprocal lattice vectors and the vanishing momentum-average are obeyed by construction. Since the basis functions cos and sin form a complete basis over the space of differentiable functions, taking $N_c \to \infty$ will allow us to represent any reasonable function (note that differentiability is required for the calculation of the Chern number). However, by comparing to the TPSC or FLEX [177] self-energies we found that already a very small cutoff N_c is sufficient to represent these functions. Specifically, we show in Fig. 6.10 that the TPSC momentum-dependence can be reproduced already with $N_c = 1$. The original data in Fig. 6.10**a** is reproduced very accurately by restricting to only very few, in this case 8, parameters, as shown in Fig. 6.10**b**.

For the actual simulation we increase the number of representable functions by setting $N_c = 2$, which yields 18 independent parameters. One such sample is shown in Fig. 6.10**c**. Clearly, these functions do not oscillate unphysically, since we do not allow high frequency contributions. At the same time the number of degrees of freedom is large enough to cover a sufficient number of possible self-energy functions. The qualitative result that we obtain through our analysis does not depend on the particular choice of this cutoff. We confirmed that while increasing N_c beyond 2 increases the target space of representable functions significantly, the fraction of interesting samples that change the Chern number decreases simultaneously. Therefore, $N_c = 2$ is the sweet spot that offers a large enough variety of functions and at the same time allows us to use a manageable number of samples.

In order to enforce more physicality on the samples, we engineered a specific distribution function from which the random variables—the parameters c, c'—are drawn

$$\rho(c^{(\prime)}_{s,l_1,l_2}) = \mathcal{N}(\mu = 0, \bar{\sigma} = \exp(-l_1 - l_2)), \tag{6.92}$$

where $\mathcal{N}(\mu, \bar{\sigma})$ denotes a normal distribution with mean μ and standard deviation $\bar{\sigma}$. The vanishing mean guarantees symmetry around 0 and therefore unbiased sign of the self-energy, while

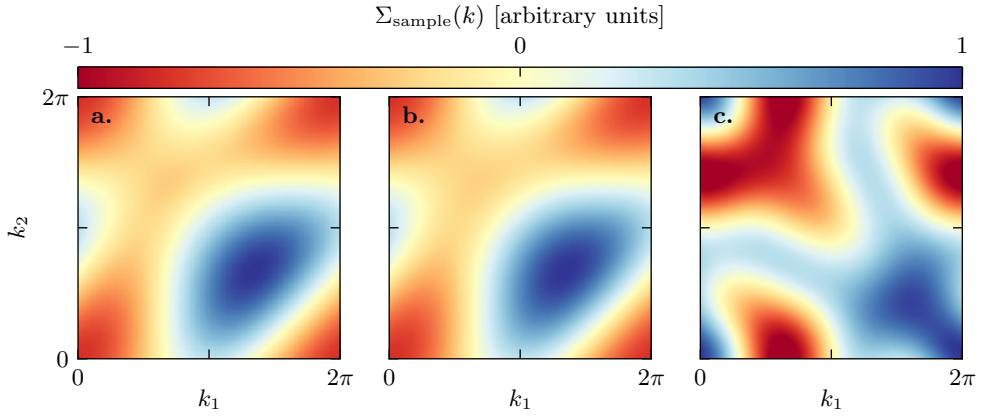

Figure 6.10: Exemplary self-energy samples as a function of $k = (k_1, k_2)$. **a.** Momentum-dependent part of the TPSC self-energy (diagonal matrix element) for an arbitrary choice of parameters $m/t = 1, U/t = 1.6$. The scale is normalized and the same in all subfigures. **b.** Fit of our general parameterization, cf. Eq. 6.91, to the TPSC data with $N_c = 1$. Apparently, the smallest non-trivial cutoff is already enough to describe the momentum-dependence produced by TPSC. **c.** Random sample with increased cutoff $N_c = 2$. [Figure adapted from Ref. [169]]

the normal distribution itself facilitates a selection of small parameters that are not exceptionally large. The standard deviation depends on the indices l_1, l_2 as $e^{-l_1-l_2}$, i.e., larger frequencies are exponentially suppressed on average. As a result, the sample functions have the properties we expect from the typical self-energy functions in the weak to intermediate coupling regime. The produced samples are highly dependent on the choice of the distribution function and, e.g., a uniform distribution would generate rather unphysical samples. Especially at larger values of N_c a decay of the size of target space for the c_{s,l_1,l_2} is essential for obtaining sensible samples.

Within our parameterization for the self-energy in Eq. 6.88 we made the assumption of a particular symmetry, namely that S_z is conserved. This allowed us to separate the two spins and reduce the description to a 2×2 matrix. Spatial symmetries on the other hand are thus far not accounted for, since the further specification of the sample functions in Eq. 6.91 does not impose any spatial symmetries on the self-energy. In order to steer the statistics obtained more into the realm of physical solutions we modify the sampling procedure such that certain symmetries of the model are conserved in the topological Hamiltonian. This can be enforced easily by deriving relations between the coefficients $c^{(\prime)}_{s,l_1,l_2}$ and sampling only independent random variables.

For the Haldane Hubbard model we generally have $\Sigma_{AA} = -\Sigma_{BB}$, cf. Eq. 6.66, which we incorporate by setting $f_0(k) \equiv 0$. In addition, the diagonal matrix elements have a mirror symmetry M with respect to the line $k_2 + k_1 = 0$

$$M : \begin{pmatrix} k_1 \\ k_2 \end{pmatrix} \mapsto \begin{pmatrix} -k_2 \\ -k_1 \end{pmatrix}. \tag{6.93}$$

We note that this applies both to the version of $H(k)$ derived earlier, cf. Eq. 6.39, and our implementation, where we chose the lattice vectors differently and therefore the momenta k_1, k_2 are exchanged. A symmetry like this generally implies certain constraints on the random variables that can be obtained by applying the symmetry operation M to the function f_3. Clearly,

$$[M f_3](k) = \sum_{l_1,l_2,s} c_{s,l_1,l_2} \cos(-l_1 k_2 - s l_2 k_1) + \sum_{l_1,l_2,s} c'_{s,l_1,l_2} \sin(-l_1 k_2 - s l_2 k_1) \tag{6.94}$$

$$= \sum_{l_1,l_2,s} c_{s,l_1,l_2} \cos(l_1 k_2 + s l_2 k_1) - \sum_{l_1,l_2,s} c'_{s,l_1,l_2} \sin(l_1 k_2 + s l_2 k_1) \tag{6.95}$$

$$= \sum_{l_1,l_2,s} c_{s,l_1,l_2} \cos(l_2 k_1 + s l_1 k_2) - \sum_{l_1,l_2,s} (-1)^{\frac{1-s}{2}} c'_{s,l_1,l_2} \sin(l_2 k_1 + s l_1 k_2) \tag{6.96}$$

$$= \sum_{l_1,l_2,s} c_{s,l_1,l_2} \cos(l_2 k_1 + s l_1 k_2) + \sum_{l_1,l_2,s} (-1)^{\frac{1+s}{2}} c'_{s,l_1,l_2} \sin(l_2 k_1 + s l_1 k_2), \tag{6.97}$$

where we used the (a)symmetry of cos and sin. By comparison with f_3 we find that

$$c_{s,l_1,l_2} = c_{s,l_2,l_1}, \qquad c'_{s,l_1,l_2} = -s c'_{s,l_2,l_1}. \tag{6.98}$$

These relations reduce the number of independent random variables and provide a recipe for enforcing the symmetry onto the samples.

For a general symmetry operation T we can follow the same procedure. The advantage over simply averaging over transformed functions is that this is (i) computationally much cheaper and (ii) one has to take care of only one operation per symmetry. The straight-forward approach of defining $\tilde{f} = f + Tf$ would work for the case $T = M$, however, care has to be taken if T has a period larger than 2, i.e., if $T^2 \neq \mathrm{Id}$ as is the case for 3-fold rotations. In general, one has a period $p \in \mathbb{N}$ and $T^p = \mathrm{Id}$. When defining $\tilde{f}$ one would have to average over $[T^i f](k)$ $\forall 1 \leq i < p$. When deriving constraints on the coefficients this is not necessary, since $Tf = f$ implies $T^i f = f$ $\forall i \in \mathbb{N}$, i.e., guarantees the idempotence of the minimal symmetry operation on the space of sample functions.

Although we did not use the averaging method described above we nevertheless want to elaborate on some more important details for completeness. Since the samples are supposed to be used for a statistical evaluation it is important that the symmetrization procedure does not unevenly change the probability distributions of the various random variables. We can formulate a couple of constraints:

(i) the symmetrization procedure must produce a complete coverage of the space of symmetric functions,

(ii) symmetric functions should be evenly distributed,

(iii) the probability distribution function of the symmetrized parameters should be well-defined, ideally the same as that used to produce the original samples.

Clearly, all of these constraints are satisfied if we sample symmetric solutions directly. However, we can show that these are satisfied also by the symmetrization procedure

$$\mathcal{T} : f \mapsto \tilde{f} = \sum_{i=0}^{p-1} T^i f. \tag{6.99}$$

Proof. Regarding (i): Let Ω be the space of all sample functions f, $\Omega_s \subset \Omega$ the symmetric subspace and $\tilde{f} \in \tilde{\Omega}$ or $\mathcal{T} : \Omega \to \tilde{\Omega} \subseteq \Omega_s$. Suppose that $Tf = f$ for $f \in \Omega$. Then,

$$\tilde{f} = \sum_{i=0}^{p-1} T^i f = pf, \tag{6.100}$$

i.e., f and $\tilde{f}$ correspond to the same function. Therefore, $\tilde{\Omega}$ contains all symmetric functions and $\tilde{\Omega} = \Omega_s$ (the other direction $f \in \tilde{\Omega} \Rightarrow f \in \Omega_s$ is trivial since all $\tilde{f} \in \tilde{\Omega}$ are also linear combinations of cos and sin and by construction symmetric).

For (ii) we have by definition $|\Omega| > |\Omega_s|$ for $p > 1$ ($T \neq \mathrm{Id}$). Therefore, each $\tilde{f}$ has a preimage of dimension > 1. We can understand the mapping $\mathcal{T}$ as a dimensional reduction,

where the resulting image $\tilde{f} \in \Omega_s$ has less degrees of freedom. Therefore, by identification of the corresponding remaining degrees of freedom we find the same number for each $\tilde{f}$, which proves the even distribution. An explicit construction has been shown before in terms of the degrees of freedom c_{s,l_1,l_2}.

Finally, the effect of $\mathcal{T}$ is a linear superposition of random variables. Suppose $Z = aX + bY$ with X, Y drawn independently from the same normal distribution. Then,

$$\rho_z(z) = \int_{-\infty}^{\infty} \rho_x(x)\rho_y\left(\frac{z-ax}{b}\right) \,\mathrm{d}x \tag{6.101}$$

$$= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} e^{-\frac{(z-ax)^2}{2b^2\sigma^2}} \,\mathrm{d}x \tag{6.102}$$

$$= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}\left[x^2 + \frac{a^2}{b^2}x^2 - \frac{2axz}{b^2} + \frac{z^2}{b^2}\right]\right) \,\mathrm{d}x \tag{6.103}$$

$$= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2b^2\sigma^2}\left[(a^2+b^2)x^2 - 2azx + \frac{a^2z^2}{a^2+b^2} - \frac{a^2z^2}{a^2+b^2} + z^2\right]\right) \,\mathrm{d}x \tag{6.104}$$

$$= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} \exp\left(-\frac{a^2+b^2}{2b^2\sigma^2}\left[\left(x - \frac{az}{a^2+b^2}\right)^2\right] - \frac{1}{2b^2\sigma^2}\left[z^2 - \frac{a^2z^2}{a^2+b^2}\right]\right) \,\mathrm{d}x \tag{6.105}$$

$$= \frac{1}{2\pi\sigma^2} \frac{\sqrt{2\pi b^2\sigma^2}}{\sqrt{a^2+b^2}} \exp\left(-\frac{1}{2b^2\sigma^2} z^2 \left[1 - \frac{a^2}{a^2+b^2}\right]\right) \tag{6.106}$$

$$= \frac{1}{2\pi\sigma^2} \frac{\sqrt{2\pi b^2\sigma^2}}{\sqrt{a^2+b^2}} \exp\left(-\frac{z^2}{2(a^2+b^2)\sigma^2}\right), \tag{6.107}$$

i.e., the distribution of the symmetrized random variable is also a normal distribution with zero mean but with modified variance $\sigma_z^2 = (a^2+b^2)\sigma^2$. The distribution can be kept the same by defining instead the normalized variable $Z' = \frac{1}{\sqrt{a^2+b^2}}(aX + bY)$. Therefore, indeed all three constraints are satisfied by the procedure of Eq. 6.100. □

In order to control the amount of momentum dependence we need to be able to generate samples with specific self-energy dispersion amplitudes d_a. It is not straight-forward to establish a relation between d_a and the parameters, so we use a more obvious approach. Given a sample Σ_{sample}, we compute the value of the dispersion amplitude and then rescale all parameters. Since we only multiply the entire matrix with a scalar the relation that exists on average between the parameters (through the different variances) is not changed. With the definition of d_a given in Eq. 5.8 we note that the computation is rather ineffecient. For real Σ the following equality holds:

$$d_a(\Sigma) = \max_{k,k'} \parallel \Sigma(k) - \Sigma(k') \parallel_\infty \tag{6.108}$$

$$= \max_{k,k'} \max_{ij} |\Sigma_{ij}(k) - \Sigma_{ij}(k')| \tag{6.109}$$

$$= \max_{ij} \max_{k,k'} |\Sigma_{ij}(k) - \Sigma_{ij}(k')| \tag{6.110}$$

$$= \max_{ij} \left[\max_k \Sigma_{ij}(k) - \min_k \Sigma_{ij}(k)\right]. \tag{6.111}$$

In general, though, the off-diagonal matrix elements of the self-energy are complex numbers with finite imaginary part. Therefore, the computation of d_a is $\mathcal{O}(N_k^2)$ in general and $\mathcal{O}(N_k)$ for real matrix elements. A separation usually requires more effort and is only worthwhile for a larger size of the self-energy matrix (here $\Sigma \in \mathbb{C}^{2\times 2}$). Since $\Sigma(\omega = 0, k) = \Sigma(\omega = 0, k)^\dagger$ is hermitian we can restrict the calculation to the upper or lower triangular matrix, which almost cuts the calculation in half, since for an $N \times N$ matrix we only have to take into account

$$N^2 \to N + \frac{N^2 - N}{2} = \frac{N^2 + N}{2} \tag{6.112}$$

matrix elements. Of course, this is mostly relevant only in the context of a Monte Carlo simulation where the same calculation is performed many times and even small savings have an immediate effect on the total run time and thus allow for an increased sample size.

6.5.2 Sampling and Analysis

We proceed by using the method described in the previous section to gather statistics of the Chern number for the Haldane model as a function of the parameters z, d_a, and therefore describing all possible self-energies. We neglect the diagonal part of the local self-energy, since that only shifts the transition along the m-axis. This means that all of our results are trivially generalized also to the case where the diagonal part $\delta\Sigma$ is included. We emphasize that the probability distribution of our choice (Eq. 6.92) for the non-local self-energy contributions introduces a certain bias towards more physical samples. This can be understood as a kind of importance sampling, where the probability of useful samples is artificially increased in order to reduce the amount of samples required to obtain useful results. Here, we use $n_{\text{samples}} = 10^7$, which mainly controls the noise in the data. Without importance sampling, i.e., by drawing parameters from a uniform distribution, the size of the sample space becomes much larger and at the same time we expect that also the fraction of samples that change the Chern number decreases, which would weaken or even destroy the signal that we are interested in. In practice, however, we found the same qualitative and quantitative result for the probability of change irrespective of the statistical distribution, where the probability of change was only insignificantly larger for the (on average) more rapidly oscillating samples obtained from the uniform distribution. This comparison was performed at $N_c = 2$. Increasing N_c beyond this value indeed reduces the measured probability of change, i.e., the fraction of interesting samples. Hence, the artificially reduced size of the sample space is not only physically reasonable but also necessary to obtain useful results. Motivated by the good quantitative agreement between the maximally unbiased uniformly distributed approach and our more refined version we conclude that the generality of the ansatz is not affected by our choice of the distribution function.

We also compared the effect of symmetries by performing one calculation without and one with symmetrization. It turns out that the results differ only marginally, while the symmetric approach led to a larger probability of change due to the reduced number of degrees of freedom. It is important to note here that the spatial symmetries in question do not protect the topological phase. Therefore, the conservation of symmetries is not required to retain the topological phase. On the other hand, the number of possible samples increases exponentially with each added parameter and we generally expect this to also increase the number of non-trivial samples. However, our methodology is based on integrated quantities that scale with the relative number of samples which change the topological phase. Given that the size of the sample space increases exponentially as a function of the number of expansion parameters, we believe that it is highly unlikely that the fraction $|\Omega_{\text{non-triv}}|/|\Omega|$ remains constant or even increases, which would indicate that trivial and non-trivial samples have the same number of degrees of freedom.

All results discussed now were obtained with symmetrized coefficients with cutoff $N_c = 2$ drawn from the probability distribution of Eq. 6.92. We proceed by studying the non-local contributions on their own first. This means we set $z = 0$ and therefore $\Sigma = \Sigma_{\text{non-loc}}$. In Fig. 6.11 we show the probability of change $P(C \neq C_0)$, c.f. Eq. 6.74, where C_{ref} has been chosen to be the non-interacting Chern number $C_0 = C(\Sigma = 0)$. The non-interacting phase

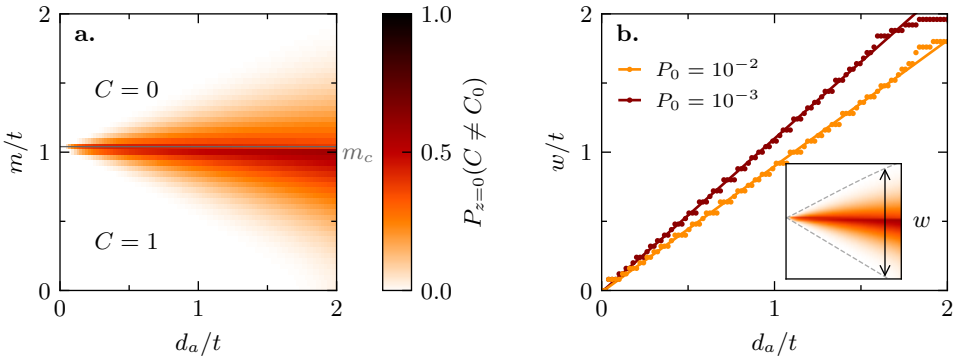

Figure 6.11: a. Probability that the perturbation $\Sigma_{\text{non-loc}}$ changes the Chern number as a function of $d_a = d_a(\Sigma_{\text{non-loc}})$. P is finite in a region around the non-interacting transition located at $m = m_c$, the width of the region increases with d_a. Overall, the non-trivial phase is more likely to change, especially close to m_c. **b.** Width w of the region with finite P as a function of d_a for $P_0 = 10^{-2}$ and 10^{-3}. The definition of w is indicated in the inset. w increases roughly linearly with d_a as shown by the linear fit (lines). [Subfigure **a.** adapted from Ref. [169]]

transition is marked by a gray line in Fig. 6.11**a** at $m = m_c$. We find that the Chern number changes only in a region approximately centered around m_c that widens with increasing d_a. The maximal probability is found always at or close to m_c. Interestingly, values for $m < m_c$ are larger, which indicates that the topologically non-trivial phase with $C = 1$ is more susceptible to a perturbation in the topological Hamiltonian. We found the same result already for the off-diagonal part of the local self-energy. In Fig. 6.11**b** we investigate the width w of the region with finite P quantitatively. The inset explains the definition of w, which we take to be simply the range of m values with finite $P > P_0$. The boundaries are almost symmetric with respect to m_c and approximately linear. Since our data is given on a discrete grid we perform a polynomial fit and obtain good results for varying degrees of up to order 3. For the width itself we plot the order 1, i.e., linear fit to demonstrate the good agreement. Clearly, the data reveals an almost linear dependence, where the slope increases with decreasing threshold value P_0. The latter is clear since the probability P is a smooth function of m. Assuming that $P(m) < P_0$ and $P_0' = P_0 - \delta$ with $\delta > 0$ we can find $\epsilon > 0$ with $m' = m + \epsilon$ and $P(m') < P_0'$. The slope will converge for $P_0 \to 0$, however, since smaller value of P become more and more susceptible to noise we make no attempt to extract the limit at this point. Instead we will perform a more quantitative analysis later.

We interpret the width w of the region of finite probability as an interval of uncertainty around the non-interacting transition. Namely, due to our unbiased approach we can conclude that a change of the Chern number as a result of non-local contributions in the self-energy is possible only for $m \in I = [m_c - w, m_c + w]$. For values outside this interval such an effect is highly unlikely if not impossible. The very simple linear relationship between the probability of change and the dispersion amplitude of the self-energy allows us to write $I \approx [m_c - d_a, m_c + d_a]$, which suggests an interpretation of d_a as the size of the error bar around the non-interacting transition. Taking into account the diagonal part of the local self-energy, the entire diagram is shifted along the m axis, which means that the critical value of the topological transition is

renormalized to $m_c+\delta\Sigma$, where $\delta\Sigma < 0$. Such a result without off-diagonal terms or an explicitly momentum-dependent self-energy is, e.g., provided by DMFT. Therefore, we can immediately use our statistical result to assign a maximal error bar to the location of the DMFT phase transition.

The power of our statistical method lies in the generality of the statement it makes. Since we did not restrict ourselves to a particular approximation for the calculation of Σ our result is universal and therefore the exact location of the phase transition *must* lie within the bounds revealed by our data.

We now add the off-diagonal term by allowing for finite z and repeat the same analysis with the full self-energy of Eq. 6.54, which now looks like

$$\Sigma(\omega = 0, k) = \begin{pmatrix} \delta\Sigma + f_3(k) & a + f_1(k) - i(b + f_2(k)) \\ a + f_1(k) + i(b + f_2(k)) & -\delta\Sigma - f_3(k) \end{pmatrix}, \tag{6.113}$$

with $\sqrt{a^2 + b^2} = z$ and $d_a(\Sigma[f_1, f_2, f_3]) = d_a$ fixed. We have seen previously that without the explicitly momentum-dependent term the local self-energy can be parameterized by $\delta\Sigma$ and z, which have opposite effects on the phase diagram. Clearly, the Chern number (Eq. 2.37) is not linear in perturbations that we add to the Hamiltonian. Thus, the result obtained with the complete self-energy is entirely unpredictable on a case by case basis. On average, though, the closer the system is to the topological transition the more sensitive it should become towards perturbations that can push it over the edge. On the other hand, the Haldane Hamiltonian from Eq. 6.39 that we use here as a proxy for topological models has a rather strong momentum-dependence. With our choice of parameters $t_2/t_1 = 0.2, \phi = \pi/2$ we have in terms of t_1

$$\begin{aligned} H(k) = (1 + \cos(k_1) + \cos(k_2))\,\sigma_1 + (\sin(k_1) + \sin(k_2))\,\sigma_2 \\ + (m - 0.4\,[\sin(k_1) - \sin(k_2) + \sin(k_2 - k_1)])\,\sigma_3, \end{aligned} \tag{6.114}$$

and therefore $d_a(H)/t_1 \approx 4$. With this number in mind we can refine our expectations. At small $d_a(\Sigma_{\text{non-loc}})$ the Chern number should be rather unlikely to change, since $\Sigma_{\text{non-loc}}$ is too small to be more than just a minor perturbation and this should not remove the poles in the Berry curvature. At $d_a(\Sigma) \approx d_a(H)$ on the other hand the perturbation is on the same order and therefore changes across the board should be very likely. Since this is a rather trivial limit we always restricted ourselves to much smaller dispersion amplitudes of the self-energy, which is a realistic assumption for the regime of intermediate correlations.

In the following, we use the local parameters m and z as the free parameters and perform calculations for a small number of values of d_a. For performance reasons we do not use a fine grid for d_a as done previously. For the choice of d_a we make the following estimate. Using the Hartree value of Σ as an approximant for the local self-energy and taking at half filling $\| \Sigma_{\text{loc}} \| \approx Un \approx \frac{U}{2}$ we can estimate values of d_a with significant momentum-dependence via the relative dispersion amplitude of Eq. 5.9 as

$$d_r \approx \frac{2d_a}{U}. \tag{6.115}$$

As we have seen in our investigation of the ionic Hubbard model we can identify strong momentum-dependence with values $d_r \approx 0.1$ and therefore $d_a \approx 0.05U$. With, e.g., $d_a/t = 0.5$ we can therefore assume that strongly momentum-dependent self-energies of up to $U/t \approx 10$ are included in our statistics. On the other hand, $d_a > 1$ should be rather unrealistic for the low-to-intermediate coupling regime where we can apply the topological Hamiltonian.

With z fixed we sample the off-diagonal part of the local self-energy from a uniform distribution as $a + ib = ze^{i\alpha}$ with $\alpha \in [0, 2\pi]$. While the average Chern number is not necessarily

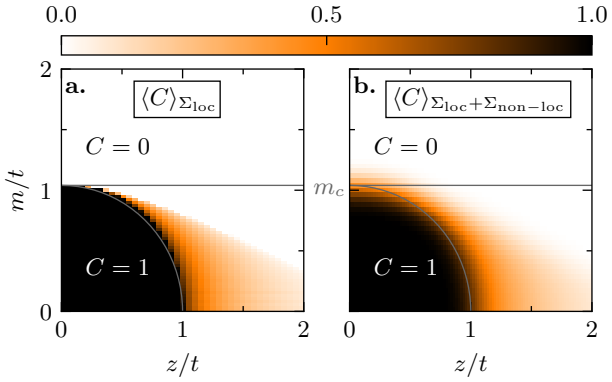

Figure 6.12: Average chern number as a function of m and z. m_c is the location of the non-interacting phase transition. **a.** Average over local self-energies only (same as Fig. 6.8). **b.** Average over total self-energy samples including local and non-local contributions with $d_a/t = 0.5$. The non-local average adds merely a blur. [Figure adapted from Ref. [169]]

a meaningful statistical indicator, we nevertheless compare in Fig. 6.12 the result for the total self-energy with that obtained earlier for the local part only, cf. Fig. 6.8. For this calculation we used $d_a/t = 0.5$, which should already describe rather strongly momentum-dependent solutions. We find that while the non-local part leads to a clear alteration of the phase diagram, the addition of the non-local part $\Sigma_{\rm non-loc}$ merely adds an additional blur onto the purely local result. The already blurred out region beyond the stable $C = 1$ phase, which arose as an average over the different phases of the off-diagonal local contributions, does not change much, only the relatively sharp transition from $\langle C \rangle > 0$ to $\langle C \rangle = 0$ is smoothened. The stable $C = 1$ phase is also largely unaffected by the additional momentum-dependent perturbation. Deep within this region we find no change at all, only at the boundary we observe a more continuous transition as opposed to the sharp step when including only local perturbations. This is most apparent at small z or even $z = 0$, where initially there was a very sharp transition from $\langle C \rangle = 1$ to $\langle C \rangle = 0$. The added non-local terms broaden this step function into a sigmoid function, which is essentially the result of Fig. 6.11. Interestingly, this observation seems to generalize also to finite z, at least in this description in terms of an average Chern number.

Clearly, the Chern number can take more than two values (although in our calculations we have not come across anything other than 0, 1, -1) and therefore $\langle C \rangle$ can only give a rough overview and does not accurately describe the precise effect of the perturbations. For this reason we use our improved estimator, the probability of change, cf. Eq. 6.74, to gain a more precise indication. To this end we set again our reference as $C_{\rm ref} = C_0$, i.e., we measure the probability $P(C_{\rm ref} \neq C_0)$ that the Chern number changes w.r.t. the non-interacting value as a result of the perturbation $\Sigma = \Sigma_{\rm loc} + \Sigma_{\rm non-loc}$.

The result of such a calculation is shown in Fig. 6.13 for $d_a/t \in \{0.25, 0.5, 1\}$. We immediately note the striking similarity to the previous result at $d_a = 0$ shown in Fig. 6.9. Again, we observe large regions of vanishing probability, i.e., stable regions, at small z/t and m/t for $C = 1$ and $m > m_c$ irrespective of z for $C = 0$. Both of these remain pretty much the same up to a small "melting" effect at the boundary, the severity of which depends on the strength of d_a. We also observe the region with $P = 1$, i.e., where we can be certain that the perturbation leads to a change of the topological classification. The shaded region in between, where $0 < P < 1$, increases in size at finite d_a.

Neglecting for the moment the quantitative aspect we concentrate on $d_a/t = 0.5$, i.e., Fig. 6.13**b**, for a qualitative discussion. The broadening or blurring of the phase transition at $z = 0$ around $m = m_c$ is clearly an effect of the momentum-dependent part of the self-energy.

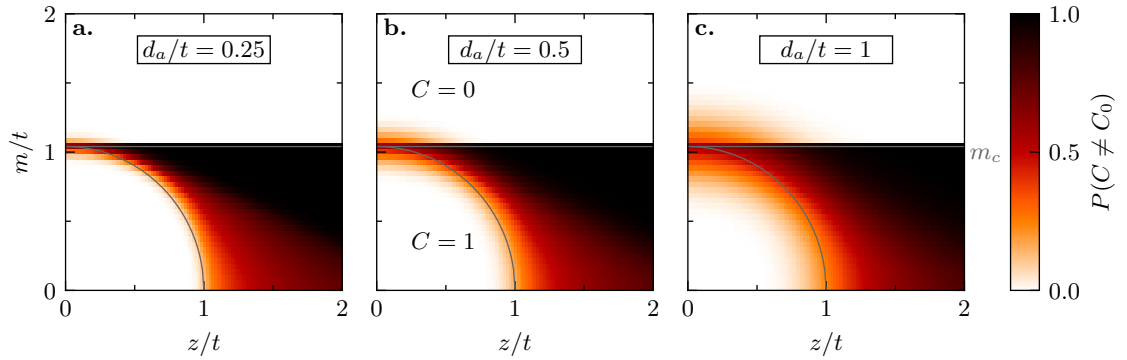

Figure 6.13: Probability of change $P(C \neq C_0)$ w.r.t. the non-interacting value under a perturbation $\Sigma(k) = \Sigma_{\text{loc}} + \Sigma_{\text{non-loc}}(k)$ that corresponds to the total self-energy for different dispersion amplitudes **a.** $d_a/t = 0.25$, **b.** $d_a/t = 0.5$ and **c.** $d_a/t = 1$. Independently of d_a the probability looks rather similar to the local result, cf. Fig. 6.9. An additional uncertainty due to the non-local part that increases in intensity as a function of d_a is visible, most clearly around $z = 0$ but also along the elliptical boundary of the stable $C = 1$ region. m_c is the critical value for the topological phase transition at $\Sigma = 0$. [Subfigure **b.** adapted from Ref. [169]]

In analogy to Fig. 6.11 this provides essentially a region of uncertainty for the exact location of the topological transition. Our statistical method does not predict exact values, which could only be obtained via an exact self-energy. Instead, by systematically going through all possibilities we canvass the parameter space for a region where the phase transition *could* happen. The result of our previous analysis at $z = 0$ provides also the width of this uncertainty to be $w \approx d_a/t = 0.5$. Although there is an asymmetry w.r.t. m_c in the sense that P is slightly larger below m_c than at the same distance above, which indicates that the non-trivial region is more sensitive towards the addition of non-local perturbations, both phases are affected similarly: the region of stability shrinks. Following the elliptical line that bounded the $C = 1$ stable region at $z = 0$ we find the same onset of finite P close to the line, which resembles melting. Following the horizontal line at m_c on the other hand we do not observe melting everywhere. In particular, at large z the transition from $P = 1$ to $P = 0$ remains sharp. This can be easily understood by remembering that $P = 1$ appears below m_c only, i.e., it indicates that there is a topologically trivial region in the formerly non-trivial regime of values of the mass m. The "transition" in P is therefore an artifact of the method of illustration and merely indicates that a topological transition w.r.t. the reference at $z = 0, d_a = 0$ takes place.

We demonstrate this in detail by plotting a "statistical phase diagram" in Fig. 6.14, where we assign a probability

$$P(C = i) = 1 - \langle \min\{1, |C - i|\} \rangle = 1 - P(C \neq i) \tag{6.116}$$

to each point in the phase diagram. $P(C = i)$ corresponds to the fraction of samples that produce a Chern number $C = i$. In our simulations over $n_{\text{samples}} = 10^7$ samples we only came across three phases, namely $C = 0, 1, -1$, for which we illustrate P separately in subfigures **a**,**b** and **c** of Fig. 6.14. Apparently, this different way of illustration resolves the supposedly sharp transition at $m = m_c$ for large z. In this part of the phase diagram only the trivial $C = 0$ phase is present, all other phases have $P = 0$. The trivial phase is certain throughout most of the phase diagram, predominantly in the upper half, while the $C = 1$ phase is found in the stable region bounded by the elliptical line. The $C = -1$ phase can only appear for $z > 1$ and is not particularly likely even there due to the strong dependence on the phase of the off-diagonal part of the local self-energy that is averaged over. The colored regions therefore contain a mixture of all three phases

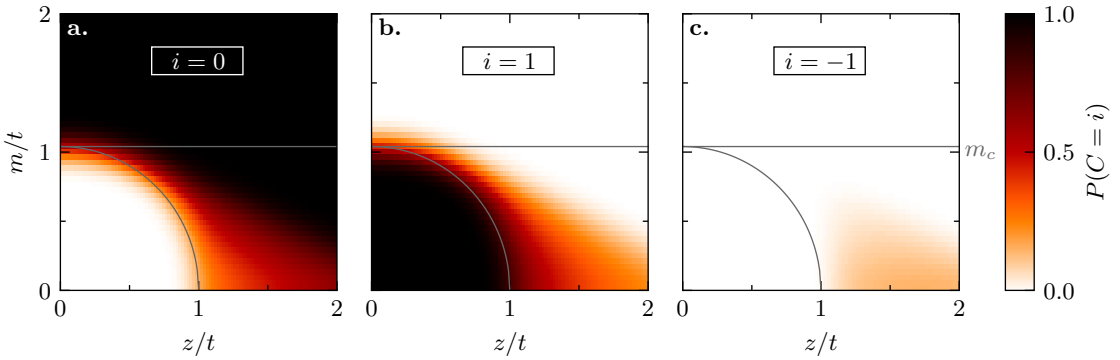

Figure 6.14: Statistical phase diagram for a perturbation with the total self-energy at $d_a/t = 0.5$. We plot for each point the probability that a specific phase is found. Possible phases are **a.** $C = 0$, **b.** $C = 1$ and **c.** $C = -1$. All other Chern numbers have $P = 0$ everywhere. To help the comparison we added lines at $m = m_c$ and at the transition for $d_a = 0$. $C = 0$ is certain in the upper part of the diagram, $C = 1$ within the stable region at small m, z. $C = -1$ can only appear beyond $z = 1$ and is rather unlikely.

which is caused by an uncertainty through $\Sigma_{\text{non-loc}}$ and Σ^{AB}_{loc}. Moreover, upon comparison with Fig. 6.8 we deduce the shape of the corresponding diagram without the explicitly momentum-dependent perturbation $\Sigma_{\text{non-loc}}$, where only the broadening is removed. Hence, the inclusion of the explicit momentum-dependence of the self-energy leads to an uncertainty of the exact location of the topological phase transition and therefore a melting of the stable regions.

In order to investigate this effect further we now drop the average over the phase α in $\Sigma_{\text{loc}} = ze^{i\alpha}$ and instead perform a number of simulations at fixed α to reduce the origin of uncertainty to only the explicit momentum-dependence described by $\Sigma_{\text{non-loc}}$. Then, we investigate the probability of change of Eq. 6.74 with a reference $C_{\text{ref}} = C_{\text{loc}}$, which corresponds to the result at $\Sigma = \Sigma_{\text{loc}}$. The resulting probability measure $P(C \neq C_{\text{loc}})$ encodes the uncertainty brought about by $\Sigma_{\text{non-loc}}$ only.

A variety of distributions of $P(C \neq C_{\text{loc}})$ are shown in Fig. 6.15 for different values of $\alpha \in \{0, \pi/4, \pi/2, 3\pi/4, \pi\}$ and $d_a/t = 1$. Keeping the phase fixed removes a degree of freedom in the sampling procedure, which allowed us to reduce the number of samples to $n_{\text{samples}} = 10^6$ for more efficient computations. The quality of the statistics is not affected by this as we can see from the smooth distribution of P across the entire diagram. While we could already observe earlier (when looking at the effects of the total self-energy) that the added non-local part leads to an additional uncertainty around the local phase transition, we have now more immediate proof of this. In fact, the measure $P(C \neq C_{\text{loc}})$ computed as

$$P(C \neq C_{\text{loc}}) = \langle \min\{1, |C - C_{\text{loc}}|\} \rangle_{\Sigma = \Sigma_{\text{loc}} + \Sigma_{\text{non-loc}}} \tag{6.117}$$

is finite only in the immediate vicinity of the local transition, which we have marked in the figures with gray lines. Consequently, the additional non-local perturbation $\Sigma_{\text{non-loc}}$ can change the Chern number only close to the local transition. The distribution itself (cut orthogonally to local transition line) has a bell shape with the maximum close to the local transition as shown in Fig. 6.15**f**. In all cases the distribution is not entirely symmetric and we find instead a slightly larger weight below the transition, i.e., in the topologically non-trivial $C = 1$ phase, indicating that this phase is less stable w.r.t. such perturbations. This increased weight, however, is only significant very close to the transition.

While all simulations yield very similar results apart from the exact shape of the local transition line, the case is different for $\alpha = \pi$. This is the only value in our selection for which the $C = -1$ phase appears. Comparison with Fig. 6.8**c** reveals that $C = 1$ is possible only for

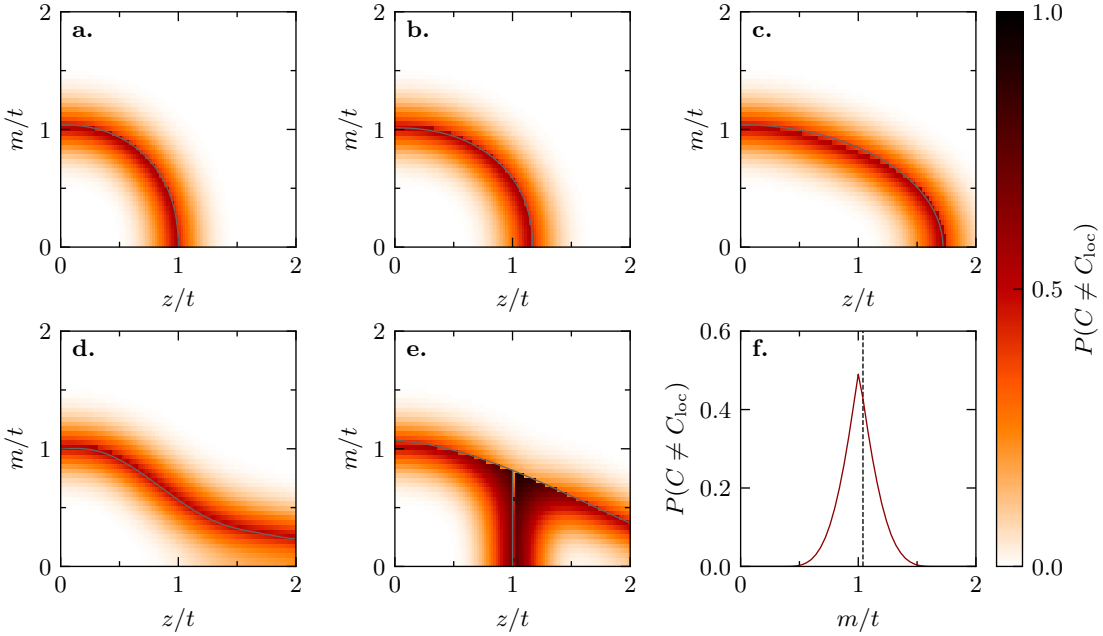

Figure 6.15: Probability of change with a reference $C_{\mathrm{ref}} = C_{\mathrm{loc}}$ encoding the uncertainty due to $\Sigma_{\mathrm{non-loc}}$ only. Different values of **a.** $\alpha = 0$, **b.** $\alpha = \pi/4$, **c.** $\alpha = \pi/2$, **d.** $\alpha = 3\pi/4$ and **e.** $\alpha = \pi$ are shown. **f.** Cut at $z = 0$, same for all α. The probability is finite only in an envelope around the local transition. The local transition is marked with solid gray lines. All calculations are at fixed $d_a/t = 1$. [Subfigures **a.**,**c.** adapted from Ref. [169]]

$z < 1$ and $C = -1$ only for $z > 1$. Thus, the diagram in Fig. 6.15**e** contains three topologically distinct phases separated by transition lines. We note that there is a region where the probability of change is especially large, unlike anything we found for other values of α. This is apparently an effect of the different topology of the diagram, i.e., the presence of a triple point where three phases meet. The maximum of the probability of change is found approximately at this triple point $(m_{\mathrm{tp}}, z_{\mathrm{tp}}) \approx (0.81t, 1t)$, which is of course metallic, but in the vicinity of which the topological phase is especially sensitive to momentum-dependent perturbations. Moreover, the $C = -1$ phase is slightly less stable as indicated by larger probability of change at similar distance to the transition line.

We now extract upper error bounds from our simulations by defining enveloping functions that contain the regions of finite probability. Apparently, the probability decays with the distance measured along the normal to the local transition line. Given a graph

$$g(l) = \begin{pmatrix} x(l) \\ y(l) \end{pmatrix} \tag{6.118}$$

parameterized by $l \in [0, 1]$ the derivative with respect to l always points along the graph, since

$$g(l + h) - g(l) \approx h\partial_l g(l). \tag{6.119}$$

We can then compute a normal vector at any l as

$$n(l) = \begin{pmatrix} -\partial_l g_2(l) \\ \partial_l g_1(l) \end{pmatrix}. \tag{6.120}$$

For the cut through the data we can then evaluate for fixed l the parameter vector

$$p_l(s) = g(l) + s\frac{n(l)}{\| n(l) \|} \tag{6.121}$$

for $s \in [-d_{\max}, d_{\max}]$, where $d_{\max}$ is the maximal distance from the transition line.

In analogy to our investigation of the case with $z = 0$, i.e., $\Sigma = \Sigma_{\text{non-loc}}$, as shown in Fig. 6.11, we define the width w of the uncertainty interval around the local transition as

$$w = \max(I) - \min(I), \qquad I = \{s : P(p_l(s)) = 0\}. \tag{6.122}$$

Numerically, of course, floating point numbers cannot be treated exactly. Therefore, we settle for a less strict condition and limit I to those points where $P(p_l(s)) < 10^{-3}$.

We compute this explicitly for $\alpha = 0$ by setting $x(l) = m(l), y(l) = z(l)$ and using the transition lines we have fitted to the data for m, z. From a visual inspection of Fig. 6.15 we already know that the data for other values of α behaves similarly. The transition line is approximately given by

$$m(z)/t = m_c(1. - (z/t)^{\frac{1}{0.45}})^{0.45}, \tag{6.123}$$

with $m_c/t = 1.04$. The expression in Eq. 6.123 has been obtained through a fit to the data. The resulting width of the uncertainty interval $w(l)$ for all points along the transition line can be extracted by interpolating the data through a bivariate spline of degree 3 and subsequently evaluating the probability

$$P_l(s) = P_{\text{spline}}(p_l(s)) \tag{6.124}$$

using Eq. 6.121. In Fig. 6.16 we show the result of such a calculation. In Fig. 6.16**a** we illustrate the slices through the transition line (solid line) as dotted lines. The values of these points are computed via Eq. 6.121. We mark with dashed lines the boundary of the uncertainty region, where we use the threshold 10^{-3} to decide whether the topological phase is stable ($P < 10^{-3}$) or unstable ($P > 10^{-3}$). Fig. 6.16**b** shows the probabilities along these cuts for different values of

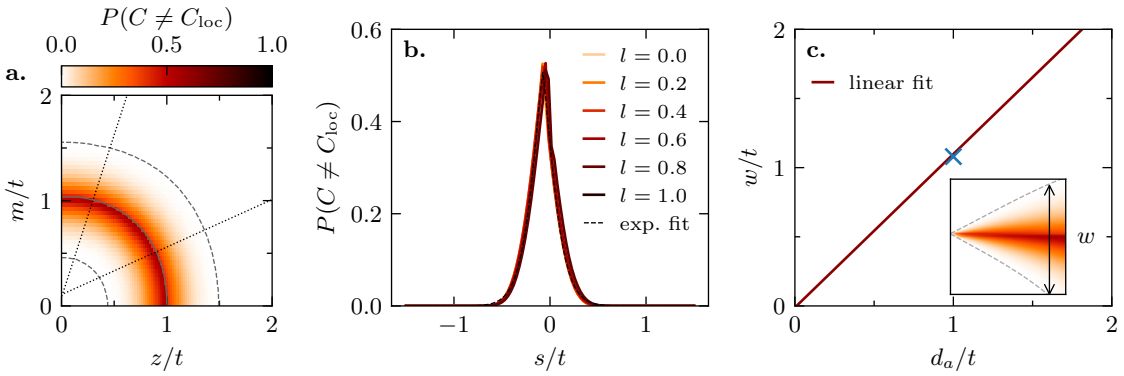

Figure 6.16: Probability of change $P(C \neq C_{\text{loc}})$ along orthogonal slices through the local transition line. **a.** Slices are illustrated (dotted lines) for two points on the line and the resulting uncertainty interval is marked by the dashed lines. **b.** Probability of change for slices through various points along the line. We find approximately the same values independent of the position l on the transition line. Hence, $z = 0$ can be used for reference. The probabilities decay exponentially. **c.** Width as a function of d_a, cf. Fig. 6.11. We mark the point extracted for $d_a/t = 1$, which lies exactly on the $z = 0$ line extracted earlier.

l along the line. An exponential fit reveals that the probabilities decay exponentially from the maximum value located slightly below $s = 0$. We can therefore give the following approximate relation

$$P(C \neq C_{\text{loc}}) \approx P_{\max} e^{-\left|\frac{s}{\xi}\right|^{\nu}}, \tag{6.125}$$

where we obtained for the exponent $\nu \approx 1.3$ and for the length scale $\xi/t \approx 0.16$. Clearly, the fit is not perfect, albeit coming very close, which suggests that the exact behavior is not represented

exactly by a single exponential. However, we can regard Eq. 6.125 as an upper bound, where we neglect contributions with larger exponents and smaller length scales.

As expected from a visual inspection of the data, the probabilities are approximately independent of l, which allows us to use any value for reference. In particular, we can use $l = 0$, which corresponds to $z = 0$ and is therefore independent of both z and α. We have done this calculation already, for clarity we provide the plot again in Eq. 6.125**c**, where the linear fit through the widths of the unstable region (difference between upper and lower boundary) is plotted against d_a. The boundary is shown in the inset for reference. For this calculation at $d_a/t = 1$ we also mark the extracted width to show explicitly that it lies on the line and note again that $w \approx d_a$.

6.5.3 Separability of the Chern Number and General Consequences for the Phase Diagram

Since the width of the uncertainty region is independent of the off-diagonal perturbation Σ_{loc}^{AB} we conclude that the local and non-local parts of the self-energy, which we have defined through the decomposition in Eq. 6.54, have two different effects on the resulting topological classification that are essentially separable. We summarize this important result in the following decomposition of the total Chern number C for the interacting system

$$C = C_{\text{loc}} + \delta C_{\text{non-loc}}, \tag{6.126}$$

where C_{loc} denotes the local Chern number obtained with the topological Hamiltonian $h_{t,\text{loc}} = H_0 + \Sigma_{\text{loc}}$ and $\delta C_{\text{non-loc}}$ is a random variable that takes values $\delta C_{\text{non-loc}} \in \mathbb{Z}$ with probabilities $P(\delta C_{\text{non-loc}} \neq 0)$ decaying exponentially as a function of the distance to the local transition. This decomposition allows us to state quite generally that the phase transition is located in an interval around the local transition, with the width of this interval of uncertainty given approximately by the self-energy dispersion amplitude d_a. Provided that the self-energy dispersion amplitude is moderate—as expected throughout most of the phase diagram away from the Mott phase where the topological Hamiltonian is not applicable—the local transition is most influential.

We show in Fig. 6.17 the location of the maximal probability as a function of d_a. Here, $s_{\max}$ is a position defined via Eq. 6.121, where $s = 0$ corresponds to the local phase transition. For reference we provide again the probability of change in Fig. 6.17**a**. In Fig. 6.17**b** we plot the length scale ξ extracted from an exponential fit of the data to Eq. 6.125 for all values of d_a and find an approximate linear dependence that is underlined by the good agreement with a linear fit that is also shown. Apparently, increasing the amount of momentum-dependence leads to a proportionally increased uncertainty in the topological classification. The representation in terms of ξ rather than w is more convenient, since, unlike w, ξ is invariant under the details of the sampling protocol. As we saw in our comparison between symmetric and unsymmetric data, only the value of $P_{\max}$ changes, while ξ remains the same. The width w on the other hand depends on $P_{\max}$ and P_0 (the threshold for zero-probability).

Regarding the maximum of the probability, i.e., the center of the uncertainty interval, we show in Fig. 6.17**c** that the maximum is initially exactly at the local transition line until $d_a/t \approx 0.5$. For more dispersive perturbations the maximum is shifted slightly into the topologically non-trivial regime, i.e., towards negative s. However, we do not find a further increase of $s_{\max}$ for larger d_a, which indicates that the interval of uncertainty is always approximately centered around the local transition.

We note that the values for the probabilities depend somewhat on the sampling procedure. In fact, using symmetrized instead of unsymmetrized data led to a difference of up to a factor

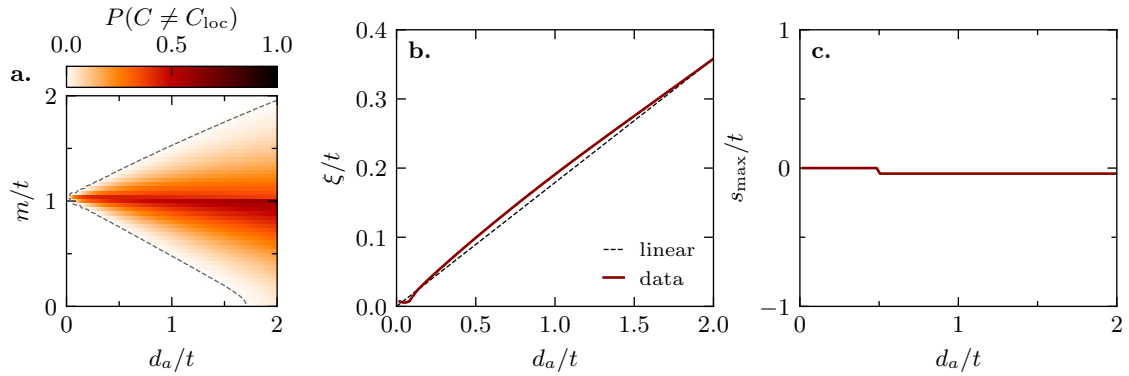

Figure 6.17: Location of the maximal probability as a function of d_a. **a.** We show for reference again the probability of change at $z = 0$. **b.** Length scale ξ of the exponential decay as a function of d_a. For the values studied, ξ is approximately proportional to d_a, as demonstrated by the linear fit. **c.** s_{max} (distance of the maximal probability from the local transition) extracted from the data. $s = 0$ corresponds to the location of the local transition. For $d_a/t \in [0, 2]$, which basically covers the entire important range, s_{max} is close to 0.

10 in the tails of the distribution. We have verified, though, that the exponential decay of the probability remains regardless of the specific sampling procedure used. Since with our definition the uncertainty region is defined to contain only points above a threshold of $P_0 = 10^{-3}$, points outside are still allowed to have $10^{-3} \times n_{\mathrm{samples}} = 1000$ cases of changed Chern numbers. We therefore rephrase our earlier statement that the Chern number can change only within a region of width w slightly. In fact, the Chern number is most likely to change only close to the local phase transition and the probability for a change decays exponentially with the distance to the transition on a length scale $\xi \propto d_a$.

It turns out that we have automatically adopted a Bayesian perspective, cf. Sec. 4.3.1, here, where we assign not a single most likely location to the phase transition, but instead a confidence interval, which is described through a probability distribution obtained as a normalized probability of change. This means that while we still acknowledge that the exact location of the topological phase transition is unknown, we can assign a probability[2]

$$P(\gamma|\gamma_{\mathrm{loc}}) = \frac{P(\gamma_{\mathrm{loc}}|\gamma)P(\gamma)}{P(\gamma_{\mathrm{loc}})} \tag{6.127}$$

to any path γ while knowing γ_{loc} as a marginalization over all possible paths. In Fig. 6.18**a** we illustrate this in terms of a m–U phase diagram, where the confidence interval of the predictive posterior distribution $P(\gamma|\gamma_{\mathrm{loc}})$ is indicated as a shaded region around the local transition, which would correspond to the maximum likelihood solution. Due to the exponential decay of probabilities with increasing distance perpendicular to the local transition, paths like γ_1 that deviate less are more likely. Since the self-energy dispersion amplitude vanishes at $U = 0$ and grows as a function of U, the confidence region broadens towards the right. The predictive distribution for fixed U as a function of m is illustrated in Fig. 6.18**b**. In contrast to the usual assumption of a Gaussian prior we have shown here that we are, in fact, dealing with an exponential distribution that has an exponent $1 \lesssim \nu \lesssim 2$. Using the identity

$$\int_0^\infty e^{-\left(\frac{s}{\xi}\right)^\nu} \mathrm{d}s = \frac{\xi}{\nu} \int_0^\infty t^{\frac{1}{\nu}-1} e^{-t} \mathrm{d}t = \frac{\xi}{\nu}\Gamma(\nu^{-1}) = \xi\Gamma(\nu^{-1}+1), \tag{6.128}$$

[2]Note that this interpretation is meant figuratively, since we measured the "posterior" immediately and did not have to optimize a suitable hypothesis.

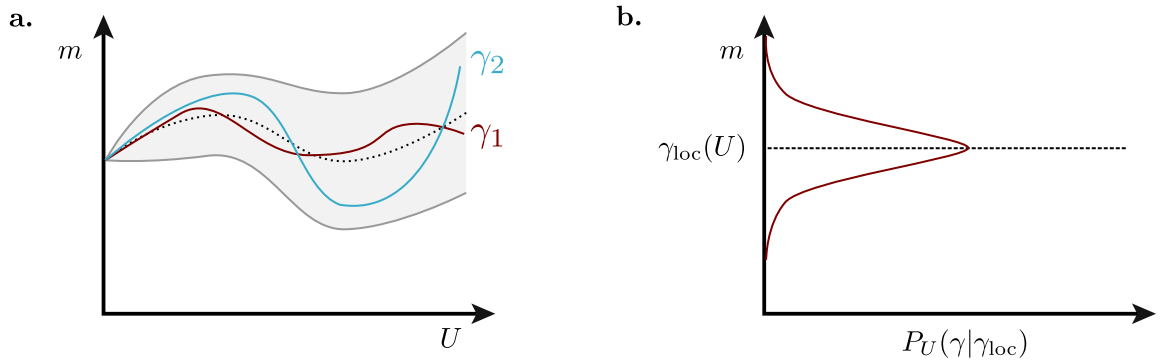

Figure 6.18: a. Illustration of the confidence region around the local phase transition $\gamma_{\rm loc}$ (dotted), in which we expect the exact phase transition to lie. Since the self-energy dispersion amplitude increases with U for weak to intermediate coupling and the length scale ξ is proportional to d_a, the confidence interval becomes broader towards larger U. We draw two possible solutions $\gamma_{1,2}$. **b.** Probability distribution along the m-axis, which is an exponential distribution centered approximately at $\gamma_{\rm loc}$. Given this distribution, the posterior of γ_1 in **a.** is larger than that of γ_2, since it deviates less from the local transition.

where

$$\Gamma(\nu) = \int_0^\infty x^{\nu-1} e^{-x} \mathrm{d}x = \frac{1}{\nu}\Gamma(\nu+1) \tag{6.129}$$

is the gamma function, we obtain the probability distribution function

$$\rho(s) = \frac{1}{2\xi\Gamma\left(1+\frac{1}{\nu}\right)} e^{-\left(\frac{|s|}{\xi}\right)^\nu}. \tag{6.130}$$

The cumulative distribution function is then obtained as

$$F(s) = \int_{-\infty}^{s} \rho(x)\ \mathrm{d}x = \frac{1}{2\Gamma\left(1+\frac{1}{\nu}\right)} \int_{-\infty}^{s/\xi} e^{-|s|^\nu}\ \mathrm{d}s, \tag{6.131}$$

which can be solved numerically. In Fig. 6.19**a** we show the cumulative distribution function for $\nu = 1.3$ and for comparison $\nu = 2$, which corresponds to a Gaussian distribution. The mean vanishes due to the mirror symmetry of $\rho(s)$ w.r.t. $s = 0$. We now compute the variance

$$\mathrm{Var}[s] = \langle s^2\rangle_\rho - \langle s\rangle_\rho = \int_{-\infty}^{\infty} s^2\rho(s)\mathrm{d}s \tag{6.132}$$

$$= \frac{1}{\xi\Gamma\left(1+\frac{1}{\nu}\right)} \int_0^\infty s^2 e^{-s^\nu}\mathrm{d}s \tag{6.133}$$

$$= \frac{\xi^2}{\nu\Gamma\left(1+\frac{1}{\nu}\right)} \int_0^\infty t^{\frac{3}{\nu}-1} e^{-t}\mathrm{d}t = \frac{\xi^2}{\nu\Gamma\left(1+\frac{1}{\nu}\right)}\Gamma\left(\frac{3}{\nu}\right) \tag{6.134}$$

$$= \frac{\xi^2\Gamma\left(1+\frac{3}{\nu}\right)}{3\Gamma\left(1+\frac{1}{\nu}\right)}, \tag{6.135}$$

which yields the standard deviation $\sigma = \sqrt{\mathrm{Var}[s]} \propto \xi$. σ/ξ is shown in the inset of Fig. 6.19**a**, where we see that the standard deviation for $\nu = 1.3$ is very close to 1, i.e., $\sigma \approx \xi$. To better

illustrate the meaning of this result we plot in Fig. 6.19**b** the probability distribution function and mark the $\xi, 2\xi$ and 3ξ regions. The coverage is rather similar to the Gaussian distribution and we find that $\approx 99.1\%$ of samples lie within 3ξ of the local transition. Comparing this to the earlier result, where we found that $\xi \approx 0.2 d_a$ and more specifically $6\xi \approx 1.2 d_a$, see Fig. 6.17, the estimate of a width $w \approx d_a$ corresponds roughly to a 2.5-sigma confidence interval covering $\approx 97.7\%$ of cases.

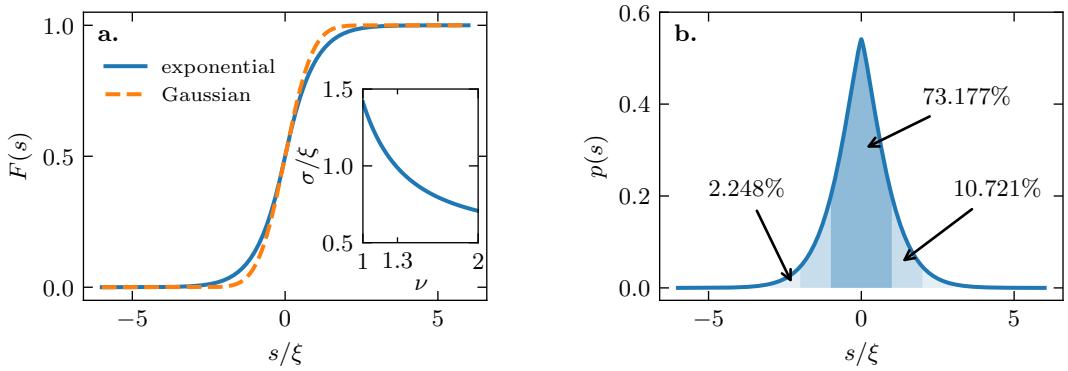

Figure 6.19: a. Cumulative probability distribution function (Eq. 6.131) for the exponential distribution with $\nu = 1.3$ and for comparison a Gaussian with $\nu = 2$. In the inset we show the standard deviation σ as a function of the exponent ν. Larger ν increases the coverage in an interval of width $n\xi$. **b.** Probability distribution function for the exponential distribution with $\nu = 1.3$. We mark the $1\xi, 2\xi$ and 3ξ regions and indicate the coverage gained. The 3ξ interval covers $\approx 99.1\%$ of cases.

6.6 Conclusion

We conclude this chapter by summarizing the main results. Inspired by the stark contrast of results for the phase diagram of the Haldane-Hubbard model provided by different numerical methods we investigated the effect of different perturbations to the Hamiltonian, which appear naturally in the context of the topological Hamiltonian. By decomposing the self-energy into a local (k-independent) and explicitly non-local (k-dependent) part, we found that the local part of the self-energy primarily shifts the mass term in the non-interacting Hamiltonian towards positive m. This is due to the expected larger magnitude of the diagonal vs. off-diagonal matrix elements of the local self-energy. Astonishingly, this already lays out the general topology of the phase diagram in terms of a positive slope of the CI to BI transition line. This had already been mentioned in Ref. [140]. The off-diagonal part on the other hand leads to the opposite effect. By destabilizing the topologically nontrivial phase, the transition line is effectively shifted towards smaller m. We have studied this effect systematically and found that this instability is most significant close to the phase boundary, while farther away the topological phase proves to be rather robust towards these perturbations.

Lastly we conducted a systematic study of the effects of explicitly non-local perturbations via a statistical study. We found that in essence the topological phase is rather stable. A parameterization of the non-local part in terms of the self-energy dispersion amplitude reveals quite generally that the Chern number is, in principle, given by the local Chern number neglecting all non-local terms in the self-energy plus an integer-valued random variable that is exponentially suppressed as a function of the distance to the local transition. We have performed fits to the data and used the extracted exponent to obtain in a quantitative analysis a confidence interval that should be rather independent of the specifics of the sampling procedure. Comparing typical values of d_a with our confidence intervals we conclude that the "true" topological

phase transition lies very close to the local transition throughout the largest part of the phase diagram. At large U and small m, where the momentum-dependence becomes rather strong, a larger deviation is statistically possible. However, we still expect the influence of the local part through varying values of the magnetization to be a stronger source of error.

Chapter 7

Engineering Topological Phases

In this chapter we present an application of a statistical method in a different context. Besides the study of correlation effects and, in particular, the description of systems where local theories seem to fail in capturing all relevant properties, there is another area of research that caught our interest, namely the problem of finding or artificially manufacturing materials with topological properties.

Historically, a large proportion of the common literature focused on models—e.g., the Haldane, Hofstadter, Kane-Mele, Bernevig-Hughes-Zhang models to name a few—rather than realistic materials. Notable exceptions are, e.g., the proposal and experimental confirmation of a realization of the quantum spin Hall effect in HgTe quantum wells [144, 178] and other recent developments like the proposal of a higher-order topological insulator in Bismuth [179]. On the other hand, a lot of research is focused on the description of existing materials and the characterization of their properties that often reveals candidates for interesting physics by chance. Here, we follow along one avenue that aims to combine the two sides using machine learning and statistical methods with the goal of eventually predicting new candidates for topological materials or ways to engineer them.

This chapter is organized in the following way: We start with a short introduction of the state-of-the-art in machine learning applications to condensed matter physics and the prediction of topological candidates, and explain the general motivation for our statistical approach. In Sec. 7.2 we link the prediction of topological properties to the understanding of phase diagrams in high dimensions and then investigate what type of information exactly we should require our method to provide, by defining an abstract sequence of tasks that need to be accomplished. After introducing the concept of interpretability (which constitutes a weakness of most generic machine learning approaches) in Sec. 7.3, we evaluate the possibility of applying clustering algorithms to understand topological phase diagrams in Sec. 7.4. The remainder of the chapter then revolves around supervised approaches that operate on data sets that we generate in a particular systematic way. This process is documented in detail in Sec. 7.5.

Taking one of our data sets as an example, we try once again to apply a generic machine learning algorithm to investigate what type of information we are able to learn realistically with these standard methods. This time, we train decision tree and random forest models. The complete analysis is documented in Sec. 7.6.

In Sec. 7.7 we then introduce our statistical approach as an alternative to traditional machine learning methods and show that the same information (and possibly more) can be extracted directly from the data set without the necessity to train a model first. The method is built around the example of the Haldane model that we systematically reconstruct from the data as the prototypical topological model on the honeycomb lattice.

With the honeycomb lattice merely serving as the test bed for the conception of the method-

ology, we demonstrate the power of our method in Sec. 7.8, with an application to the kagome lattice. We show that our statistical approach allows us to learn a type of phase diagram that enables us to understand the relationship between complex hopping-configurations in a high-dimensional parameter space and the topological phase on a qualitative level. Here, the framework of tight-binding parameters defines a language that is universally understood and therefore promises to enable new discoveries by bringing together experiences from theoretical and application-oriented experts alike.

The chapter closes with additional remarks about possible adaptations of the method to the description of interacting systems that goes beyond the discussion of Chapter 6, and straightforward applications to real materials in combination with ab-initio calculations.

A large part of this chapter describes our evaluation of different methods and can therefore distract from the main message. For readers only interested in the statistical method we recommend to focus on Sections 7.2, 7.7 and 7.8.

Parts of the results discussed in this chapter were published as Ref. [173]:

Thomas Mertz and Roser Valentí
Engineering topological phases guided by statistical and machine learning methods
Phys. Rev. Research **3**, 013132 (2021)

7.1 State of the Art

We summarize here shortly the current state of the art in topological phase discovery with and without machine learning. For more information on applications of machine learning to solid state systems we recommend the very detailed review by J. Carrasquilla [180].

The current surge of interest in machine learning or artificial intelligence is mainly driven by the advances in deep learning [181], which has been enabled through careful optimization of neural network design and of course the developments in hardware-accelerated algorithms [182]. While the prediction through a neural network essentially requires a lot of linear algebra calculations that can be performed rather efficiently—especially on modern graphics processors—the same cannot be said about most algorithms in computational physics, where problems are usually non-linear and require iterative procedures that are numerically very costly and knowledge of one particular solution does not necessarily translate to similar problems. Therefore, one approach that aims at benefiting from the efficiency of AI-driven methods is the attempt to accelerate certain often repeating calculations by offloading them at least in parts to neural networks and other machine learning models [183–192]. In particular, in variational wavefunction methods, complicated machine learning models have also proven to yield lower-energy solutions than the traditional physically motivated models, albeit at the cost of physical intuition and interpretability [193].

On the other hand, there are problems that are not yet well-understood. Especially through unsupervised learning it is conceivable that discoveries could be made via the search for hidden patterns in large data sets ("big data")—a task that is too unwieldy to be done by hand. So far, most progress in this field has been devoted to setting the stage, i.e., proving that machine learning methods can indeed be used to discover and explain new effects. This is done by studying certain known cases and demonstrating that without helping the algorithm too much,

i.e., with as little bias as possible, a result compatible with the known physical laws is returned. For example, unsupervised learning algorithms have been shown to be capable of predicting non-trivial properties such as the existence of a phase transition in the 2-dimensional Ising model and the identification of the magnetization as the corresponding order parameter [194]. Similar ideas have recently been applied also to the XY-model, where instead the presence of topological vortices was detected as the most dominant feature in the data [195–197]. Another successful approach that requires using a small amount of prior information about the nature of topological phases has recently been demonstrated to be capable of identifying distinct topological phases based on the properties of the Hamiltonian [198]. In addition, by using deep learning on input Hamiltonians it has been shown that complex models can capture the physics behind even complicated order parameters like the Chern number [199–202].

Another interesting direction is the evaluation of experimental data sets. For a long time, insights were obtained by fitting model functions to raw data via standard optimization methods. While these methods are still relevant, they are now increasingly complemented with artificial intelligence, which promises to achieve predictability through the generalization property inherent to many machine learning models that is not present in more specialized model functions. Researchers across many fields are now starting to apply these more refined methods [203, 204].

Developments towards the discovery of new candidates for interesting topological phases are branching out in many different directions. A large cluster of attention is, e.g., focused on methods related to the topological quantum chemistry [42, 205, 206], which makes predictions for the presence or absence of possibly topological band crossings based on the spatial symmetries of the lattice. Given a particular symmetry group, the classification reveals whether topological states can be found or not. Since classifications are tabulated for all symmetry groups, this is a very convenient way to narrow the search down to only relevant space groups. On the other hand, the realization of a topological phase still depends strongly on the actual values of parameters, which then requires substantial computations. In addition, the theory makes no statement about the stability of topological phases if spatial symmetries are broken. In real materials, we always expect the perfect symmetry to be broken, if only slightly. An AI-driven scheme to identify interesting materials has also recently been proposed in the context of superconductivity [207], which could possibly also be applied to identify topological materials.

One of the main obstacles to the search for topological materials is that many properties are known primarily from the study of models that do not necessarily have a one-to-one correspondence in actual materials. Usually, the lattice type and symmetries dictate which kind of topological model may correspond to a given crystal and via mapping of the band structure to tight-binding parameters one hopes to land somewhere in the topological region of the corresponding phase diagram.

The most straight-forward way to search for topological properties would be to simply compute the topological invariants for all thinkable materials and catalog the results. However, since the required density functional theory calculations are far too costly numerically to allow for such large-scale surveys, one has to rely on alternative methods. Fortunately, many materials are already well-studied and the corresponding data is known to at least a small number of experts. One can therefore make educated guesses by, e.g., probing specifically materials that are known to feature Dirac points close to the Fermi surface. This approach is a rather old-school method and the chances of success are limited primarily by the experience of the researcher. Any expert can only know a limited number of different compounds, which yields a data set of $\mathcal{O}(10)$. Given only such a small amount of data to work with, the chances for success are likely not vastly better than simple trial and error, which has been realized also by the community. Assuming that one is not in the comfortable situation of possessing such broad knowledge, one can instead canvass existing crystal databases and the information therein to scan for possible

candidates [208–211]. This corresponds in a way to making use of the combined experience of the entire community and is enabled by machine learning techniques that allow high-throughput evaluations of data.

More interesting literature regarding applications of both supervised and unsupervised machine learning techniques to the field of topological phases can be found in Refs. [212–219].

The idea that we are investigating in this chapter is to apply the general concept of a statistical prediction method that we introduced in Chapter 6 to the discovery of topological phases, with the aim of bridging the gap between models on one hand and realistic materials on the other. To this end, however, we first need to develop an understanding of the type of information we ask the method to provide.

7.2 Understanding What We Understand

We set out to investigate possible ways to engineer topological phases, eventually, of course, with a practical application in mind that can help with the selection of candidate materials and other experimental parameters. This task can only be achieved if some kind of information is procured that can help our understanding of the underlying physical mechanisms. We neglect here the notion of a machine that produces exact predictions of particular chemical compounds and appropriate external conditions under which a topological phase is found. Albeit being entirely possible in principle, such a machine would require an insurmountable quantity of computational power that cannot be produced realistically—unless, of course, one is satisfied with an answer such as "42". We intentionally draw an analogy to the popular science fiction novel "The Hitchhiker's Guide to the Galaxy" by Douglas Adams [220] here, since the question about the meaning of "life, the universe and everything" is equally ambiguous. Setting this fantasy aside, it turns out that the task we are trying to complete is not particularly well-defined, since we lack a notion of the type of information that we seek to obtain. This definition of desired information and with that the question that we ask our machinery to answer should generally be in line with what is realistically achievable. In this section, we elaborate on this problem and motivate the choices we made for the following analysis.

In order to build a machine learning setup that can assist in the prediction and subsequently engineering of topological materials, we have to first get a grasp of what question exactly the algorithm should answer. We begin this discussion with topological models for which phase diagrams are known and analyze their information content. The successful principle behind the use of phase diagrams can be boiled down to a reduction in complexity that allows us to easily grasp complex relationships between physical parameters that are somehow related to an intuitive understanding of the crystal structure on one hand, and properties that arise from the solution of the corresponding Schrödinger equation on the other. Given a typical tight-binding model, the number of parameters is limited to only a handful of parameters such that a phase diagram can be drawn easily by varying the parameters independently and mapping out the entire phase space.

This concept seems simple enough and we have seen in Chapter 6 that the Haldane model can be understood in terms of a phase diagram as a function of four parameters m, t_1, t_2, ϕ. By using one parameter as the scale we are left with three parameters, e.g., $m/t_1, t_2/t_1, \phi$, which allows for a graphical representation of the physical properties described by different configurations in terms of a three-dimensional phase diagram by drawing the phase transition according to Eq. 6.51. The resulting phase diagram is shown in Fig. 7.1. The non-trivial phase is enclosed by the gap closing surface, where we differentiate the $C = +1$ and $C = -1$ phases in color. We can recognize the main features, i.e., the $C = +1$ phase can only be found for $0 < \phi < \pi$, while for the $C = -1$ phase we have the restriction $\pi < \phi < 2\pi$, assuming that ϕ is expressed modulo

2π. Given a value of ϕ, we can then express a constraint on the remaining two parameters such that they lie within one of the enclosed volumes. Clearly, $m/t_1 \gg t_2/t_1$ will produce a point on the outside, which means that this is a property of the trivial phase.

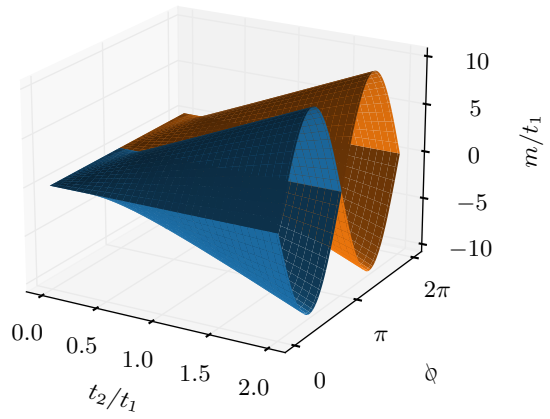

Figure 7.1: Three-dimensional representation of the Haldane phase diagram in terms of all four parameters. t_1 is used as a scale, the other parameters t_2, m, ϕ define a coordinate system, in which we draw the surfaces where the band gap closes according to Eq. 6.51. The topologically trivial phase lies above (below) the surface for $m > 0$ ($m < 0$). The non-trivial phase is enclosed by the gap-closing surfaces and we use the colors blue and orange to differentiate between the two distinct phases $C = +1$ and $C = -1$, respectively.

We note that three-dimensional representations are often inferior to a collection of two-dimensional projections or slices. This is primarily rooted in the dimensionality of the media we work with, which allow for an easier integration of two-dimensional plots. However, also the fact that our perception in general is limited to three spatial dimensions favors a lower-dimensional choice for graphical representations of data, since this allows us to assume an ideal observer's role with every single point in view at the same time. In general, given a d-dimensional world we can assume this role for graphical illustrations of $d' < d$ dimensions, since viewed from the remaining dimension, the lines connecting our eye (observer) with the source of the reflected light (data) will never cross. The proof is straight-forward, since given the definitions

$$v_{\text{data}} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_{d-1} \\ 0 \end{pmatrix}, \qquad v_{\text{obs}} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \quad \Rightarrow \quad v_{\text{obs}} - v_{\text{data}} = \begin{pmatrix} -v_1 \\ -v_2 \\ \vdots \\ -v_{d-1} \\ 1 \end{pmatrix}, \tag{7.1}$$

every "light ray" originating from different data points also has a different direction. Since all rays terminate at the observer's position they intersect there, which excludes the possibility of any further intersections. This has important implications in the context of phase diagrams. Provided that we cannot increase the number of spatial dimensions available to us we can only achieve a perfect overview over two parameters at the same time. Adding the time dimension we can increase this number to three, however, at the cost that we are now observing data at different times which somewhat limits our ability to directly put data points into relation with one another. Clearly, everything beyond three parameters becomes extremely difficult to understand.

Following this argument, the diagram shown in Fig. 7.1, albeit comparatively simple, does not reveal the entire physics at a glance without the accompanying equations that were used to generate it. Nevertheless, we are often satisfied with the information conveyed by just this simple illustration. We tend to simplify the information contained within to more memorable features. In this case, the information that one remembers can be

1. large m/t_1 is characteristic for the trivial phase regardless of the other parameters,
2. $0 < \phi < \pi$ only allows $C = 0, +1$, while $C = 0, -1$ are found for $\pi < \phi < 2\pi$,
3. the robustness of either topological phase increases (linearly) with t_2/t_1,
4. $t_2 = 0$ or $\phi \in \{0, \pi, 2\pi\}$, i.e., real next-nearest neighbor hopping, implies a metallic phase and, in particular, the absence of a topological phase.

This listing is not necessarily complete and, of course, the exact solution from Eq. 6.51 contains more details. Nevertheless, the more general qualitative information above is already enough for an understanding of the underlying relationships between parameters and the topological phase. Any scientist provided with this information can immediately apply this knowledge to make predictions for possible realizations of the Haldane phases in materials.

On the other hand, realistic systems are often much more complicated and since all models are valid only through the tight-binding approximation, one never finds an exact correspondence to a particular model. Nevertheless, understanding the interplay between parameters on a qualitative level can help in developing an intuition also for more complicated systems.

In light of these considerations, we find it justified to restrict our expectations to a qualitative description of the interplay between parameters and the topological classification, with the option to possibly extend this to a more quantitative refined approach later. We have already identified that our understanding works best for low dimensionality of the data, which allows for an easily memorable graphical representation. Hence, we suggest that an attempt to model or automate and support this understanding can in principle be decomposed into several individual steps that we illustrate in Fig. 7.2. Starting from a usually high-dimensional input data set that has

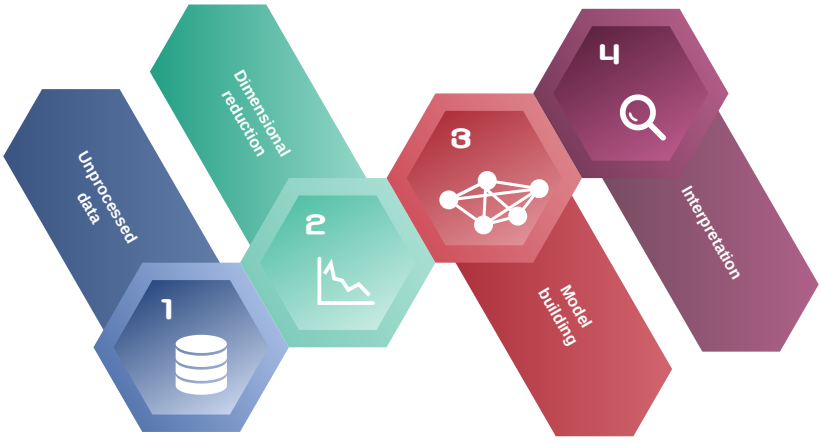

Figure 7.2: Decomposition of the process of understanding. Unprocessed high-dimensional data is used as the input. On this data, dimensional reduction has to be performed to make the results more comprehensible. Model building corresponds to the evaluation and perhaps graphical preparation of the relations between parameters. The result can then be inspected and the information gained would ideally be applied towards the prediction of new physics.

to be obtained in step 1 we try to first fish out those parameters or combinations thereof that

are most valuable for the description of the data with the goal of getting closer to the realm of understandable inter-parameter relations. The task of selecting parameters is called dimensional reduction and corresponds to step 2. In step 3, one would then make an attempt to build a model that describes the data and in the optimal case holds predictive power. The final step is then the observation of the result and interpretation of the model.

We note that this is more or less the way any machine learning project is laid out. However, there are a couple of intricacies that differentiate our concept from a large number of machine learning applications. While it is true that dimensional reduction is often used, the reasoning is entirely different here. Usually, one tries to reduce the dimensionality of the data for the purpose of lowering the computational complexity of the following learning phase and at the same time increasing the success rate for such a task, since the size of the underlying optimization problem is greatly reduced. In other words, by reducing the amount of parameters one can use a less complicated model that can be trained more efficiently and due to the removal of extra dimensions one is more likely to end up with an optimal solution. Our philosophy is different from that approach in that we strive for the lowest possible dimension for the sole purpose of being able to understand the final result, which is more in line with the field of data analytics. As we have seen before, the complexity of grasping complicated relations increases significantly beyond three parameters. Performing this step prior to the model building phase allows us to possibly get a better understanding of the data set itself, which might even enable us to skip phase 3 and immediately move to the interpretation of the data.

For the model building phase, which corresponds to the generation of an automated understanding of the relations between parameters and topological phases, we rely—in contrast to the contemporary trend towards deep learning—on simple models that allow for a reconstruction of the learned information. We discuss the merits and problems of this approach later in the context of "interpretability".

We stress here that according to the above considerations, a machine learning application focused on the reproduction of the Chern number as presented in, e.g., Ref. [200] does not necessarily help our own understanding of the distribution of the topological phases over the parameter space, as it simply constitutes an alternative way of computing the topological invariant and can at best be used for improved performance compared to conventional algorithms. New

7.3 Interpretability

In the context of machine learning and artificial intelligence, we believe it is important to also address the topic of interpretability [221–224] of the applied models—in particular, pertaining to scientific applications, where the generation of knowledge is the primary concern. The term "interpretability" refers to the amount of transparency of a machine learning model to a human observer. It encodes how likely it is to comprehend the decision-making process of the machine. We strongly recommend the book by Christoph Molnar [225] that provides a detailed overview over the state of the art of improving the understanding of complicated black box models.

The most impressive industrial applications of machine learning enable complex computations from pattern recognition in computer vision to natural language processing or automated driving. Realizing these concepts with traditional algorithms would require an enormous amount of rules to be implemented, possibly yielding highly unstable software, whereas the machine learning model learns such rules by itself if it is provided with suitable data. For these applications, in particular, if matters of safety are involved, it would be essential to know upon which rules exactly the model bases its decisions, however, for the completion of the task at hand this is not necessary.

This is in stark contrast to the scientific application, where we are mostly interested in how

the decision is made and the result is of lesser relevance since it could in most cases also be obtained in another more direct way. In order to explain why this is, we look at two different cases, where i) the decision rule is already known, and ii) it is not known. In case i), we clearly do not need to extract the rule since it is already known from the start. Such an undertaking can therefore only be motivated by the desire to deliver a proof of principle or increase the productivity of a repeated task by exploiting the efficient evaluation of trained machine learning models. In the scientific context, especially the latter is very tempting, since many common computational tasks are very difficult to solve. However, this temptation fades in comparison to case ii), where machine learning offers ways to deliver new insights and to help with making new discoveries. One such avenue, where this aim is taken quite literally, was explored recently by Krenn and Zeilinger, who developed a semantic neural network engine that processes published literature and uses the information therein to generate new research ideas that align with the current interest in the field [226]. In maybe all cases that promise new scientific discoveries, the result obtained by the machine learning model will be less important than the reasoning behind its construction. In particular, in the field of topology, machine learning methods could assist in the discovery of new order parameters, since finding structure in data is literally the reasoning behind unsupervised learning. A demonstration of this would, e.g., be the aforementioned discovery of the spin structure factor as an order parameter of the phase transition of the Ising model [194].

The new discovery of a topological phase is of theoretical value only if it can somehow be understood how it is defined. Simple cases like the Ising model can be explained through the application of the principal component analysis (PCA), which also delivers an expression for the order parameter. More complicated systems, however, require more complicated methods where the comprehension of the learned information becomes rather difficult. The discussion of interpretability begins right there. The popularity of neural network models is mainly fueled by their general applicability to many different problems—as opposed to, e.g., linear models. However, the size of the model itself, i.e., the number of weights, grows with the complexity of the problem it is supposed to explain. Hence, it is not unthinkable that once a model is trained such that it describes the data set very well, the complexity of the data has been merely transformed into the complexity of the model. For generic models this is, in fact, very likely to happen. A reduction in complexity on the other hand requires careful selection of features such that effectively a compression of the original data is achieved. This can be thought of as a non-linear basis transformation.

We will not go into detail about approaches that try to alleviate this fundamental problem of complex models such as deep neural networks. Such approaches are, e.g., discussed in Ref. [225]. Instead, we discuss in the following section a different model that is more interpretable by design.

7.4 Unsupervised Learning Approaches (Clustering)

It is often advisable when developing new ideas to start small and focus on well-known cases for prototyping. We follow this logic here and try to somewhat automate the generation of understanding for topological models that are already limited to relatively few parameters before moving on to more complicated systems. In this section we will focus on unsupervised learning methods and motivate that they are ideal candidates for phase 2 of Fig. 7.2, i.e., dimensional reduction. We will discuss several possible algorithms, how to apply them and investigate their pitfalls.

The aim that we want to accomplish in phase 2 is dimensionality reduction, which can be achieved in a number of ways. Typical examples are clustering, where subsets of points are assigned cluster indices based on their distribution in the feature space, usually taking into

account a particular distance metric [227–229]. Assuming the number of clusters is much smaller than the initial data set size, the algorithm can successfully reduce the dimensionality of the data set by condensing the number of data points to a smaller number of characteristic points:

$$X \in \mathbb{R}^{n,m} \mapsto \tilde{X} \in \mathbb{R}^{l,m} \tag{7.2}$$

with $l \ll n$.

On the other hand, there exist algorithms such as the Principal Component Analysis (PCA) [230, 231], whose principle of function is the reduction of m, i.e., the size of each individual data point. This is achieved by determining a transformation f of data points X of the form

$$f : \mathbb{R}^{n,m} \to \mathbb{R}^{n,p} \tag{7.3}$$

with $p \ll m$. In the case of PCA, the map f is a linear transformation of the data matrix, however, more involved algorithms such as autoencoders [232, 233] can be used to represent non-linear transformations.

Clustering

In the context of data science, clustering is seen as a way to extract information out of structured data. When applying machine learning or data science methods to physical systems, one critical choice one has to make is what type of data to use as an input for the algorithm. We will focus here on the most general input of system parameters that define the Hamiltonian since a solution of this generality would be outfitted with highly predictive capabilities and also be trivially interpretable. By choosing as features the parameters of the physical model we address the problem of finding relations between them directly, which greatly simplifies phase 4 in Fig. 7.2. Other approaches exist, though, that are highly successful in characterizing topological phases by, e.g., choosing the eigenstates of the Bloch Hamiltonian as the input data [198, 217, 234]. Being very specific in the choice of the input data, i.e., skipping the step of constructing the Hamiltonian and diagonalizing the corresponding matrix, reduces the demands placed upon the following machine learning task, however, at the same time the predictive power of the resulting machine is rather limited, since the mapping $t \to \{|\psi\rangle\}$, where t stands for the model parameters and $\{|\psi\rangle\}$ for the corresponding eigenstates, is typically not invertible. Hence, acquiring an understanding of the relationships between eigenstates does not necessarily provide the same in terms of the system parameters and predictions $\{|\psi\rangle\}$ made by the machine might not be trivially representable in terms of a simple model.

We briefly remind ourselves that unsupervised learning, in contrast to supervised learning, operates on unlabeled data, i.e., data points are coordinates only. These data points can be thought of as a data matrix $X \in \mathbb{R}^{n\times m}$, where n is the number of data points and m the number of features, i.e., the dimension of each data vector. This means that one has a collection of data points whose distribution should be analyzed by the machine learning algorithm. Clearly, without assigning a label to the data, differentiation between different classes of points can only happen in several ways that all depend on a form of distance metric. Points that are close to each other should belong to the same class, while points at larger distances are less likely to be connected. Assuming that this approach works, we can expect it to produce a sort of characterizing feature, i.e., a representative data point, that describes the entire class of connected (equivalent) data points. This sounds very promising, since it corresponds essentially to the mathematical concept of equivalence classes, upon which the topological classification is based.

Assuming that a representative for each particular equivalence class can be found, we then obtain a description of the data set in terms of l representative parameter configurations x_i,

$i \in \{1, \ldots, l\}$, where l is the number of equivalence classes present in the data set, and the distance metric used in the clustering algorithm. Apparently, not only will we be able to classify additional points by computing their distances to all x_i and choosing that class, which produces the smallest value, but we will also have an idea of the characteristics of each class encoded in the representative configuration. Ideally, this would allow us to develop a deeper understanding of the particular relationships between parameters in the individual topological phases. We illustrate this in Fig. 7.3, where a large data set of dimension $n \times m$ is mapped to a small set of l m-dimensional vectors, where $l \ll n$, and a distance function. In the optimal case that the learning algorithm achieves 100% accuracy, the right hand side would contain the same information as the original data.

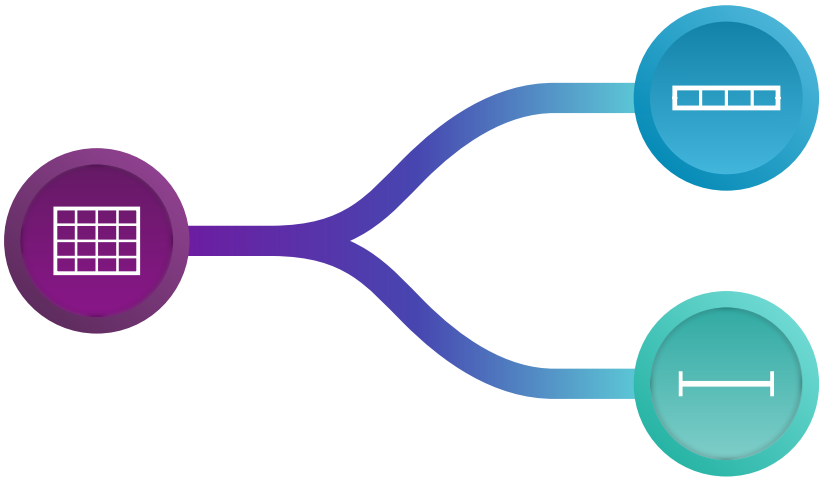

Figure 7.3: Illustration of the mapping performed by the clustering algorithm. The large data set (left) is mapped to a small number of representatives (top right) that are of the same size as individual data points and a distance function (bottom right) that contains information about how other points in the same equivalence class relate to the representative.

Apparently, clustering works as a kind of data compression, i.e., the formerly verbosely stored data containing a lot of redundant information is mapped to a more optimal form that requires much less storage and is also much easier to understand. The output data contains a small number of representatives in the form of m-dimensional vectors, i.e., the clustering algorithm maps $\mathbb{R}^{n \times m} \to \mathbb{R}^{l \times m}$ with $l \sim \mathcal{O}(1)$. The dimensionality of the result in terms of the number of features m seems to be unchanged, however, given that each class is described by a single representative, one can obtain relations between the parameters from the coordinates of the representatives. Hence, while the topological classification results in an enormous dimensional reduction of a data set of size $n \times m$ to just a set of size $l \sim \mathcal{O}(1)$ containing the class labels, the problem of describing the relations between parameters requires a further analysis of the components of the representatives in combination with the distance mapping. Therefore, a clustering algorithm could be employed in phase 2 of our schematic workflow from Fig. 7.2.

An ideal algorithm would then assign labels $\lambda \in \Lambda \subset \mathbb{Z}$ to each of the data points based on the topological class they belong to, with there being a bijection g between the topological class index $c \in C \subset \mathbb{Z}$ and the cluster label

$$g : \Lambda \to C, \lambda \mapsto c. \tag{7.4}$$

We now look at constraints for cases where this ideal solution cannot be found, i.e., there is no optimal bijective solution for g or when the problem of unsupervised learning in terms of clustering is fundamentally ill-defined.

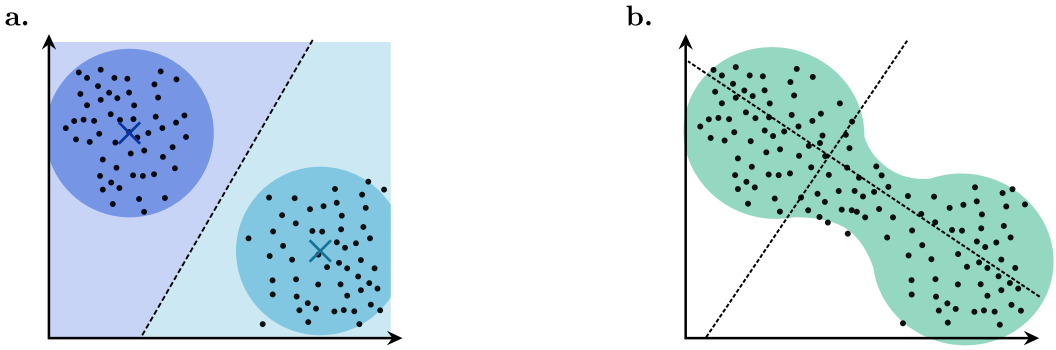

Figure 7.4: We show two different data topologies. **a.** Two clusters of data (black points) can be separated from each another based on the distances between data points. Clusters are marked schematically by colored circles, with their centers or representatives marked as "×". Every point below the dashed line belongs to the lower cluster due to the smaller distance to its center and vice versa. **b.** There is no separation between clusters and almost any line can be used as a separation, provided there are no further constraints.

Clearly, if a distance measure is applied to distinguish topological phases from one another then this algorithm can only succeed if there is a significantly large separation between data points corresponding to any two phases. We illustrate this in Fig. 7.4**a**, where we show two well-defined clusters with a gap in between, into which a well-defined transition line can be placed. In contrast, the case shown in Fig. 7.4**b** features only one connected region, where (given no further information) all transition lines are thinkable solutions. For centroid-based methods, as illustrated here, the distance w.r.t. the cluster center is what determines the affiliation of points to clusters. Assuming that there is a gap (absence of data points) in the data of size w, we have for points x_a, x_b belonging to different clusters with centers c_a, c_b

$$d(x_a, x_b) \geq w, \tag{7.5}$$

where d is the distance function. It then follows immediately for points lying on the line $x_t = tc_a + (1-t)c_b$ for $t \in [0,1]$ that the distance $d(x_t, c_a)$ increases discontinuously when moving from one cluster to the other. In the opposite case, where such a discontinuity does not exist, however, the location and number of clusters is entirely ill-defined.

We can formalize the requirement for a well-defined clustering problem in terms of the metric d as $\exists c_1, c_2 \in \mathbb{R}^m, w > 0 \in \mathbb{R}$:

$$\begin{aligned} d(x_i, c_1) + w < d(x_i, c_2) \quad & \forall x_i \in X_1 \subset X, \\ d(x_i, c_1) > d(x_i, c_2) + w \quad & \forall x_i \in X_2 \subset X. \end{aligned} \tag{7.6}$$

Eq. 7.6 expresses the condition that at least a pair of centroids, i.e., centers of clusters, must exist that allow for a separation of the data set X into subsets X_1, X_2, where the additional separation w describing the size of the discontinuity in the distances is a measure of the quality of the clustering solution. How successful this is depends on the choice of d in addition to the composition of the data. Assuming well-separated data points a mean-based clustering, such as the k-means algorithm [132, 235, 236], can be successful. In the case of Fig. 7.4**b**, however, this will certainly fail to deliver a satisfying solution.

We demonstrate this by performing an actual k-means calculation on two random data sets in Fig. 7.5, where we have two features x_1, x_2 and assume two clusters. In Fig. 7.5**a**, the clusters are well separated and the algorithm succeeds in finding a suitable line. All points corresponding

to the same cluster have the same color. In Fig. 7.5**b**, the data is overlapping, i.e., there is no gap in between and therefore the solution obtained by the algorithm is not necessarily meaningful. Looking at the data without any coloring we would expect there to be only one cluster, since structure is largely absent from this data set.

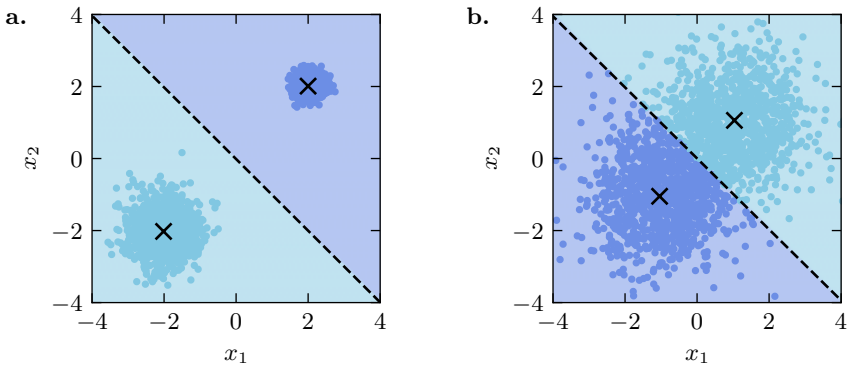

Figure 7.5: Result of a k-means clustering calculation. **a.** Two clusters are well-defined and a sensible solution is returned. **b.** The solution returned by the algorithm is rather artificial, since the data does not impose a distinction between two clusters. Equal colors mean equal clusters and the cluster centers are marked by "×".

From these considerations we have seen that all clustering methods, independently of the distance function used, must in some sense rely on the existence of a gap between data points, which could be the absence of data points in some region, but can also be expressed more generally as the presence of low density regions. Low density here means low compared to the density within the clusters or to the average density. Based on the composition of the data, the optimal method varies and there is no single general purpose algorithm.

We now apply clustering to our topological model. Assuming that our data set consists of parameters characterizing particular models or materials that we want to classify in terms of their topology, usually the entire feature space is populated with data points as data would be generated using some Monte Carlo or grid based scheme. For topological insulators and superconductors, each data point that has a gap in the spectrum of the Hamiltonian (gap function) can be assigned a topological index. This means that the existence of metals, which we assume irrelevant for this classification task, is an important ingredient that enables the successful solution of the problem.

Thankfully, topological invariants can only change under the condition of a gap closing, i.e., the existence of a metallic phase at phase boundaries is guaranteed. However, there is no guarantee for this boundary to generate a low-density region that is required for clustering algorithms to work. We illustrate how the presence of metallic phases can create gaps between topological phases in Fig. 7.6. Apparently, we are presented with only two cases: a metallic separation line that does not allow for clustering, as shown in Fig. 7.6**a**, or an extended metallic region that clearly separates the two topological phases and allows many clustering algorithms to find a satisfactory solution, see Fig. 7.6**b**. The general case of low density regions does not exist (with densities > 0), since there is no mixing of metallic and non-metallic points.

For realistic higher-dimensional data, the presence of a gap also provides a means to validate the result without the need for graphical inspection, which would be virtually impossible for dimensions larger than 3, by simply comparing distances to the different representatives. Assuming that the data is shaped as in Fig. 7.6**b**, it is clear, that mean-based or centroid algorithms like k-means that we described earlier will not work, since not all clusters have convex shape. This means that we can find points in cluster A that are closer to the center of cluster B than

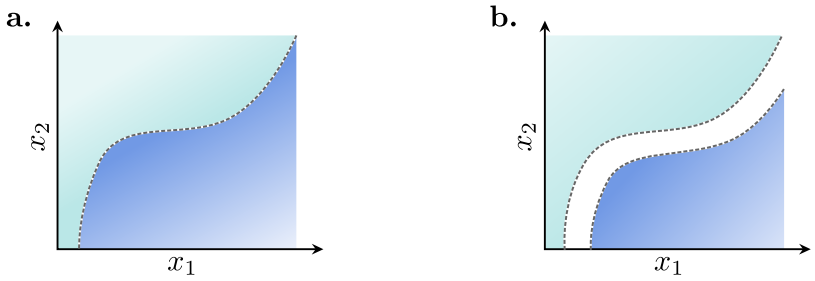

Figure 7.6: Two types of phase diagrams in terms of two arbitrary parameters x_1, x_2. Each diagram contains two topological phases shaded in different colors. **a.** The two phases are separated by a metallic line. **b.** The metallic region is extended in two dimensions. The two cases differ generally in that clustering algorithms can work for **b.**, but not **a.**, since the latter lacks a clear separation of data points.

that of cluster A, despite the existence of a separation between the two. Hence, we cannot use the distance to the cluster center as a distance measure. The solution is rather straight-forward, since we see immediately that what distinguishes points belonging to cluster A from points belonging to cluster B is that their neighbors also belong to cluster A. Here, connectivity-based algorithms should work much better, where we define the distance of a point x to a cluster C with representative $c \in C$ as

$$d : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}, (x, c) \mapsto \min\{\| x - x_i \| \, | x_i \in C\}, \tag{7.7}$$

which guarantees the existence of a discontinuity in the distances when moving from one cluster to the other. With this type of density metric any shape of clusters can be realized, however, the definition of a threshold is necessary, since the data is discrete and therefore all points naturally have a finite distance to their neighbors.

We now investigate as an example the Haldane model

$$H = t_1 \sum_{\langle ij \rangle} c_i^\dagger c_j + t_2 \sum_{\langle\langle i,j \rangle\rangle} e^{i\phi_{ij}} c_i^\dagger c_j + m \sum_i \mathrm{sign}(i) c_i^\dagger c_i, \tag{7.8}$$

cf. also Eq. 6.35. Here, we have a feature space of size $m = 4$, where $x_i = (t_1, t_2, \phi, m)$. We can straight-forwardly compute a data set on this rather small feature space using a grid based method, where we use an equidistant grid. At fixed t_1 and ϕ we obtain a two-dimensional diagram shown in Fig. 7.7**a**, where we color each grid point depending on whether it is in the data set (insulator) or excluded (metallic). It turns out that all data points are insulating, i.e., the data set does not separate into clusters at all. We had seen this already in our discussion of the Haldane phase diagram in Sec. 6.2, where we found that the topologically distinct phases are separated only by lower-dimensional submanifolds, i.e., lines for a two-dimensional diagram or surfaces for a three-dimensional diagram. As a consequence, any line drawn from a point in one phase region to a point corresponding to another phase encounters metallic solutions only at isolated points. The distance between any two phases is therefore 0. As a result, neither clustering algorithms discussed so far are applicable.

The solution lies in providing additional information. In Fig. 7.7**b**, we plot the band gap for the same range of parameters and find that it is always finite except for points lying on two distinct lines through the feature space, cf. Eq. 6.51. While this means that we cannot apply any clustering algorithm based on our criterion Eq. 7.6, the problem lies even deeper. Since our regularly distributed data contains no underlying structure whatsoever, we cannot apply any unsupervised learning algorithm. The band gap, however, can be used to artificially produce structure in the data by excluding points that have a very small gap. In Fig. 7.7**b**, we observe

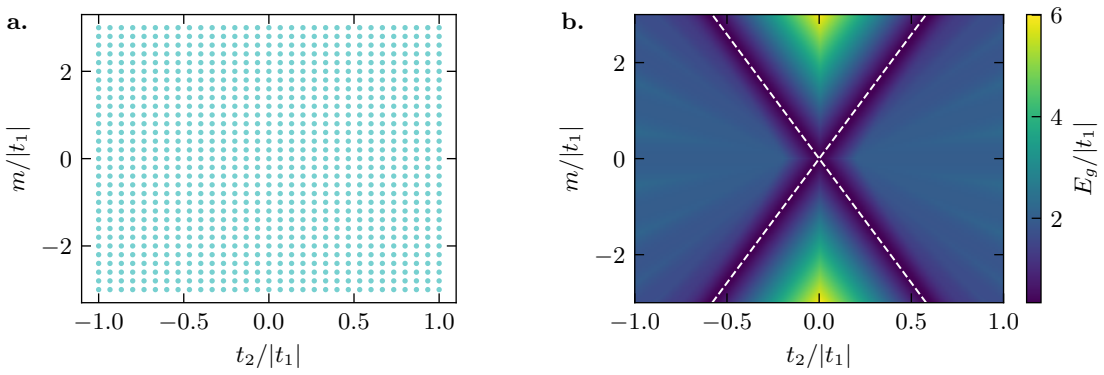

Figure 7.7: We plot the data for the Haldane model. **a.** Data grid, where all insulating data points are shown, where insulators are defined through $E_g > 10^{-2}$. Apparently, all data points are insulating and we did not find a single one with a sufficiently small gap. **b.** Band gap E_g as a function of the free parameters t_2 and m. The band gap vanishes only on points lying on two lines that cross in the origin. However, we find that around those gap closing points the gap goes to zero rather smoothly. $\phi = \pi/2$ was fixed during these calculations.

that the gap E_g is continuous in the features and slowly approaches 0 in the vicinity of the two lines. We can therefore use the band gap as a substitute for the density of points.

The band gap is difficult to compute numerically, since we are using a finite resolution in momentum-space. Hence, we are not guaranteed to resolve the points where the band gap closes. For the Haldane model, in particular, we have shown in Sec. 6.2 that the band gap can only close at the K and K' points, which are located at $(1/3, 2/3)$ and $(2/3, 1/3)$ in terms of the reciprocal lattice vectors. Fractions like this are difficult to treat numerically, since they are not exactly representable as floating point numbers. Furthermore, a general algorithm should be able to treat other models, too, where this information might not be known. Therefore, we choose here to not make use of our knowledge of the location of the Dirac points and instead approach the problem from a more generic perspective. We show in Fig. 7.8 two ways of introducing structure into our data by taking into account the metallic solution. In Fig. 7.8**b**, we simply take the previous regularly spaced grid of points and introduce the additional constraint that the band gap must always be larger than a threshold value E_{min} for all data points. We can extract a reasonable value by looking at an arbitrary gap closing and estimate roughly the error due to the finite grid. We show slices along t_2 for fixed m in Fig. 7.8**a**, and based on this define metals through $E_g/t < 0.4$, which gives the metallic region a small finite width. From Fig. 7.8**b** we can confirm that this produces data that can clearly be separated into different regions by eye.

We now use the band gap data as a density of points. Obviously we can no longer use a regularly spaced grid. Instead, we choose to sample from a distribution that will produce the same density profile as the gap data. In order to achieve this we apply the following simple algorithm:

1. shift the minimum of the gap to compensate for numerical inaccuracy,
2. define a local probability as
$$p(t_2, m) = E_g(t_2, m)/\max\{E_g\}, \tag{7.9}$$
3. draw random coordinates (t_2, m) from a uniform distribution on some interval and numbers $p_a \in [0, 1]$,
4. accept points if $p(t_2, m) > p_a$.

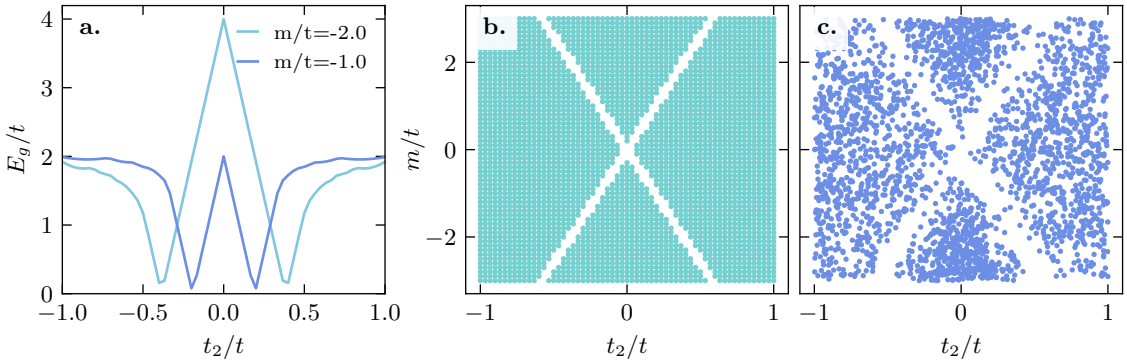

Figure 7.8: Modified data obtained by taking into account the band gap. **a.** Values of the band gap for several values of m/t. **b.** We use the original regularly spaced grid but place an additional constraint on the data points: the band gap must be above a specific threshold value that depends on the numerical accuracy we expect. Here, we chose $E_{\min}/t = 0.4$ for a relatively coarse k-grid of 41×41. **c.** Data obtained stochastically, by sampling from a distribution given by the gap data.

Step 1. is performed by setting $E_g \to \max\{0, E_g - E_{\min}\}$. However, this time we do not need to artificially increase the value of $E_{\min}$, since the density will naturally be low around the minimum. Therefore, we choose here $E_{\min}/t = 0.2$. Instead of a probability distribution function that is normalized over the entire patch in parameter space we choose to normalize probabilities locally according to Eq. 7.9. If the values of E_g in regions with finite gap differ greatly we could also choose to normalize according to

$$p(t_2, m) = \min\{1, E_g(t_2, m)/\tilde{E}_g\}, \tag{7.10}$$

with $0 < \tilde{E}_g \leq \max\{E_g\}$. In this case the density of the data will be more homogeneous than the gap values. Steps 3. and 4. correspond to a Metropolis-type algorithm [237, 238], where we first propose a random step and accept it only with a probability given by $p(t_2, m)$. We show now that this produces a density corresponding to Eq. 7.9.

Proof. We define $x_1 = t_2, x_2 = m$. By defining the local probabilities as in Eq. 7.9 we made sure that p is basically proportional to the gap values, i.e., this corresponds simply to a normalized density. Choosing coordinates uniformly guarantees that every point is equally likely to be proposed. With p_a uniformly distributed in $[0, 1]$ the acceptance condition is met with probability $P(p_a < p(x_1, x_2)) = p(x_1, x_2)$. The probability density of accepted points is then given by

$$\rho(x_1, x_2) = \frac{p(x_1, x_2)}{\int p(x_1, x_2) \mathrm{d}^2 x}, \tag{7.11}$$

and therefore, the probability density is proportional to the value of the gap with an average value of $\frac{p(x_1,x_2) N \delta_1 \delta_2}{\int p(x_1,x_2) \mathrm{d}^2 x}$ for N proposed data points per field of size $\delta_1 \delta_2$. □

We now evaluate how unsupervised learning can be applied to this data (an example is shown in Fig. 7.8**c**). Connectivity-based algorithms are very susceptible to noise, since the presence of a single point within the gap can act as a bridge to connect two otherwise separate clusters. For the regularly spaced grid this is not a problem, however, in order to decide if a pair of points is connected the algorithm needs an input for the maximal distance $d_{\max}$ between connected points. If we have a grid of points $v_i = \sum_j (a_j + i_j \delta_j) \mathbf{e}_j$ with grid spacings δ_j for components j the distance between any two points (assuming the Euclidean distance) is

$$d(v_i, v_j) = \sqrt{\sum_k (i_k - j_k)^2 \delta_j^2}. \tag{7.12}$$

Hence, points that are directly adjacent to one another along any axis have distances δ_j. If δ_j are very different from one another for different j this can lead to problems where no proper choice for $d_{\max}$ that is less than the gap width but larger than all δ_j exists. This is encountered, e.g., for the original data grid shown in Fig. 7.8**b**, where we used an equal number of points in both dimensions despite the different scale. This problem can be remedied by either defining a modified distance function or choosing a uniform grid with $\delta_j = \delta \, \forall j$.

We show in Fig. 7.9 three results obtained by unsupervised learning algorithms. In Fig. 7.9**a**, we use the uniformly spaced grid and apply the DBSCAN algorithm [239,240]. In contrast to the connectivity-based algorithm, DBSCAN is density-based and therefore puts faith only in those data points that are surrounded by a particular minimum density. Outliers that don't satisfy this constraint are classified as noise and do not belong to any cluster. Apparently, DBSCAN succeeds in classifying four different regions separated by metallic solutions. Regarding the

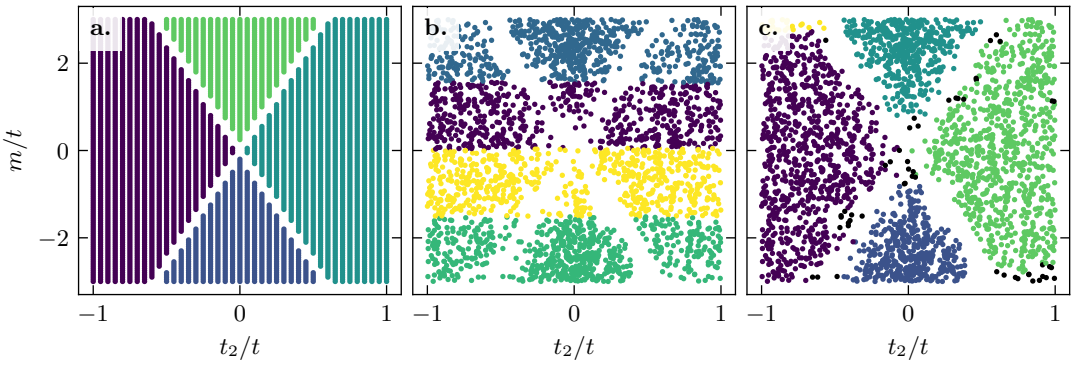

Figure 7.9: Clustering applied to the data that is aware of the gap values. Points belonging to a single cluster have the same colors. **a.** Uniform regularly spaced data points are distinguished correctly by DBSCAN, which finds exactly four different classes of points. **b.**,**c.** Noisy data obtained by sampling from the probability density proportional to the gap function. In **b.**, we apply k-means, which fails to distinguish phases properly. **c.** Result of the DBSCAN algorithm for the same data. The solution is rather good, although an additional fifth cluster appears (yellow). Black points are labeled as noise.

noisy data sampled via the local probabilities of Eq. 7.9, we investigate the performance of both DBSCAN and k-means. In Fig. 7.9**b**, we show the k-means result. Apparently, the centroid-based algorithm fails to capture the structure of the data. We can explain this by considering the expected optimal solution. Due to the symmetry of the data, the two centroids corresponding to the triangular shaped regions will have coordinates $t_2 = 0$ and $|m|/t > 2$. The remaining two centroids must be located at $m/t = 0$ and $|t_2|/t \approx 0.7$. As a consequence, the lines marking the boundaries of the clusters cannot lie within the gap, which makes this solution as plausible as any other. We note that this algorithm could succeed if the separation between the clusters were increased.

In contrast to k-means, DBSCAN succeeds in obtaining a reasonable clustering solution, as shown in Fig. 7.9**c**. While there are a number of points that reduce the separation between clusters, due to the finite probability within the "gapped" region, careful tuning of the parameters leads to their classification as noise, so that all clusters are distinguished properly. A minor weakness in the result is the misclassification of a small number of points at small t_2 and large m as a fifth cluster. This could not be resolved by tuning the parameters.

We have now seen that, even though there does not strictly exist a separation in the coordinates between different topological classes, the application of clustering algorithms is still possible by making use of the band gap as a point density. In the optimal case where the correct result is obtained, however, this approach cannot reveal the relationship between the different

phases. In fact, from Fig. 6.4 we know that the two clusters centered around $t_2 = 0$ correspond to the same topologically trivial phase, which is not revealed by the data shown here. In order to account for the topological classification one would have to compute the topological invariant for a representative of each cluster.

To summarize, clustering of the model parameters itself cannot be applied generally, due to the lack of a metallic region of finite width. Using the band gap as additional input data can resolve this problem, however, since the information about the topological phase is still not contained in this extended data set, a complete classification is not possible. The topological information is, however, contained in the eigenstates of the Bloch Hamiltonian. It is therefore possible to perform clustering on the eigenstates based on a distance measure that is able to detect the discontinuous behavior of the eigenstates at the topological transition. This has been shown to work, e.g., in Ref. [217].

We wrap up this discussion of unsupervised learning approaches by assessing the usefulness of such methods in the context of improving our understanding of topological phase diagrams. Apparently, what we are able to accomplish is the labeling of insulating phases that are separated by metallic regions without explicit knowledge of the topological invariant. This reduces the size of the data set significantly, and through the computation of a small number of topological indices one would arrive at a complete phase diagram. Methods like this could prove to be very helpful in discovering unknown phenomena by scanning large parameter spaces for hidden structure. However, we note that the unknown information must already be encoded in the distance measure, which makes such an approach rather tedious without any prior knowledge. Even if a possible candidate is found after testing many possible metrics and data compositions, one would then have to find a physical explanation for the type of structure that this metric probes.

We have here focused on the type of dimensionality reduction of Eq. 7.2. In Sec. 7.7 we will comment on the second type of dimensionality reduction (cf. Eq. 7.2) that instead of the number of data points reduces the dimension of each data points.

7.5 Data Generation

In the following, we will discuss supervised approaches with which we try to navigate around the difficulties of interpretability that we discussed in Sec. 7.3, while still providing valuable information regarding the original question of understanding the high-dimensional general phase diagram.

Since the data that we need is not abundantly available from, say, DFT calculations [241,242], we need to generate our own data[1]. Here, care has to be taken that we do not already imprint a bias on the data set and thereby influence the outcome. It turns out that this is not as simple as it sounds, since a maximally unbiased approach conflicts with other constraints that we impose to obtain a clear signal.

The data set of interest here is composed of a set of n_{features}-dimensional feature vectors $\mathbf{x}_i$, where n_{features} is the number of features and is controlled by the number of variables we allow in our system. This, of course, depends also on the type of parameterization used. In accordance with the usual notation we define the data matrix X of the data set through

$$X = (\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_{n_{\text{samples}}-1})^T, \tag{7.13}$$

[1]While materials databases do exist, they typically contain $\mathcal{O}(10^3)$ samples, and therefore, the corresponding feature space is rather sparsely populated. In addition, using such a database we would merely investigate known results and not follow the spirit of our statistical approach, wherein we aim to investigate *all* possibilities.

which is an $n_{\text{samples}} \times n_{\text{features}}$-dimensional matrix with n_{samples} being the number of samples, i.e., the number of independent features vectors. Specific individual features are denoted by $x_j := X_{ij} = [\mathbf{x}_i]_j$ in reference to the feature in general with no regard to the individual sample. Since our algorithm is supervised in nature, we rely on additional data in the form of an n_{samples}-dimensional vector Y containing the labels corresponding to each of the data points. The data set is then defined as the tuple (X, Y), which generally define a relationship through the map

$$f : \mathbb{R}^{n_{\text{features}}} \to \mathbb{Z}, f(X) = Y. \tag{7.14}$$

Here, we assume real-valued features and that f is performing a classification in the sense that labels are integer scalars.

With the task at hand being the understanding of topological phases and having seen in our discussion of unsupervised approaches that such information is very difficult to come by without knowledge of the topological invariant, we choose here to label the data with the corresponding topological invariant. We denote the classifier that computes the topological invariant by C. The model is defined through a Bloch Hamiltonian $H_{\mathbf{x}_i}(k)$ that itself depends on the features $\mathbf{x}_i$. We can make Eq. 7.14 explicit as

$$y_i = C[H_{\mathbf{x}_i}(k)] =: C(\mathbf{x}_i), \tag{7.15}$$

where $f = C \circ H$ is a function of the features. Since we generate the data ourselves, we have to assume that the classifying function C is known. Hence, it would make little sense to use a conventional supervised learning approach that is aimed at learning C. Instead, we use the data solely to sort the data into separate data sets depending on their class label, which circumvents the difficulty with non-finite gaps between clusters that we encountered in Sec. 7.4. With this difficulty resolved, we can then analyze the composition of the different data sets and describe the structure and relationships of features within each one of them. A further usage of the label is therefore not required.[2]

For the generation of data we define the data points as

$$\mathbf{x}_i = \mathbf{x}_{\text{ref}} + \delta_i, \tag{7.16}$$

where $\mathbf{x}_{\text{ref}}$ is an arbitrary reference point that is the same for all samples in the data set and δ_i are perturbations to the coordinates of the reference point. The δ_i are sampled symmetrically around the reference point from a suitable random distribution, chosen such that we obtain both decent coverage of the parameter space around x_{ref} and interpretable statistics.

Naturally, the question arises why we use random samples instead of a regular grid. Assuming we did use a grid, that would require a fixed amount of points (fixed resolution) m_i for each dimension i. The total number of points is then given by

$$n_s = \prod_{i=1}^{n_f} m_i = \mathcal{O}(m^{n_f}), \tag{7.17}$$

where $m = m_i/a_i$ and a_i are proportionality constants defined for each feature such that $m_i = a_i m$. Hence, the total number of points required scales exponentially with the number of features. Since our algorithm is supposed to work also for large dimensions, where it unfolds its true potential, since those are the problems that are the most difficult to understand, an approach

[2]The label does indeed only serve to reveal the underlying structure of the data set, that is a consequence of the topology of the system. Thus, we will usually refer to the Chern number simply as "the label" or y, since the definition of C is irrelevant and should, for the purpose of our analysis, be considered unknown.

with exponential scaling is not acceptable. We will explain later on how the sample size is chosen to avoid encountering the same problem and what this means for the quality of the statistics.

The description of quantum materials in terms of models is usually based on the tight-binding representation, i.e., a selection of amplitudes in a Wannier representation is given names $t_1, t_2, t_3, \ldots \in \mathbb{C}$ and these are used to refer to an entire class of matrix elements of the real-space Hamiltonian, e.g., for a single site unit cell

$$H = \begin{pmatrix} 0 & t_1 & t_2 & t_3 & \ldots \\ t_1^* & 0 & t_1 & t_2 & \ldots \\ t_2^* & t_1^* & \ddots & \ddots & \\ t_3^* & \ddots & \ddots & & \\ \vdots & \ddots & & & \end{pmatrix}. \tag{7.18}$$

where we set the onsite energy ε to zero. The matrix representing the Hamiltonian in the Wannier basis is a hermitian $L \times L$ matrix, where L is the size of the system in sites. Translational symmetry requires that all matrix elements along diagonals are the same, which for a hermitian matrix leaves exactly L parameters. It makes sense to order them according to the distance between the sites they represent in increasing order, which means that t_1 is the matrix element between nearest neighbors, etc. (cf. Sec. 3.1.4). For a single site unit cell, the diagonal matrix element is proportional to the identity and can therefore be set to 0 without loss of generality. The remaining $L - 1$ parameters are limited to only the first few representing the shortest distances, since these are expected to be the strongest contributors to the physical properties—an assumption that is motivated from the definition of the hopping matrix elements, cf. Eq. 3.58. For this rather basic example, only a handful of parameters exist anyway, which allows for a simple description of the phase diagram. If we increase the size of the unit cell, we effectively replace the entries of the matrix Eq. 7.18 with submatrices of the same size as that of the unit cell. Translational symmetry is no longer required for the matrix elements contained therein, which means that the number of free parameters in increased.

In general, we want to look at cases that are represented more conveniently through overlap integrals of orbital wavefunctions, that compose the matrix elements of the Hamiltonian. This is typically used for multiorbital materials, where the number of significant hopping matrix elements is much larger and models are therefore much more difficult to formulate. We denote the corresponding matrix elements as $t_{ij}(\mathbf{R})$, where $\mathbf{R}$ is the displacement vector that connects different unit cells, and i, j are the indices of the respective sites/orbitals. In this representation, the number of parameters is limited by the lattice graph itself, which—through the associated connectivity—determines how many possible values of $\mathbf{R}$ with the same lengths exist and how these lengths $|\mathbf{R}|$ increase as a function of the degree of neighbors. Secondly, $i, j < n_{\text{unit}}$ are bounded by the unit cell size. The increased number of parameters w.r.t. the single site formulation stems from the possibility to break translational and point group symmetries, which allows for the discovery of new models that are intrinsically unsymmetric.

The feature vector is now defined as

$$\mathbf{x}_i = (t_{ij}(\mathbf{R}) \mid \forall \mathbf{R}, i, j), \tag{7.19}$$

where allowed values of $|\mathbf{R}|$ are bounded to a small sensible value. The data thus represents a view onto a generalized phase diagram as described in Sec. 7.2. The number of features is typically much larger than four, and therefore, a graphical representation is out of the question. What remains are traces of the relationships between individual parameters that arise out of the mapping of Eq. 7.15 and that we want to extract systematically.

We note that an understanding of the phase diagram in terms of these relationships is necessary to be able to make any kind of predictions. A prediction in this case corresponds, in principle, to an inversion, i.e.,

$$f^{-1}(y_i) = \mathbf{x}_i. \tag{7.20}$$

Given f^{-1} it would then be trivial to obtain a feature vector for any desired class label y_i. However, since $y_i \in \mathbb{Z}$, i.e., countable, and $\mathbf{x}_i \in \mathbb{C}^{n_f}$ (uncountable), f is not invertible. Any kind of prediction can therefore not be unique and must be based on regions in feature space rather than particular points.

The choice of features that we made in Eq. 7.19 is motivated by our aim to analyze the relationships between parameters. For an approach that attempts to train a model, this choice of the data set is rather inconvenient, since the model would have to encompass the creation of the Hamiltonian matrix, diagonalization thereof and computation of the topological invariant in terms of the eigenvectors. The complexity of this task is sufficiently high that only very complex neural network type approaches are likely candidates, which makes an interpretation of the resulting model very difficult. If this were the aim we would recommend an approach like in Refs. [198,217,234], where the components of eigenvectors of the Hamiltonian are used as the features.

Interestingly, training an arbitrarily complicated model such that it reproduces the data typically achieves nothing regarding the desired information, since the complexity of the data set is simply shifted into the parameters of the model. Having completed the computationally expensive task of training the model one then faces a similar problem as before in analyzing the data set that is now composed of model parameters instead of the original features. We demonstrate one such approach in Sec. 7.6.

So far, we defined the features to be simply the hopping parameters that are generally complex numbers. However, it is often more convenient to work with real features. We therefore transform the feature vector via a mapping $g : \mathbb{C} \to \mathbb{R}^2$. This is, of course, not unique, however, a certain natural choice is given by

$$g(x_j) = (\mathrm{Re}\,[x_j]\,, \mathrm{Im}\,[x_j]) \qquad \text{or} \qquad g(x_j) = \left(|x_j|, -i \log\left(\frac{x_j}{|x_j|}\right)\right), \tag{7.21}$$

where the logarithm in the second option refers simply to the phase ϕ of the complex number $x_j = |x_j| e^{i\varphi}$, which we will denote as $\varphi = \varphi(x_j) \in [-\pi, \pi]$. The relationships between the two different representations of complex numbers are rather complicated

$$\begin{aligned} \mathrm{Re}\,[x]\,(|x|, \varphi) &= |x| \cos(\varphi), \\ \mathrm{Im}\,[x]\,(|x|, \varphi) &= |x| \sin(\varphi), \\ |x|(\mathrm{Re}\,[x]\,, \mathrm{Im}\,[x]) &= \sqrt{\mathrm{Re}\,[x]^2 + \mathrm{Im}\,[x]^2}, \\ \varphi(\mathrm{Re}\,[x]\,, \mathrm{Im}\,[x]) &= \arctan\left(\frac{\mathrm{Im}\,[x]}{\mathrm{Re}\,[x]}\right), \end{aligned} \tag{7.22}$$

since with one exception they involve trigonometric functions. Therefore, we will be using both options, thereby doubling the number of features and introducing a certain redundancy. We then select the best representation, where two out of the four are enough to uniquely define the original complex number. Since the Hamiltonian is hermitian by definition, certain features that appear only in the diagonal matrix elements must be real. We treat these features separately and capture only the real part.

Sampling of Features

The requirements for the sampling procedure are twofold. First, we should canvass a large portion of the configuration space in order to increase the potential for predictions. Second, the configurations that we use as samples should form a rather homogeneous cloud in order to avoid the introduction of too much of a bias. By choosing a symmetric distribution, which means symmetric w.r.t. arbitrary reflections through the reference point or rotations around an axis piercing the complex parameter plane at the reference point, we remove the bias from the distribution. The bias is not completely gone, but instead controlled by $\mathbf{x}_{\text{ref}}$, which serves as the center of the data cloud. A completely unbiased approach would be obtained by setting $\mathbf{x}_{\text{ref}} = 0$ and sampling over all possible feature vectors $\mathbf{x}_i \in \mathbb{C}^{n_{\text{features}}}$ uniformly, however, we will see later that with this choice the desired information is rather inaccessible. Hence, a certain bias is necessary.

We will use two different probability distributions to sample the features from:

1. scaled normal distribution

$$\rho_1(\mathbf{x}) = \prod_{i=1}^{n_f} \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{1}{2}\frac{(x_i - x^i_{\text{ref}})^2}{\sigma_i^2}}, \tag{7.23}$$

2. uniform distribution

$$\rho_2(\mathbf{x}) = \prod_{i=1}^{n_f} U(\Omega_i), \tag{7.24}$$

   where Ω_i is the sample space for feature x_i.

Clearly, both choices sample features independently, such that we can define probability distributions per feature as

$$\rho_{1,i}(\mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{1}{2}\frac{(x_i - x^i_{\text{ref}})^2}{\sigma_i^2}}, \tag{7.25}$$

and

$$\rho_{2,i}(\mathbf{x}) = U(\Omega_i). \tag{7.26}$$

The sample spaces Ω_i and the standard deviations σ_i are chosen similarly. Here, we define $\Omega_i = B_{\alpha|x^i_{\text{ref}}|}(x^i_{\text{ref}}) \subset \mathbb{C}$ for complex features. The notation $B_r(x)$ refers to the solid sphere or ball of radius r, centered around x, where we generally use the Euclidean norm to describe distances. The parameter $\alpha := r_i/x^i_{\text{ref}}$ controls the spread of the distribution in units of the initial value. With α fixed across all features, this implies that features with a large component in x_{ref} are allowed to vary more than those with smaller components. The intuition behind this procedure is to try to limit both the number of unphysical features and one particular source of redundancy. Assuming that x_k corresponds to $t_{ij}(\mathbf{R})$ with $|\mathbf{R}| \gg 1$ we would expect the value of $|x_k|$ to be rather small compared to other features corresponding to hoppings across smaller distances. In order to satisfy this criterion for "physicality" on average we therefore correlate the largest allowed values for features with the initial values. This does not enforce a hard constraint on the individual features, but makes these configurations less likely. Since topological invariants depend, as we know, only on the eigenvectors of the Hamiltonian, we are free to rescale the energy axes as we see fit, i.e., $H \to cH$ with $c \in \mathbb{R}^{>0}$. This transformation changes neither the eigenvectors nor their order, and therefore, leaves topological invariants unchanged. Given that we allow all values within a certain radius around $\mathbf{x}_{\text{ref}}$, this applies, in particular, to the features with large initial values. A sample of the form (features arranged by distance $|\mathbf{R}|$ in increasing order)

$$\mathbf{x} = (\epsilon_1 x_1, \epsilon_2 x_2, \dots x_k), \tag{7.27}$$

where $0 < \epsilon_i \ll 1$, would be equivalent to

$$\mathbf{x}' = (x_1, \frac{\epsilon_2}{\epsilon_1} x_2, \dots \frac{1}{\epsilon_1} x_k), \tag{7.28}$$

where $|x'_k| \gg |x_1|$. According to our earlier definition this corresponds to an unphysical configuration, however, we allow it in order to reduce the bias, albeit with a reduced probability. In addition, restricting the maximal values of features differently, reduces the amount of samples that are related by scale transformations, and therefore, reduces the redundancy in our data set.

We illustrate the choice of the sampling space in Fig. 7.10. Obviously, we can only draw this for up to three dimensions, for which we obtain the cylindrical structure shown in the image. We opt for a shape of the form

$$\Omega = \left\{ \mathbf{x} \in \mathbb{R}^{n_{\text{features}}} \,\middle|\, \frac{|x_i|}{|x^i_{\text{ref}}|} \leq \alpha \; \forall i \in \{1, \dots, n_{\text{features}}\} \right\}. \tag{7.29}$$

For complex hopping parameters this means that sampled values all lie within a circular region around the reference point, while the radius r_i of this circle is controlled by the modulus of the corresponding coordinate of the reference point. For real parameters this corresponds simply to an interval of width r_i around x^i_{ref}. For the Gaussian distribution we set $\sigma_i = r_i$, and therefore, the volume shown in Fig. 7.10 corresponds to the 1σ region.

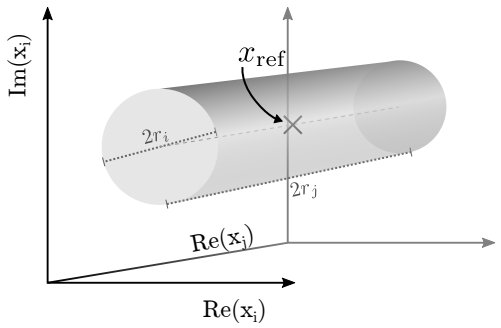

Figure 7.10: We illustrate the coverage of the configuration space Ω produced by the uniform distribution. Here, two features x_i and x_j are shown. Projected onto x_i, all samples lie within a distance r_i from the reference point x_{ref}. For the Gaussian distribution, the volume depicted here marks the 1σ region. [Figure reproduced from Ref. [173]]

Feature engineering

It is common in machine learning to engineer additional features from the bare input data in terms of arbitrary functions of features, as a way to improve the performance of a simple model. We applied this by transforming the originally complex hopping parameters to different sets of real features in Eq. 7.22, which are related to each other via nonlinear functions. It would be possible to go further than that and we expect that this would be a key ingredient in order to learn more, however, this goes beyond the current scope and we will comment on possible ways to achieve this later.

7.6 Supervised Learning (Tree-based Classifiers)

We now discuss the application of a traditional machine learning algorithm, namely the random forest, to the data. The random forest is one of the most common methods used in machine

learning [243] and can be thought of as an ensemble of decision trees that are a common model for classification. In fact, decision trees seem like the perfect model for our purpose as we will explain in the following.

The understanding of phase diagrams in terms of the relations between parameters that we want to achieve can often be realized by means of ranges of parameters for which certain phases can be observed. Writing down a particular model for this purpose is rather difficult, since non-linear models are usually required to capture anything but the simplest relationships. Usual classification models in the form of logistic regression work on the basis of a regression model whose roots determine the threshold for distinguishing between different classes. Assuming a linear model

$$h_w(x) = \sum_{i=1}^{n_{\text{features}}} w_i x_i + w_0, \tag{7.30}$$

the roots are given by

$$\mathbf{w} \cdot \mathbf{x} = 0, \tag{7.31}$$

where $x_0 = 1$, which corresponds to a generalized plane in n_{features}-dimensional space with normal $\mathbf{w}$. Projection on any two dimensions will therefore yield straight lines as a separation between phases. Clearly, this will not be accurate in a general setting, which means that the training phase leads to a significant error that we then have to evaluate in order to draw conclusions. The obvious route to improve the model's performance is to add additional non-linear features, e.g., quadratic terms $\propto x_i x_j$, and retrain the model to evaluate the resulting error. The training phase quickly becomes very costly and with an increasing number of features we are increasingly lost trying to capture the underlying relationships learned, since the simple interpretation as a plane, cf. Eq. 7.31, is no longer valid.

The decision tree, on the other hand, is intrinsically non-linear, since the classification is based on a series of boolean decisions, i.e., *yes/no* questions, that are readily comprehensible and can be combined to form a broader picture. Each question is of the form $x_i > v$, i.e., it separates the data into two subsets based on a threshold value. With a single question this is similar to the linear classifier, however, since this can be done an arbitrary number of times, the categorization of the data performed through the tree is, in general, highly non-linear. Through the tree-shape and its inherent boolean decision-making process, the result retains a certain interpretability, though. Information about the algorithmic learning process can be found in Refs. [244, 245]. One weakness of this type of model is that decision trees tend to overfit the result [246] by trying to account for every single data point, which typically requires a very deep tree structure that loses a lot of its simplicity due to its sheer size. For conventional applications, where the data naturally contains noise, this means that the model incorporates faulty information. In our case, however, we compute the topological invariant for every point and can therefore be certain that the data is noise-free. Nevertheless, a simpler description of the data that gives a rough understanding would be preferred over a deep tree structure that is similarly difficult to read as the data itself.

7.6.1 Decision Trees

We now evaluate the use of a decision tree for the description of a phase diagram[3]. To this end we choose a particularly simple use case that does not yet include the full complexity of

[3]Since decision trees are the go-to models in machine learning, they have, of course, already been applied in a topology context, see Ref. [234]. This approach follows a rather different motivation, though, and we note that we were not aware of earlier work until after finishing our own. Full credit for the idea of using random forests goes to Daniel Guterding.

high-dimensional systems with many features. Here, we simply use the Haldane model in the form of Eq. 6.35, where we define the features as

$$\mathbf{x} = (t_1, t_2, \phi, m). \tag{7.32}$$

Apparently, all x_i are real so that there is no need to perform additional feature engineering. We keep in mind, though, that there is an inherent redundancy in this representation, since we neglect the freedom of a global scale parameter. Therefore, many seemingly different data points, in fact, correspond to the same data point in terms of the eigenvectors that are created. Physically, it is debatable whether we call the configuration with $t_1 = 1, t_2 = 0.1, m = 0.5$ distinct from $t_1 = 2, t_2 = 0.2, m = 1$ (in arbitrary units). Since the Hamiltonian is the same up to a constant factor, though, it makes sense to merge points that are related by a scale transformation into a single data point in order to avoid an overly complicated model. On the other hand, such exact correspondences are rather unlikely, given the stochastic process with which we generate the data. One way around this would be to fix the scale explicitly by setting, e.g., $t_1 = 1$ and thereby reducing the number of features by one.

We first investigate the performance of the decision tree classifier on our data set. The creation of the data set is performed using the uniform distribution of Eq. 7.24 with a spread parameter of $\alpha = 1.5$. The reference point is chosen as $\mathbf{x}_{\text{ref}} = (1., 0.2, \pi/2, 1.05)$ in arbitrary units, where the Chern number vanishes. However, $\mathbf{x}_{\text{ref}}$ is very close to the phase transition at $m/t = x^4_{\text{ref}} \approx 1.04$, which means that we expect to find several phases in the vicinity. This is a basic requirement that guarantees the suitability of our data set for the study of the phase diagram. For the size of the data set we choose $n_{\text{samples}} = 100000$, which controls the number of random samples generated. Of course, the number of samples belonging to the different classes is not necessarily the same as this depends primarily on the reference point. In most cases it is reasonable to assume that the data set will be intrinsically biased in a way that not all labels appear the same number of times, i.e., there can be an overabundance of some labels, while others are scarce. The training algorithm will always try to maximize the performance, i.e., the rate of correct classifications overall, which means that an imbalance in the distribution of labels in the training data can lead to an imbalance in the performance between different class labels.

We investigate these effects of imbalance below as our data set contains the following number of samples:

$$n_{y=0} = 69852, \quad n_{y=1} = 21763, \quad n_{y=-1} = 8385. \tag{7.33}$$

Clearly, there is an overabundance of trivial samples with the topological samples holding only a fraction of ≈30%. In order to validate the result, we split the data set into two parts: the training and test sets. It is customary to use a third set, the so-called validation set, to optimize hyperparameters, i.e., the parameters that determine the composition of the model (e.g., depth of the tree) [131, 247, 248]. Since we are not optimizing the hyperparameters right now, we have no use for this additional set, and therefore, include more points in the test set. In Table 7.1 we show the composition of the training and test sets for six different choices, all generated from the initial data set.

We first look at an unbalanced data set, which we call set 0*, where we split the original data set arbitrarily into two, according to $n_{\text{test}}/n_{\text{train}} = 2/3$. Clearly, the ratios between the abundances of the individual labels are unaffected by this. In order to test the dependence of the performance on the balance of the data set we need to create balanced data sets with the same number of samples for each label. The smallest number $n_{y=-1}$ is the limiting factor here and with an 1:1 splitting between training and test size we obtain set 1 with a training set size of $n_{\text{train}} = n_{y=-1}/2$ and sets 2-4 with sizes $n_{\text{train}} = n_{y=-1}/3, n_{y=-1}/10, n_{y=-1}/30$, respectively. Since sets 0* and 1 differ greatly in the total training set size we decided to compare instead sets 0 and 1, where this additional source of error is not present. We proceed by training a decision tree

	total train	$y = 0$	$y = 1$	$y = -1$	total test	$y = 0$	$y = 1$	$y = -1$
set 0*	66666	46568	14508	5590	33334	23284	7255	2795
set 0	12574	8784	2736	1054	87426	61068	19027	7331
set 1	12576	4192	4192	4192	12576	4192	4192	4192
set 2	8385	2795	2795	2795	8385	2795	2795	2795
set 3	2514	838	838	838	2514	838	838	838
set 4	837	279	279	279	837	279	279	279

Table 7.1: Number of data points per label for the different data sets used. Sets 0 and 0* are unbiased sets that retain the ratios of points between the different labels. Sets 1 through 4 are created using symmetric numbers of samples for each of the labels, but with different total set sizes. We typically use the same size for the training and test sets, except for set 0*, where $n_{\text{train}} = 2n_{\text{test}}$ and set 0, where we chose the same training set size as for set 1 and assigned all other samples to the test set.

classifier on these different choices of training data with the SCIKIT-LEARN implementation [249]. In Fig. 7.11**a**, we show the resulting accuracy computed as

$$\text{accuracy} = \frac{\#\text{correct classifications}}{\#\text{test set}}, \tag{7.34}$$

which is evaluated only over the test set. As expected, the unbalanced training set yields a very good accuracy for the overly abundant class with label $y = 0$, which reaches values in the range $90 - 100\%$. The accuracy is shown here as a function of the maximal depth that we allow for the tree. Clearly, the larger the tree gets, the more accurate the prediction will become on the training set, since in the extreme case where the depth becomes so large that the number of leaf nodes equals the number of data points, there will be a separate rule for each sample. This is, of course, counterproductive, since we will have only transformed the data set into a tree structure, ideally containing the same information, however, the generalization of the model to new data is expected to be quite poor. Furthermore, the deeper the tree, the lower its interpretability as there are just too many rules to follow through to understand how decisions are made. Therefore, reliable performance at low depths is favored over good performance at large depths.

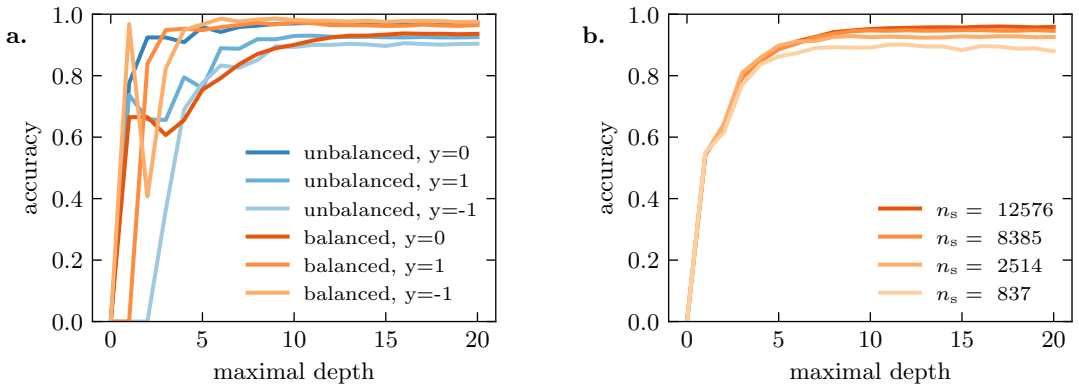

Figure 7.11: Accuracy of a decision tree model trained on our data sets from Table 7.1 as a function of the maximal depth of the tree. **a.** Two runs using set 0 (unbalanced) and set 1 (balanced), respectively. We discriminate the accuracy between the different topological classes. The performance of the model on topological vs. trivial samples crucially depends on the data set used. **b.** Several runs using different sizes n_s of the balanced data sets 1-4. The performance is rather robust w.r.t. the data set size.

The uptick trend of the performance as a function of the maximal depth of the tree is visible in the data, however, we also observe peaks in the accuracy for particular classes at low depths

that we want to investigate in more detail later. Coming back to the comparison between unbalanced and balanced data sets, we find that while for the unbalanced set the performance on the $y = 0$ class is especially good, the contrary is the case for the balanced set, where the accuracy on the topologically non-trivial classes outperforms the accuracy for the trivial class. This is certainly interesting and we can hope to find useful information in the rules learned by the training algorithm. Notably, none of the solutions found by the algorithm are optimal in the sense that maximal differentiation between classes is obtained alongside minimal complexity of the tree. This task corresponds to the problem of finding global extrema of neither convex nor concave non-linear functions, which is NP-hard [250–252].

Having established that the balanced data set yields a much better performance on the, for our purposes, more important non-trivial phases, we now proceed by using only balanced data sets. In Fig. 7.11**b**, we show the accuracy of the balanced data sets 1-4, i.e., as a function of the size of the training set. Since our model depends on only four features, the number of required samples is rather low, which is reflected in the very good performance across the board. Only at larger depths starting at around 5 we find a discrepancy between the different amounts of training data, where more data generally leads to a better performance. We note that the total accuracy levels out at a depth of around 10 with a rate of correct classification of $\approx 96\%$. Typically, we would expect that the accuracy decreases again for larger maximal depths as a consequence of overfitting the data [131, 253, 254], which the total accuracy does not reflect. However, the class-resolved accuracy shown in Fig. 7.11**a** does show a small reduction with increasing depth for the non-trivial phases, which is compensated for in the total accuracy by an appropriate increase of the performance on the trivial data.

We now want to investigate what information the algorithm has managed to learn, with a particular focus on what we can learn from the optimized representation of the model. Before we do this, however, we discuss the meaning of the hyperparameters, i.e., the maximal tree depth, and their effect on the complexity of the model. The decision tree is a binary tree, i.e., every node splits into at most two child nodes. We can distinguish two types of nodes: leaf nodes that do not have any children and internal nodes that have exactly two children. The former will carry a class label that corresponds to the majority of labels of those samples whose decision paths lead to that leaf node. Internal nodes on the other hand carry the binary conditions placed on particular features in order to split the data. This information, i.e., which sequence of features and which threshold values lead to a classification, is what we are interested to learn from the model. The complexity of the model is therefore inherently related to the number of internal nodes, which is controlled by the depth of the tree. Given a depth d, the maximal number of nodes is given by

$$n_{\text{nodes}} = \sum_{i=0}^{d} 2^i = 2^{d+1} - 1, \tag{7.35}$$

where $i = 0$ corresponds to a tree containing just the root node. The leaf nodes are those nodes that terminate a given path through the tree, of which there are at most 2^d. The number of internal nodes, and with it the complexity of the tree, is therefore bounded by

$$\text{complexity} \propto n_{\text{internal}} \leq 2^{d+1} - 1 - 2^d = 2^d - 1. \tag{7.36}$$

A lower bound is also readily constructed by considering the case of a single path of depth d, with all nodes not on this path being leaf nodes

$$\text{complexity} \propto n_{\text{internal}} \geq d. \tag{7.37}$$

Since these two bounds lie very far apart, we try to estimate the average complexity over all possible trees of depth d. To this end we start, as illustrated in Fig. 7.12, with the simplest

construction where $n_{\text{internal}} = d$, c.f. Fig. 7.12**a**. The tree has initially $d-1$ leaf nodes that we can turn into internal nodes, thereby growing the tree as shown in Fig. 7.12**b**. The problem now reduces to finding the number of trees with the same depth and a particular number of internal nodes, namely $d \ldots 2^d - 1$. This can be solved by counting the ways of turning leaf nodes above the final depth into internal nodes. While this is simple at first, it quickly becomes very difficult, since there are many conditions to check. A simple recursive relation does not exist for this reason. Therefore, we simply conjecture at this point that the distribution is symmetric in n_{internal} around the mean value $\bar{n}_{\text{internal}} = (2^d - 1 + d)/2$, and therefore, the average complexity can be approximated by

$$\langle \text{complexity} \rangle \approx \bar{n}_{\text{internal}} = \frac{1}{2}(2^d - 1 + d) \sim 2^{d-1}. \tag{7.38}$$

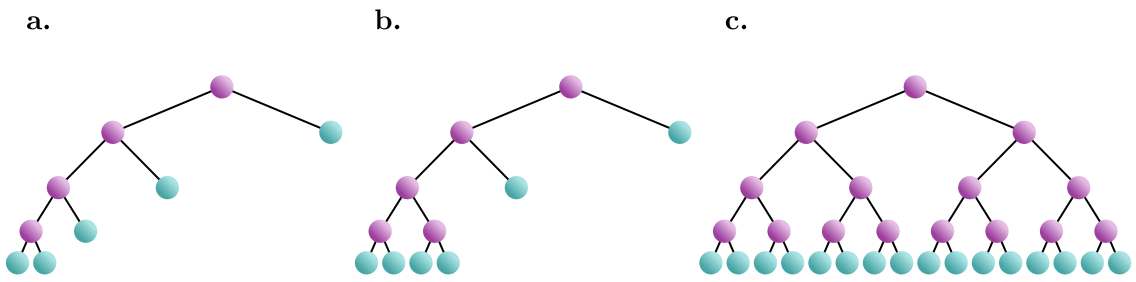

Figure 7.12: Construction of trees with fixed depth and fixed number of internal nodes. Starting from the minimal tree with d internal nodes in **a.**, one replaces leaf nodes at depth $< d$ successively through internal nodes as in **b.**, until one arrives at the complete tree shown in **c.** Internal nodes are shown in magenta, leaf nodes in teal.

The complexity of the model therefore scales exponentially with the depth of the tree. In Table 7.2 we give a few values of the average complexity from Eq. 7.38 to illustrate what value of d makes sense in order to be able to understand the decision-making process of the model. Apparently, values around $d = 10$ are not particularly useful for the purpose of interpreting the

d	0	1	2	3	4	5	...	10	...	20
$\langle \text{complexity} \rangle$	0	1	2.5	5	9.5	18	...	516.5	...	524297.5

Table 7.2: Average complexity of the decision tree model from Eq. 7.38 as a function of the maximal tree depth. We assume that the learning algorithm exhausts the maximal tree depth. It is to be expected that realistic models are closer to the complete tree, i.e., the complexity might be even larger in practice.

resulting tree structure. Instead, a complexity $\mathcal{O}(10)$ would be preferred, which can be achieved by setting $d \sim 5$. We note that even with this comparatively simple model the performance on the topological data set is already very good.

First we investigate the point $d = 1$, where the $y = -1$ phase shows an almost perfect accuracy. The resulting tree is particularly simple, since only a single internal node exists, i.e., only a single comparison rule is learned. The tree structure is shown in Fig. 7.13. Apparently, the two degrees of freedom in the model (choice of the feature and threshold value) were chosen such that $x_4 = m$ is compared against the threshold value 1.027 to split the data set into two parts. On the right, i.e., the condition is false and $m > 1.027$, we find 28.7% of the samples. However, strikingly, 77.8% of these belong to class $y = 0$ and only 3% to class $y = -1$, i.e.,

this is what the model believes to be class 0. On the other hand, for $m \leq 1.027$ (left) we find the majority of the samples and, in particular, roughly 96% of the $y = -1$ samples, which reflects the good accuracy on the test set. However, the considerable number of 15% of wrongly labeled $y = 0$ samples and the fact that $y = 1$ points are all wrongly labeled impact the overall performance negatively. In principle, our data sets are randomized, however, the accuracy on the training and test sets need not always be the same. A good way to achieve better estimates would be to create many data sets by randomly shuffling the points between the training and test sets and report the average accuracy. Here, we do not go into this detail, since we are only interested in how the model learns and the precise value of the reported accuracy is not important.

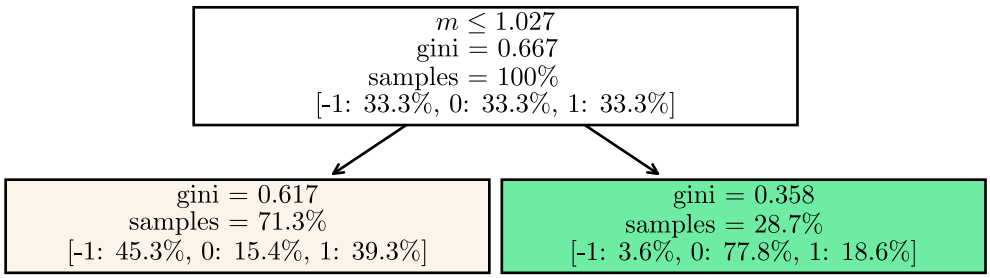

Figure 7.13: Tree structure of a pathological example with depth $d = 1$. There is only a single rule based on the feature $x_4 = m$. The two leaf nodes are not enough to represent three classes, but the model performs relatively well in the classification of the $y = -1$ samples in the data set. The impurity of the data at each node is given by the Gini index, cf. Eq. 7.39, "samples" is the fraction of samples reaching each node. We also specify the fractions of samples with labels $y = -1, 0, 1$, respectively.

If we take a closer look at the tree shown in Fig. 7.13, we notice the entry "gini". This measure of the impurity of a given node corresponds to the Gini coefficient [255–257], which is defined as

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \langle x \rangle}. \tag{7.39}$$

G is essentially a measure of the spread of the data, not around the mean like in case of the standard deviation, but between any pair of points. It is defined as the average difference between any two points divided by the mean.

Now, we increase the complexity of the tree by setting the maximal depth to 5, for which, according to Fig. 7.11, the model already provides consistently good performance. The resulting model has 63 nodes and of those 31 are internal nodes, i.e., the complexity corresponds to the upper bound of 31. The consequence for our discussion is that we cannot illustrate the model in the same way as Fig. 7.13, since the resulting graph would not fit onto a single page. If this type of complexity is already reached for such a simple problem with only four features, our expectation for the scaling behavior with increasing dimension of the feature space is beginning to look rather grim. It is therefore necessary to develop a different measure of the information learned by the model that is more indirect than the simple inspection of the model's decisions.

A promising approach is the use of the impurity information. In general, the impurity should decrease as a function of the depth of the node, since the algorithm's main goal is the reduction of impurity. A similar approach was used by Ref. [234], where instead of the Gini coefficient the entropy of the data computed for each node was used to measure the impurity. Although they

do not specify this to a greater extent, we believe that they compute

$$S = -\sum_y p_y \log(p_y) = -\sum_y \frac{n_y}{n} \log \frac{n_y}{n}, \tag{7.40}$$

where n_y is the number of samples reaching this node with label y. We define the impurity reduction for a feature x_i in terms of the Gini coefficient as

$$\delta G(x_i) = \sum_{j \in S_i} w_j G_j - (w_{j\to\text{left}} G_{j\to\text{left}} + w_{j\to\text{right}} G_{j\to\text{right}}). \tag{7.41}$$

Here, the sets S_i contain all nodes for which the feature x_i was chosen by the algorithm, respectively. G_j is then the Gini coefficient of the node where the corresponding rule involving x_i appears, and $G_{j\to\text{left/right}}$ are the Gini coefficients of the left and right child nodes, respectively. Since the impurity is in a way an intensive quantity, i.e., it is not proportional to the number of samples that actually reach a particular node, the impurity values are weighted by the fraction of samples at node j: $w_j = n_{\text{samples},j}/n_{\text{samples}}$. The value of $\delta G(x_i)$ therefore quantifies the total reduction of the data impurity through rules involving feature x_i, which to a certain degree describes the information content regarding the topological classification that is contained in a particular feature.

We obtain the data shown in Fig. 7.14 for multiple classes of trees, for which we allow different maximal depths. The Gini coefficient is presented in Fig. 7.14**a** as a function of the depth. Since there are a multitude of nodes at a given depth, we plot the average over all nodes at a particular depth

$$\langle G \rangle(d) = \sum_{\text{depth}(i)=d} G_i. \tag{7.42}$$

The standard deviation serves as a measure of the statistical error and the error bars shown correspond to the 1σ confidence level that contains most data points. We observe the expected trend of decreasing impurity as we progress deeper into the tree, i.e., the algorithm works correctly. The data shown here was extracted for a tree of $d_{\max} = 20$. In the plot we only show points up to $d = 18$. In fact, at $d = 18$ the Gini index reaches 0, i.e., the data has been purified and the algorithm terminates. Early termination is a sign that the learning procedure achieved 100% success rate in the training set, however, for new data points the accuracy is typically lower, albeit still very good in this case ($\approx 96\%$ total accuracy). Due to the difficulty of finding global minima, the algorithm employed is a greedy algorithm that selects the best choice locally. As a result, the corresponding graphs for smaller maximal depths are exactly the same up to an earlier termination at the respective $d_{\max}$. In Fig. 7.14**b**, we show the impurity reduction extracted from the optimal tree model for different values of $d_{\max}$ normalized to 1. From the minor variation in the data for different sizes of trees we see that the details are rather robust towards growing deeper trees. This is a consequence of the weights w_j in Eq. 7.41 that assure that the nodes at smaller depths account for the majority of the impurity reduction. Apparently, rules referring to the phase $x_2 = \phi$ are most discriminatory and therefore account for the bulk of the impurity reduction, while x_1, x_3 are ranked similarly. The nearest-neighbor term on the other hand is much less important. Intuitively this makes sense, since the t_2 vs. m phase diagram is independent of the value of t_1. We note that a sequence of experiments taking into account different depths, as shown here, is rather insightful, since the bulk of the impurity reduction is expected to happen in the early stages of the learning process, while later only incremental changes occur that can blur the information due to the increasing number of nodes at lower depths. This is expected to be most severe in case of overfitting, where the model adapts to even small details in the data.

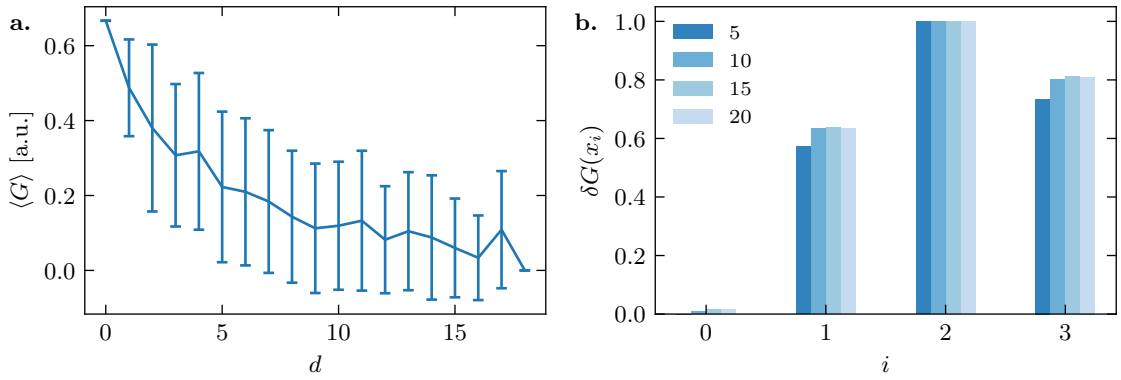

Figure 7.14: **a.** Gini coefficient, cf. Eq. 7.39, as a function of the depth of the tree, plotted as the average over all nodes of the particular depth. The error bars mark the 1σ confidence interval. As expected, the impurity decreases the deeper one descends down the tree. A tree with a maximal depth of 20 was used. At depth 18 the impurity reaches 0, therefore the algorithm terminates. **b.** Impurity reduction, cf. Eq. 7.41, for different maximal depths $d_{\max} \in \{5, 10, 15, 20\}$ as a function of the four features from Eq. 7.32. The values are normalized to the maximal value per experiment. $x_0 = t_1$ is scored consistently low, the other three perform similarly depending on the model.

Finally, we want to take another look at the understanding of the phase diagram that the model has actually captured. If all information were contained, the model should be able to predict the correct phase for a given set of input features, i.e., it should be able to interpolate, but also extrapolate. The evaluation of the accuracy over the test set shown in Fig. 7.11 already captures the model's interpolation capabilities, since both the test and training sets were generated out of the same probability distribution, i.e., both sets of samples lie in the same region of phase space. Given the reference point $\mathbf{x}_{\text{ref}} = (1., 0.2, \pi/2, 1.05)$ and the spread parameter $\alpha = 1.5$, we can determine the bounds for each of the four features

$$\begin{aligned} x_0 = t_1 &\in [-0.5, 2.5), \\ x_1 = t_2 &\in [-0.1, 0.5), \\ x_2 = \phi &\in [-\pi/4, 5\pi/4), \\ x_4 = m &\in [-0.525, 2.625). \end{aligned} \tag{7.43}$$

The half-openness of the intervals is a consequence of the numerical random number generator and does not affect the result.

We use the model trained with $d_{\max} = 20$ and let it create a phase diagram similar to Fig. 6.4. This can be achieved by creating a data set consisting of a regular grid of points and evaluating the model on this grid to obtain predictions for each point. The resulting diagram is shown in Fig. 7.15. At first glance the result looks rather wrong, however, there are some important issues to be aware of. Let us take a look at the t_2 vs. m diagram shown in Fig. 7.15**a** first. Apparently, the transition line is reproduced rather well above $m = 0$. This is not the case for negative m, however. Comparing the values where the performance of the predictor deteriorates with the bounds of the data set Eq. 7.43, we find that the two quantities coincide. Therefore, the bad performance of the model in parts of the phase diagram can be attributed to a high generalization error that in this case corresponds to extrapolation. Points within the bounds of the data set are predicted rather well, although likely none of them were included precisely in the training set. The bad extrapolation performance can be explained by taking a closer look at the decision tree structure. A decision is made taking into account a particular sequence of binary comparison rules. All of these rules combined can be represented as a cuboid in feature space that is assigned a particular class label. Necessarily, this means that the model

function is piecewise constant and since there is no reduction in impurity to be gained by placing constraints outside of the regime provided by the training data, the edge cases are extended to infinity. This implies that a classification obtained for data points at the boundary of the data set will be used as the prediction for points lying outside. In case a particular region is not fully enclosed within the bounds of the data set, this leads to bad extrapolation behavior by construction. The piecewise constant property of the model is apparent in Fig. 7.15**a**. Since it is unrealistic to cover the entire feature space, the only way to improve upon this result would be to include other features, in terms of which the phase transition lines can be represented through thresholds. The definition of these engineered features, however, requires knowledge of the solution and is therefore not straight-forward in a general case.

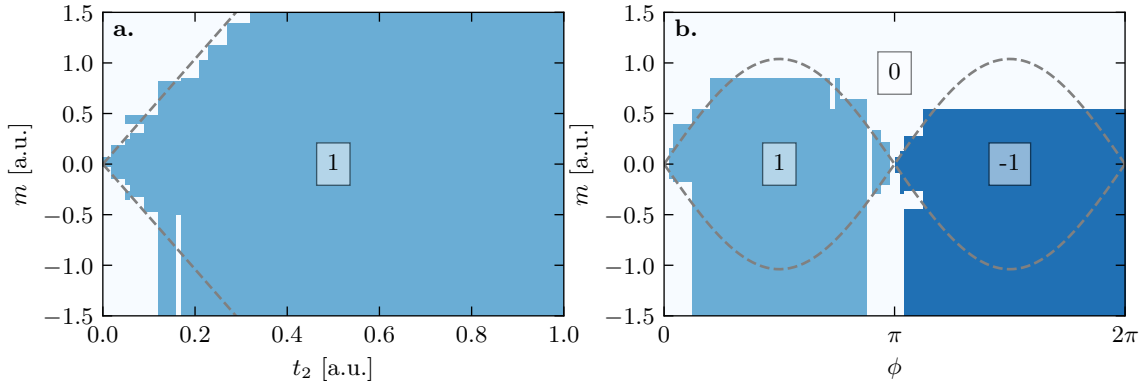

Figure 7.15: Phase diagram produced by the decision tree predictor with $d_{\max} = 20$. **a.** Class label as a function of t_2, m with $t_1 = 1, \phi = \pi/2$ fixed. **b.** Class label as a function of ϕ, m with $t_1 = 1, t_2 = 0.2$ fixed. All quantities in arbitrary units. The exact phase boundaries are shown with dashed lines. The model performs rather well as a function of ϕ, however, extrapolations beyond the bounds of the training set are incorrect.

The same arguments apply to the ϕ vs. m diagram shown in Fig. 7.15**b**, where the top of the $y = 1$ region is rather well approximated but the bottom is extended to negative infinity due to bad extrapolation. Incidentally, the behavior around $\phi = \pi$ is well explained and the extrapolation within the $y = -1$ class region works rather well considering that training data was available only up to 1.25π.

There are several conclusions to be drawn from this study. Most importantly, the model can only be as good as the data. In this case, we disregarded previous knowledge of the phase diagram on purpose, as to construct a generic case study of applications where this knowledge is not available. As a consequence, we cannot expect the model to work particularly well outside of the region of feature space that was provided within the training set. For points within the training set, the model is able to capture the essential properties of the phase diagram. However, with 835 nodes in total and 417 internal nodes the model is rather complex, albeit way less so than expected initially, cf. Table 7.2. Nonetheless, a reconstruction of particular decision paths or an evaluation of different paths is much too cumbersome to be done manually. The initial benefit of the decision tree model of being extremely transparent is suddenly much less interesting due to the amount of complexity contained in the model. Since all splitting rules by definition perform axis-parallel cuts, the description of the phase diagram requires a large number of rules for any transition line that does not lie parallel to an axis[4]. However, as we have seen in Fig. 7.14, this information can be further compressed into an easily interpretable

[4]We note that the process of "pruning", which is typically applied to reduce the complexity of a model, is ineffective for this particular problem, since we are guaranteed to require an infinite number of rules for perfect precision if the transition lines are not axis-parallel.

vector representation $(\delta G(x_i)|i = 0 \dots 3)$ that describes the the weight or importance of each of the features in the decision making process of the model. Clearly, information is lost during this process, however, the information gained is valuable as it makes tangible the correlation between the features on one hand and the topological classification on the other. Extrapolated to higher-dimensional models this method allows the selection of those tight-binding parameters that have the highest impact on the topological classification and thereby improves the understanding of the high-dimensional phase diagram.

7.6.2 Random Forests

Before we go on to discussing our statistical method that is in some ways superior to the traditional machine learning methods discussed in this section, we want to comment on further improvements to the decision tree model. Decision trees are unbiased in nature, however, they have a tendency to overfit [246]. According to the so-called "no-free-lunch" theorem [258] there is no single model that fits any task. Considering this, the results obtained with the simple decision tree algorithm suddenly do not look too bad anymore. Moreover, there is a way to improve on the propensity to overfit: the random forest [259].

One of the problems of the decision tree learning algorithm is the non-optimality of the solution on a global scale, i.e., algorithms that offer practical scaling with the number of features are greedy algorithms that only guarantee to make the best choice locally. The global optimum, however, could require a less favorable choice at a higher node only to generate a much better split further down the decision path. Moreover, decision trees have a rather high variance [246, 247], which means that training two trees on different subset of the same original data set can lead to two completely different tree structures. Of course, this is also a consequence of each individual tree following a greedy optimization strategy and the best choice can vary strongly with the composition of the data set.

In order to allow for more variety, one constructs the random forest as an ensemble learning algorithm that consists, as the name suggests, of an ensemble of different trees. The idea is to reduce the variance by effectively averaging over the predictions of a large number of uncorrelated trees. Clearly, if all trees are drawn from the same statistical distribution, the average or majority vote will lie much closer to the truth than the predictions of a random individual tree. The creation of an ensemble of independent and identically distributed trees is achieved via two methods introduced in Ref. [259]: *bootstrapping* or *bagging* and *random splitting*.

The term bootstrapping refers to the creation of independent trees via their separate training on different data sets. Due to the aforementioned strong dependence of the tree structure on the details of the data set, this can be achieved by using random subsets of the original data set. The bootstrapping process works such that it draws random samples from the original data set with replacement so that one obtains n_{trees} identically distributed data sets that allow for the training of independent trees out of the same original data set.

An additional factor of randomness is introduced via the random splitting technique. Even with the bootstrapped input data it is possible that a small number of features is very descriptive in a sense that they have the largest correlations with the class labels. As a result, all trees will prefer to choose these features when defining a splitting rule in an internal node, which increases the correlations between trees, since they will all look rather similar. This is ameliorated by allowing each tree in a random forest to use only a subset of all features for finding the best split. There are two variants of this procedure described in the literature. In the one discussed in the original publication by Breiman [259], one fixes a random subset of features for each tree initially, that is valid throughout the entire training process. In contrast, as described, e.g., in Ref. [247], one could also draw a random subset at each split node. The former means that all

trees operate only on a subset of features, while the latter allows each tree to effectively use all features, however, only at random for each decision. Despite the differences, both methods achieve the same goal of decorrelating the trees.

We note that random splitting seems rather counterproductive in the present context considering our interest in precisely these most descriptive features that the added randomness seeks to suppress. However, when viewed as a whole, the ensemble still retains this information as we will see below.

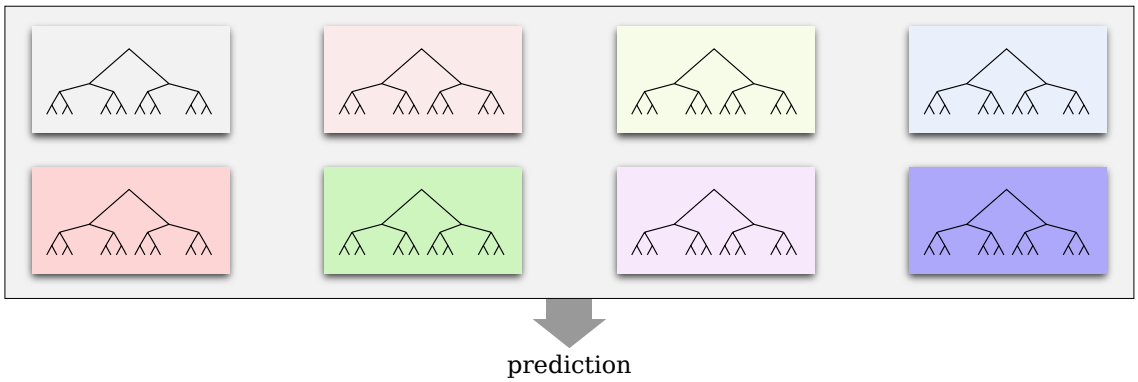

Figure 7.16: Illustration of a random forest consisting of an ensemble of independent trees that are trained on bootstrapped data. Each tree is trained on a different data set and uses different features for splitting. The prediction of the ensemble corresponds to the average prediction of the individual trees.

In Fig. 7.16 we illustrate the random forest classifier. It consists of an ensemble of independent trees that, given input data, each make their own independent predictions. The prediction of the forest is then obtained by averaging over the individual outcomes of the trees. This is either done by a majority vote or by averaging the distributions of the predictive leaf nodes.

We now train a random forest model using the data set 1 from Table 7.1. The random forest has a number of hyperparameters, namely all of those of a normal decision tree plus the size of the ensemble, i.e., the number of trees, the number of samples to draw during the bootstrapping stage and the number of random features to consider for each split. Here, we choose a number of $n_{\text{trees}} = 100$ trees and $n_{\text{bootstrap}} = n_{\text{samples}}$, i.e., each tree is trained on a data set that contains as many samples as the original data set. For the number of features $n_{\text{features}-\text{split}}$ we choose a value of 2. We will elaborate on the meaning of this parameter later. The depth of the trees is unlimited, i.e., they are grown until all leaf nodes are pure (contain only samples with a single class label).

From the trained model we obtain again a phase diagram for the Haldane model in analogy to Fig. 7.15, which is shown in Fig. 7.17. Clearly, the fit to the exact transition lines is much better than with the single tree model and we can no longer see the axis-parallel boundaries of the phases, which are typical for the piecewise constant decision tree model. In fact, this can be understood as a consequence of the averaging procedure. For a single tree we have already motivated that the entire feature space is decomposed into cuboids, each one corresponding to a leaf node and with that to a particular topological phase. The model function therefore does not vary within each cuboid. In case of the forest, each single tree has the same property, however, the location and size of the cuboids vary from one tree to another. Due to the averaging, the cuboids lose their regular shape and the resulting function becomes smoother the larger the ensemble is. We note that the bad extrapolation behavior is still observed for data that lies outside the bounds of the training set. This is expected, since all trees share this incapability to extrapolate, and therefore, the average prediction is the same as the prediction of any individual tree.

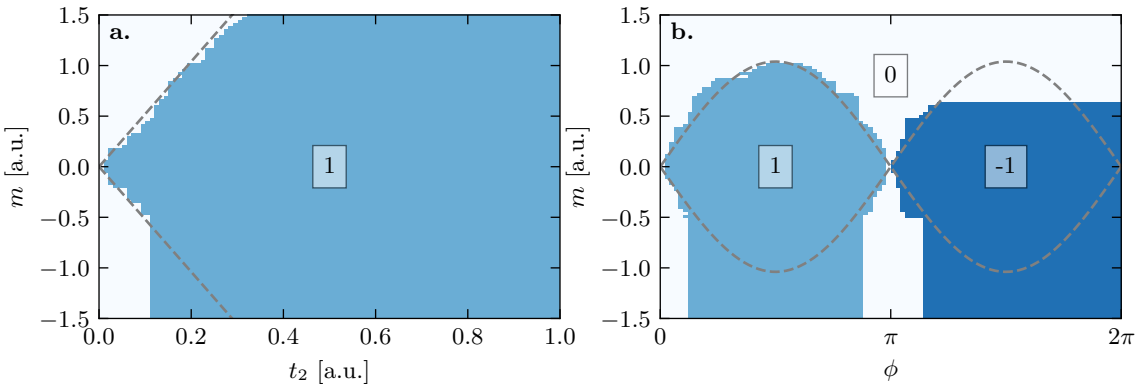

Figure 7.17: Phase diagram of the Haldane model learned by the random forest classifier for the data set 1 of Table 7.1 with the bounds of Eq. 7.43. The predicted phase boundaries are much closer to the exact location (shown as dashed lines) as for a single decision tree. Extrapolation is still not possible, as indicated by the large errors, e.g., for negative m. **a.** $t_1 = 1, \phi = \pi/2$ and **b.** $t_1 = 1, t_2 = 0.2$. t_1, t_2, m in arbitrary units. Hyperparameters: $n_{\text{trees}} = 100, n_{\text{bootstrap}} = n_{\text{samples}}, n_{\text{features-split}} = 2$.

Again, we are interested more in the information that the learning algorithm extracted from the data set rather than the actual prediction that we knew already with much higher precision. Unfortunately, while the random forest is superior to a single decision tree in terms of lower variance, this advantage comes at the cost of interpretability. Clearly, we can no longer look at the decision rules in terms of a tree, since there are now a multitude of independent trees that each make their own decisions. However, we can still use an importance measure such as the impurity reduction of Eq. 7.41 to quantify how much influence each single feature has on the outcome of the classification. To this end, we average over the values obtained for each individual tree according to

$$\delta G_{\text{forest}}(x_i) = \frac{1}{n_{\text{trees}}} \sum_{j \in \text{trees}} \delta G_j(x_i). \tag{7.44}$$

The resulting distribution over the four features is shown in Fig. 7.18**a** for four different values of $n_{\text{features-split}}$, which correspond to all possible values that this parameter can take, namely $1 \leq n_{\text{features-split}} \leq n_{\text{features}}$. $n_{\text{features-split}} = 1$ corresponds to a fully random tree, where in each node a random feature is used irrespectively of the entropy reduction incurred. This allows for highly non-optimal solutions and is therefore less desirable. $n_{\text{features-split}} = n_{\text{features}}$ on the other hand means that at each node all features are available, which corresponds to the usual tree learning algorithm without the additional randomness. In this case, the values 2 and 3 make the most sense and we observe that the reduction of Gini impurity is very similar for both choices.

The reduction of Gini impurity has the disadvantage that it is computed using the samples of the training set. Models that overfit the data can therefore substantially increase the impurity reduction of features that are, in fact, rather unimportant. A similar measure of importance, albeit evaluated on the test set, is given by the so-called permutation importance [259]

$$\text{pi}_i = s - \frac{1}{n_{\text{perm}}} \sum_{j=1}^{n_{\text{perm}}} s_{j,i}, \tag{7.45}$$

where s is the success rate (accuracy) of the model on the test set and $s_{j,i}$ are success rates on n_{perm} modified test sets that are constructed as follows. Given a feature index i, the i-th column of the data matrix $X = (x_{ki})$ is shuffled randomly. As a consequence, the data is not necessarily correct anymore, since the values of feature x_i were randomly swapped between different samples

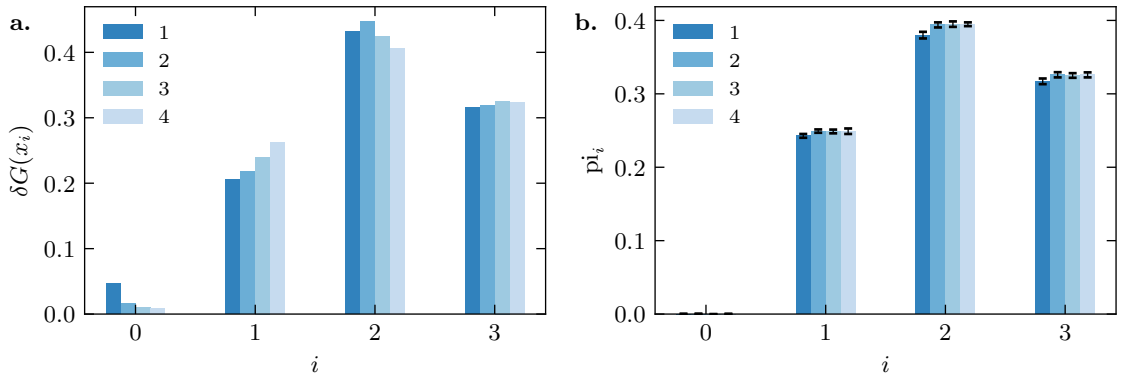

Figure 7.18: a. Gini impurity reduction, cf. Eq. 7.44, for the random forest model for all four features i and different values of $n_{\text{features-split}}$ (colors). The values are rather similar for all choices of $n_{\text{features-split}}$ and the result is in agreement with that of the single decision tree, cf. Fig. 7.14. **b.** Permutation importance (Eq. 7.45) per feature for different $n_{\text{features-split}}$. Values are very similar to the Gini impurity reduction but less dependent on $n_{\text{features-split}}$. Error bars show the standard deviation over the ensemble.

without touching the other features. The idea is rather simple. If a feature is not important for the classification at all and its value therefore irrelevant, the shuffling procedure does not change the success rate, and thus, the permutation importance vanishes. If, on the other hand, a feature is descriptive of the data, then the success rate is expected to change. In order to account for the possibility of lucky shuffles that shuffle only within classes or only outside, one averages over a number of n_{perm} different iterations.

We plot the permutation importance defined in Eq. 7.45 in Fig. 7.18**b** for all possible values of $n_{\text{features-split}}$. The result is rather independent of the number of features considered for each node and only the forest of random trees ($n_{\text{features-split}} = 1$) deviates slightly. This implies that for this simple model the classifier works rather well irrespective of the additional randomness. The permutation importance corresponds qualitatively and quantitatively to the reduction of the Gini impurity, where the phase x_2 is most important, followed by x_3 and x_1, and x_0 is not important at all. We attribute the good agreement to the fact that our data contains no noise by construction and is generated randomly. Therefore, overfitting is not possible.

We close this discussion of tree-based supervised learning of phase diagrams with the conclusion that an extraction of exact correlations between features is rather cumbersome if not impossible for realistic model dimensions. Here, dimensions refers not only to the number of features but also the number of hyperparameters of the model. However, we have seen that information about the importance of individual features can be extracted from a model once it is optimized. This notion is not limited to tree-based algorithms, which we chose based on the premise of interpretability. In fact, the permutation importance can be computed for any classifier, which includes much more flexible logistic regression or neural network models that can handle this type of data much better. A commonality between all of these methods, though, is that the model has to be optimized first. In the following section we discuss our purely statistical approach that extracts similar information without the need to train a particular model.

7.7 Statistical Method

Unless stated otherwise, we will use in the following the features $\mathbf{x} \to \delta\mathbf{x} = \mathbf{x} - \mathbf{x}_{\text{ref}}$ *for symmetry reasons. This shifts the point of reference to zero, and therefore, corresponds only to a global shift of the origin of the phase diagram.*

We have seen during our analysis in the previous section that it is possible—with some

caveats—to train a machine learning model with the data that we generate to learn the phase diagram. However, while we could demonstrate this for a very low-dimensional example, the expectation is that this becomes more complicated with increasing dimension, i.e., the more features are present. The choice of tree-based methods was informed by the inherent simplicity of the underlying model, which promises to provide very convenient insight into the information learned. This transparency is what we need in order to learn from the method, since, in principle, the trained model is just a fit to data that we can easily generate anyway. It turned out that this information is, in general, too complex to comprehend or interpret, however, we found more compressed ways to extract information about the importance of individual features for the topological classification, namely the impurity reduction and permutation importance measures. We will now extract the same information out of the data set itself without the necessity of a middle man (aka the machine learning model).

To begin with, our aim is to extract those features from our labeled data set that are most characteristic for a particular phase. A good measure would gauge the relevance of each feature by its discriminatory power between different class labels. This idea immediately reminds us of the principal component analysis (PCA), where features with the largest variance, i.e., those that describe most of the variation between different points and can therefore be used to discriminate between different classes, are picked out. Since our unlabeled data set has no inherent structure due to the stochastic sampling process that guarantees a particular distribution of points over the entire domain, the variance of features evaluated over all data points is unfortunately non-descriptive and of no value. However, by splitting up the data set into separate purified groups, i.e., grouping together points that have the same class label, we can introduce structure into each of these data sets. We illustrate this point in Fig. 7.19, where we represent the whole data set by a rather structureless disc that contains data points from all different classes combined. Taken apart, however, as shown on the right, each of the resulting subsets has a very non-trivial structure that is rather dissimilar to that of the original data set.

Figure 7.19: Illustration of the structure of the data set. Left: complete data set containing points of all class labels. The structure corresponds to that of the unlabeled data set, we illustrate different labels in terms of different colors for clarity. Right: data set taken apart into three subsets, each containing only one particular class label. Each of these subsets has a non-trivial structure that generally differs from that of the original set and between one another.

We take this argument as motivation to define the following data sets

$$X_l = \{\mathbf{x}_i \in X | y_i = l\} \quad \forall l, \tag{7.46}$$

that each contain only points from a particular class. Having started from completely random data points distributed over an arbitrary domain we have now, through our knowledge of the topological invariant, created data sets that reflect the structure inherent to every individual

topological class. By applying PCA to these data sets one obtains those features that have the largest variance. This does not really make sense, though, since our data is already split into the separate classes and we are now more interested in how these differ from each other. The largest variance is typically held by those features that are not very descriptive of a class label, since the classification is more or less indifferent towards the change of such a variable. On the other hand, those features that show the smallest variance seem much more descriptive, since their values are more confined as a consequence of the added information about the class label.

We could therefore just apply PCA straight-forwardly, only reversing the selection process such that instead of keeping the largest singular values we would keep the smallest ones. However, there are a couple of problems with this approach. Due to the construction of the original data set, the variances of different features are not the same, i.e., there are initially already features that have lower variance by construction. The selection process would have to take into account only the reduction in variance as a consequence of the information contained in the label. Secondly, PCA is a linear method, which means that it tries to find the most descriptive components over all linear superpositions of features. This makes factoring out the individual variances a little more complicated. One solution would be to rescale the data so that all variances are initially the same (for data set X).

7.7.1 Entropy Reduction

We investigate instead another method that yields a more direct measurement of the reduction of variability of the data. Given the full labeled data set (X, Y) we can define the variability of the features in terms of the entropy H, cf. Eq. 4.58, as

$$H(X) = \sum_{i=1}^{n_{\text{features}}} H(X_i), \tag{7.47}$$

since the features are sampled independently. The information about the topological class is encoded in the label vector Y that can be interpreted as another random variable. The entropy of the informed data set, i.e., equipped with the additional label data, is then given by the conditional entropy $H(X|Y)$. The conditional entropy encodes the variability of the data sets X_l, which have been projected on a particular label. With

$$H(X|Y) = H(X, Y) - H(Y) \tag{7.48}$$

and the definition of the mutual information $I(X;Y) = H(X) + H(Y) - H(X, Y)$, cf. Eq. 4.71, we have

$$H(X|Y) = H(X) - I(X;Y), \tag{7.49}$$

i.e., the mutual information measures exactly the reduction of entropy as a consequence of adding the label.

Eq. 7.49 computed without the average over Y and solved for $I(X;Y)$ yields a number per label that quantifies how much the data is compressed upon projecting out all but this one label, which provides a measure of the amount of structure that was added to the data. The quantity underlying all measures here is the probability density function (PDF)

$$\rho_l : \mathbb{R}^{n_{\text{features}}} \to \mathbb{R}, \tag{7.50}$$

which, given a label l, maps each feature vector to a corresponding probability density. In fact, ρ_l is a piecewise constant function of all features.

Proof. To see this, we realize first that given a "complete" set of features in the sense that we have no additional degrees of freedom in the model, each feature vector $\mathbf{x}$ represents a unique configuration. The probability distribution function that ascertains whether or not a particular configuration can appear in a topological phase is thus well-defined by

$$\rho_l(\mathbf{x}_i) = \frac{1}{\mathcal{N}} \begin{cases} 1 & \text{if } y_i = l \\ 0 & \text{else,} \end{cases} \tag{7.51}$$

where $\mathcal{N}$ is a normalization constant. □

The definition of Eq. 7.51 does not take into account an external source of unevenness of points that could be introduced via a sampling procedure that is non-uniform. However, one can always recover the above definition by checking if the measured probability density is finite or not. Using the uniform distribution we will always obtain Eq. 7.51 directly. Since the feature space is unbounded and we typically find a trivial phase at infinity, which corresponds to some limit, ρ_l is generally not normalizable. However, restricting ourselves only to a finite subspace $\Omega \subset \mathbb{R}^{n_{\text{features}}}$, which is already guaranteed by the construction of the data set, ρ_l can be normalized over Ω.

Clearly, the complete information about the topological phase as a function of all features and with that the complete description of the phase diagram is contained in ρ_l. Since we have no prior knowledge and are not inclined to falsify our description through the introduction of a bias, the only way to estimate ρ_l from the data is by binning. This means acquiring a large enough number of samples so that the feature space is rather densely populated and then extracting the probability density by simply counting the number of data points per unit volume. In order to guarantee a dense population, we require a constant number n of samples per feature dimension, i.e., the total number of samples is given by

$$n_{\text{samples}} \sim n^{n_{\text{features}}}. \tag{7.52}$$

Clearly, this is not sustainable, since the exponential scaling with the number of features quickly pushes this out of the manageable regime. For example, with decent $n \sim \mathcal{O}(100)$ the number of samples for only four features is already of the order of 10^8. On the other hand, we were hoping to describe dimensions much larger than that which lie far beyond anything representable graphically, however, Eq. 7.52 indicates that this will not be possible in terms of ρ_l computationally.

Instead, we define the marginal distributions, cf. Eq. 4.12, through

$$p_l : X \to \mathbb{R}, \quad X = \mathbb{R}^m, \quad m < n_{\text{features}}, \tag{7.53}$$

which define a cascade of functions for $1 \leq m < n_{\text{features}}$ that are related to the bare probability density via

$$p_l(\{x_i | i \in S\}) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \rho(\mathbf{x}) \prod_{j \notin S} \mathrm{d}x_j. \tag{7.54}$$

In the following, we will focus on the one-dimensional functions, where $|S| = 1$, i.e.,

$$p_l(x_i) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \rho(\mathbf{x}) \prod_{j \neq i} \mathrm{d}x_j, \tag{7.55}$$

that describe the distribution of a single feature, marginalized over the entire high-dimensional feature space. While the exponential scaling of the complexity for the computation of the

bare probability density ρ_l is equivalent to a grid-based approach, the stochastic nature is now becoming an asset. Numerical integration over a single variable using some sort of quadrature formula, e.g., Simpons's rule, typically has an associated error of $\mathcal{O}(h^4)$ [260], where $h = (b-a)/n$ is the step size used and a, b are the limits of the integration domain. Therefore, the error scales as $\mathcal{O}(1/n^4)$. At the same time, a stochastic Monte Carlo approach [261] decomposes the integral into

$$I \approx \frac{b-a}{n} \sum_{i=1}^{n} f(x_i). \tag{7.56}$$

Since x_i are independently sampled, the variance is given by (cf. Eq. 4.38)

$$\mathrm{Var}[I] = \frac{(b-a)^2}{n^2} \sum_{i=1}^{n} \mathrm{Var}[f(x_i)] = \frac{(b-a)^2}{n} \sigma^2, \tag{7.57}$$

where $\mathrm{Var}[f(x_i)] =: \sigma^2$ depends on the distribution underlying the x_i. The error of the Monte Carlo approximation of the integral can then be approximated by the standard deviation $\delta I = \sqrt{\mathrm{Var}[I]} \sim \mathcal{O}(n^{-1/2})$. This result does not depend on the dimension of the integral, since in the expression Eq. 7.56 only the integral width $(b-a)$ has to be replaced by the volume of the integration domain $V = \int_\Omega \prod_i \mathrm{d}x_i$, the scaling remains $\mathcal{O}(n_{\text{samples}}^{-1/2})$. For Simpson's rule on the other hand, n refers to the number of samples per dimensions, i.e., $n = n_{\text{samples}}^{1/d}$, which means that the total error is $\mathcal{O}(n_{\text{samples}}^{-4/d})$. As a consequence, for constant given error, the number of samples required does not depend on the number of dimensions for Monte Carlo integration, while it scales exponentially for quadrature formulas. This is the main benefit of using statistical means for the description of phase diagrams.

We note that the more features are integrated out, the better the approximation of the marginal distribution becomes at constant number of samples. Practically, we expect that one- and two-dimensional distributions can be obtained with good accuracy, while higher-dimensional quantities require sample sizes that take significantly longer to process. By focusing on one-dimensional distributions, i.e., all features but one are integrated out, we are essentially neglecting correlations between features. This is for sure valid as long as the features are only weakly correlated with each other.

Given the marginal distributions $p(x_i)$ of Eq. 7.55 estimated from both the complete data set and that obtained by projecting out particular labels we can now compute the entropy reduction $I(X;Y)$ as in Eq. 7.49, which provides a measure of the structure added to the distribution of each individual feature upon projecting onto a specific label. The mutual information $I(x_i, y)$ encodes the amount of information a particular feature x_i carries about the phase labeled by y and can be interpreted just like the importance measures that we defined for the decision tree.

We show a chart of the measured entropy reduction for the same data set (set 0*) that we used to train the tree and forest models in the previous section in Fig. 7.20. Subfigure **a** contains the information contained in each feature, evaluated for different class labels. Apparently, no feature contains a lot of information about the trivial class. This is expected, since our premise is that topological phases are somewhat rare, and therefore, the trivial phase must be the generic case that is realized by most configurations. Therefore, the assumption that the system is in a trivial phase is expected to be rather generic and the constraints on the features should be much less severe than for non-trivial phases. And indeed, we observe that the information content of the features for the two non-trivial phases is much larger. While $x_0 = t_1$ is found to contain no information at all, all other features contain similar information about the $y = 1$ phase. Here, $x_2 = \phi$ is ranked more important than $x_3 = m$, followed by $x_1 = t_2$. Incidentally, this reproduces the ranking found by the decision tree and random forest models, cf. Fig. 7.18. This ranking is

not consistent throughout the different phases and, in particular, the $y = -1$ phase seems to be heavily controlled by both x_2 and x_3, which is indicated by the same large value of the mutual information. This makes a lot of sense if we consider again the bounds of the data set given in Eq. 7.43 and the known Haldane phase diagram from Fig. 6.4. Clearly, the overlap with the $y = -1$ phase is much larger in terms of $x_2 = \phi$ than it is for $x_1 = t_2$, while the importance of $x_3 = m$ is estimated to be rather similar to that of x_2. As for the much larger values compared to the $y = 1$ phase, this is a largely artificial consequence of the choice of the reference point. Since $\mathbf{x}_{\text{ref}}$ was chosen close to the transition between $y = 0$ and $y = 1$ and far away from the $y = -1$ phase, the samples with $y = -1$ must be spread over a smaller region, which implies a larger reduction of entropy.

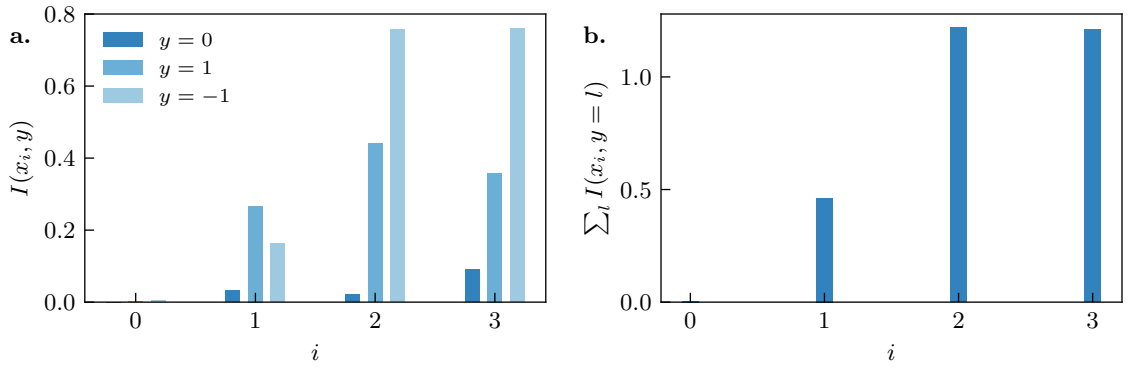

Figure 7.20: Entropy reduction (mutual information) $I(x_i, y)$ of feature x_i and class y. In **a.** we show values with full dependence on both i and y. Feature x_0 does not reduce the overall entropy at all, while x_2 and x_3 contain very strong information about the topological class, in particular $y = -1$. In **b.** we show the information score integrated over all class labels that represents the total information content.

In Fig. 7.20**b**, we show the integrated information content of each feature, where we sum up the contributions from all topological phases, which is directly comparable to the Gini reduction and permutation importance measures discussed for the machine learning models. The value is clearly dominated by the largest contributors, and therefore, we obtain a similar ranking as for $y = -1$.

We have now seen that it is possible to extract similar information out of the data set itself from information theoretic considerations alone, without the need for an expensive model training phase that is required for any usual machine learning algorithm. However, at the same time we have found that the discussed entropy reduction is susceptible to the specific choice of the sample space and is therefore only reliable when comparing phases that are similarly abundant. On the other hand, it reveals immediately that in the portion of the four-dimensional feature space that we investigated, the $y = -1$ phase is clearly rarer than the other two phases, while any of the non-trivial phases are more restrictive on the values of the features than the trivial phase. We therefore have already constructed an algorithm that provides basic information about the topological phase diagram and scales favorably with increasing dimension, i.e., is applicable also for large numbers of features, since all we need to consider are one-dimensional integrated probability densities.

7.7.2 Statistical Distance

We now introduce another way to define the importance of a particular feature that is independent of the shape of the original distribution. In analogy to data science applications and, e.g., the evaluation of experimental data in signal processing, we define the importance of a feature

as the "contrast in the signal". What exactly we mean by "contrast" is illustrated in Fig. 7.21, where we show two different cases. In Fig. 7.21**a**, we look at a single feature x_i and compare the bare distribution $p(x_i)$ with the conditional distributions $p(x_i|y = l_1)$ and $p(x_i|y = l_2)$, where the data is projected onto particular class labels l_1, l_2. From the similarity of the bare distribution to the one projected onto l_1, it is apparent that the values x_i can take for this particular class are not much restricted, and therefore, the information content in x_i pertaining the class l_1 is rather low. In contrast, $p(x_i|y = l_2)$ differs strongly from the bare distribution and also the distribution for class l_1. This indicates that x_i is very descriptive of the phase l_2 and can be used to distinguish it from other classes.

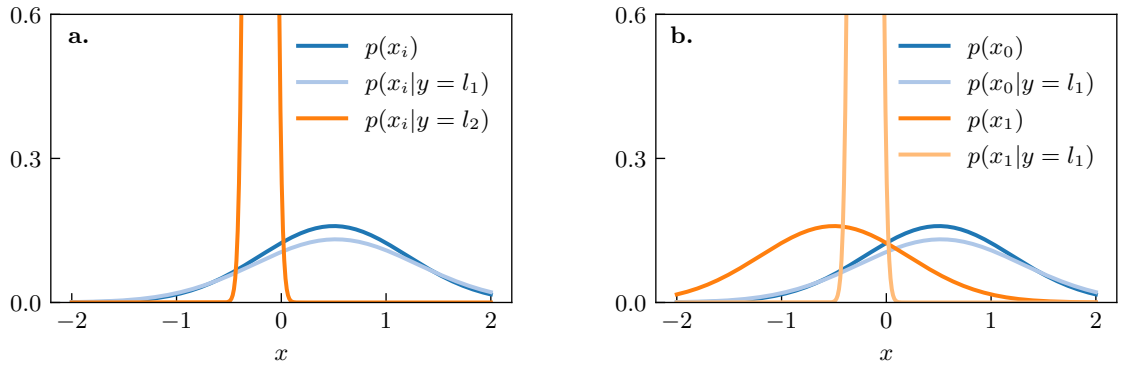

Figure 7.21: Illustration of the feature importance defined as the contrast in the signal. **a.** Marginal distribution for a feature x_i. We show a total distribution and the distributions projected onto particular labels l_1, l_2. x_i is seemingly much more important for the description of l_2 than it is for l_1, since the distributions differ strongly in the former case. In **b.** we show distributions of two features x_0, x_1, bare and projected onto label l_1 for each. For the description of l_1, it is clear that x_1 is more important than x_0. [Figure adapted from Ref. [173]]

Another view onto the same problem is presented in Fig. 7.21**b**, where we show bare distributions of two different features x_0, x_1 and the respective conditional distributions corresponding to the same class l_1. For feature x_0, the difference between the bare and conditional distributions is very small, which indicates low importance. Feature x_1, however, reacts strongly to the information about the class label l_1. Therefore, feature x_1 is a good descriptor of class l_1, while x_0 is not.

The difference between the two views lies in which quantities we compare. In Fig. 7.21**a**, we compare how the same feature reacts to different class labels, while in Fig. 7.21**b**, we compare how different features react to a particular common class label. The second case is more common if we want to find a low-dimensional description of a particular topological phase, while the first offers insight into the relevance of a single feature in general.

Having established pictorially what we want to achieve, we now formalize this analysis that for general applications should offer ways of automation. The requirements we define for our importance measure D are that of a mathematical distance:

i) $D(p, q) = D(q, p)$ (symmetry)

ii) $D(p, q) \geq 0$ and $D(p, q) = 0$ iff $p = q$ (positivity)

iii) $D(p, q) \leq D(p, r) + D(r, q)$ (triangle inequality)

where p, q, r are distribution functions. Not all properties are equally important, though. While positivity is required for an interpretation as an importance score, neither the triangle inequality nor symmetry are necessary. However, a symmetric function would be preferable in order to

reduce the ambiguity in the definition. The triangle inequality on the other hand assures that the distance D takes the smallest value possible, which in this context can be imagined as morphing one distribution in any way possible into the other should at no point in between assume a shape that is closer to both initial distributions than half their distance. So, if we let $p_{t=0} = p$ and $p_{t=1} = q$ for $t \in [0, 1]$, then

$$D(p, q) \leq D(p, p_t) + D(p_t, q) \quad \forall t, \tag{7.58}$$

and, in particular, for the "middle", where $D(p, p_{t^*}) = D(p_{t^*}, q)$,

$$D(p, p_{t^*}) \geq \frac{D(p, q)}{2}. \tag{7.59}$$

The triangle inequality therefore guarantees that we do not overestimate the importance of a parameter by neglecting the most direct path. We will see in practice that this is not always beneficial.

In Sec. 4.3.1, we have already introduced the Kullback-Leibler divergence [134, 262] as a measure of the information contained in a distribution in contrast to a ground truth. The KL-divergence was introduced by Kullback and Leibler as the "mean information for discrimination between [two hypotheses]" [134], which is exactly what we are looking for. Unfortunately, D_{KL} satisfies only the positivity requirement and neither symmetry nor triangle inequality as we have seen in Sec. 4.3.1. There is a different version, however, which Kullback refers to in his book as "the divergence" [262], that satisfies also the symmetry requirement:

$$D^*_{\mathrm{KL}} = \int_{-\infty}^{\infty} (p(x) - q(x)) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x. \tag{7.60}$$

A different measure was defined by A. Bhattacharyya a few years earlier [263, 264] and is usually referred to as the "Bhattacharyya distance" [265]. We define the Bhattacharyya distance as

$$D_{\mathrm{B}}(p, q) = -\log\left[\sum_i \sqrt{p(x_i)q(x_i)}\right] \tag{7.61}$$

for discrete distributions and for continuous distributions

$$D_{\mathrm{B}}(p, q) = -\log\left[\int_{-\infty}^{\infty} \sqrt{p(x)q(x)}\,\mathrm{d}x\right]. \tag{7.62}$$

The symmetry property is obvious, since p and q appear only as a symmetric product. For the positivity we find

$$\sum_i \sqrt{p(x_i)q(x_i)} = \sum_i \sqrt{p(x_i)}\sqrt{q(x_i)} \tag{7.63}$$

$$\leq \left(\sum_i p(x_i)\right)\left(\sum_i q(x_i)\right) \tag{7.64}$$

$$= 1, \tag{7.65}$$

where we used the Cauchy-Schwarz inequality in the form

$$\left(\sum_i v_i w_i\right)^2 \leq \left(\sum_i v_i^2\right)\left(\sum_i w_i^2\right). \tag{7.66}$$

Equality holds only if v, w are linearly dependent. With the monotony of the logarithm and $\log(x) \leq 0$ for $x \in [0, 1]$ the positivity of D_{B} follows immediately. The exact same equations apply for the continuous form. The triangle inequality on the other hand is not satisfied, which is best demonstrated by a counterexample. Suppose we are considering discrete distributions p, q, r with two events and probabilities

$$\begin{aligned} p_1 &= 1, \quad p_2 = 0 \\ q_1 &= 0, \quad q_2 = 1 \\ r_1 &= \frac{1}{2}, \quad r_2 = \frac{1}{2}. \end{aligned} \tag{7.67}$$

Then, we have

$$D_{\mathrm{B}}(p, q) = -\log(0) > D_{\mathrm{B}}(p, r) + D_{\mathrm{B}}(r, q) = 2\log(\sqrt{2}), \tag{7.68}$$

i.e., the triangle inequality is violated. There is a way, however, to modify the definition from Eq. 7.62 of the Bhattacharyya distance in such a way that the triangle equality is restored. Suppose we define the Hellinger distance [266, 267]

$$D_{\mathrm{H}}(p, q) = \sqrt{1 - \int_{-\infty}^{\infty} \sqrt{p(x)q(x)}\, \mathrm{d}x}. \tag{7.69}$$

Then, clearly the symmetry still holds and according to Eq. 7.65, the integral is bounded from above by 1, and therefore,

$$D_{\mathrm{H}}(p, q) \geq 0. \tag{7.70}$$

As in the case of the Bhattacharyya distance, $D_{\mathrm{H}} = 0$ can only be achieved if the integral is precisely equal to 1, and therefore, this implies $p = q$. In fact, the integral is also bounded from below since for probability distributions $p(x), q(x) \geq 0$ we have $\int \sqrt{p(x)q(x)}\mathrm{d}x \geq 0$, and therefore, $D_{\mathrm{H}} \leq 1$. With regard to the triangle inequality, we find that

$$\begin{aligned} \int_{-\infty}^{\infty} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x &= \int_{-\infty}^{\infty} p(x)\, \mathrm{d}x + \int_{-\infty}^{\infty} q(x)\, \mathrm{d}x - 2\int_{-\infty}^{\infty} \sqrt{p(x)q(x)}\, \mathrm{d}x \\ &= 2 - 2\int_{-\infty}^{\infty} \sqrt{p(x)q(x)}\, \mathrm{d}x, \end{aligned} \tag{7.71}$$

where we used the normalization of p and q, and therefore,

$$\int_{-\infty}^{\infty} \sqrt{p(x)q(x)}\, \mathrm{d}x = -\frac{1}{2}\int_{-\infty}^{\infty} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x + 1. \tag{7.72}$$

The definition of Eq. 7.69 then becomes

$$D_{\mathrm{H}}(p, q) = \frac{1}{\sqrt{2}} \sqrt{\int_{-\infty}^{\infty} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mathrm{d}x}, \tag{7.73}$$

which we can also express through the Euclidean norm

$$D_{\mathrm{H}}(p, q) = \frac{1}{\sqrt{2}} \parallel \sqrt{p} - \sqrt{q} \parallel_2, \tag{7.74}$$

	definition	pos.	symm.	Δ-ineq.
KL-divergence	$D_{\mathrm{KL}}(p,q) = \int_\Omega p(x) \log\left(\frac{p(x)}{q(x)}\right)$	+		
KL-divergence*	$D^*_{\mathrm{KL}}(p,q) = \int_\Omega (p(x)-q(x)) \log\left(\frac{p(x)}{q(x)}\right)$	+	+	
Bhattacharyya dist.	$D_{\mathrm{B}}(p,q) = -\log\left[\int_\Omega \sqrt{p(x)q(x)}\,\mathrm{d}x\right]$	+	+	
Hellinger dist.	$D_{\mathrm{H}}(p,q) = \frac{1}{\sqrt{2}} \parallel \sqrt{p} - \sqrt{q} \parallel_2$	+	+	+
Area dist.	$D_{\mathrm{A}}(p,q) = \parallel p - q \parallel_1$	+	+	+
Inf. dist.	$D_{\mathrm{inf}}(p,q) = \parallel p - q \parallel_\infty$	+	+	+

Table 7.3: Statistical distances, their definitions and their properties positivity, symmetry and triangle inequality. All functions satisfy positivity, and most are also symmetric. The triangle inequality is only found for the Hellinger distance Eq. 7.69, the area distance Eq. 7.78 and the infinity distance Eq. 7.80.

where $\|\cdot\|_2$ acts on the space of functions $p, q : \mathbb{R} \to \mathbb{R}$. It then follows immediately that

$$D_{\mathrm{H}}(p,q) = \frac{1}{\sqrt{2}} \parallel \sqrt{p} - \sqrt{r} + \sqrt{r} - \sqrt{q} \parallel_2 \tag{7.75}$$

$$\leq \frac{1}{\sqrt{2}} \left(\parallel \sqrt{p} - \sqrt{r} \parallel_2 + \parallel \sqrt{r} - \sqrt{q} \parallel_2 \right) \tag{7.76}$$

$$= D_{\mathrm{H}}(p,r) + D_{\mathrm{H}}(r,q). \tag{7.77}$$

We offer a fourth variant that is somewhat based directly on our consideration of Fig. 7.21, where we looked at how different two distributions appear to the eye. This can be quantified in terms of the area between the graphs of two distribution functions

$$D_{\mathrm{A}}(p,q) = \int_{-\infty}^{\infty} |p(x) - q(x)|\,\mathrm{d}x, \tag{7.78}$$

which is identical to twice the total variation distance D_{tot} [268]. Apparently, D_{A} is symmetric and positive. The latter property follows immediately from $\int f(x)\,\mathrm{d}x = 0 \Rightarrow f(x) = 0\ \forall x$, since f itself is positive. We can rewrite D_{A} in terms of the 1-norm as

$$D_{\mathrm{A}}(p,q) = \parallel p - q \parallel_1, \tag{7.79}$$

which implies the triangle inequality by the same argument as for the Hellinger distance, see Eq. 7.77. In analogy to these examples, we can also define a distance in terms of the infinity-norm

$$D_{\mathrm{inf}} = \parallel p - q \parallel_\infty = \sup_{x \in \mathbb{R}} |p(x) - q(x)|. \tag{7.80}$$

Most of these definitions are special cases of the so-called f-divergence [269, 270]

$$D_f(p,q) = \int_\Omega f\left(\frac{p(x)}{q(x)}\right) q(x)\mathrm{d}x, \tag{7.81}$$

where f is a convex function.

We summarize the properties of the different "distances" in Table 7.3. In the following we will compare results of the various measures in order to select the one that suits our needs best. To this end, we start with a very artificial problem, namely the probability distributions from Fig. 7.21. The illustrative figure was created using a Gaussian-shaped distribution with different variance and mean parameters. We focus here on the functions shown in Fig. 7.22**a**, which

correspond to those of Fig. 7.21**a**, and compute the distances between pairs of distributions. Here, we simply choose labels $l \in \{0, 1, 2\}$ to distinguish the three different distrubtions with no particular meaning assigned to the label. We take as a reference the distributions p_0 and compute the distances to p_0, p_1, p_2. The distance $D(p_0, p_0)$ serves as a sanity check of our implementation, since this distance is guaranteed to vanish due to the positivity of all distance functions. For reference, we provide the parameters of the three different functions in Table 7.4.

	μ	σ
p_0	1.0	0.5
p_1	1.1	0.52
p_2	0.1	-0.2

Table 7.4: Parameters (mean μ and standard deviation σ) of the Gaussian distributions of Fig. 7.22.

The distances computed with all symmetric measures from Table 7.3 are shown in Fig. 7.22**b**. As expected, $D(p_0, p_0)$ vanishes and $D(p_2, p_0)$ is larger than $D(p_1, p_0)$ for all choices of the distance measure, however, the different measures differ in the details. We use a logarithmic scale here in order to represent all values in the same plot. This is necessary, since D_{KL} is an order of magnitude larger than the other values for $D(p_2, p_0)$. At the same time, $D_{\mathrm{B}}(p_1, p_0)$ is significantly smaller than all other values. This indicates that the contrast obtained by D^*_{KL} and D_{B} is expected to be much larger than for the other measures. In fact, the values for $D_{\mathrm{H}}, D_{\mathrm{A}}$ and D_{inf} differ by about two orders of magnitude between the two labels, while for D^*_{KL} and D_{B} we observe a difference of three orders of magnitude.

A large enough contrast is necessary in order to be able to distinguish important features from those that are less important in an automated algorithm, where we do not inspect the distribution functions visually. The observation of a clear jump in the distances between different subsets of features can then be used to define a threshold, below which features are considered unimportant. On the other hand, there are also advantages to D_{H} and D_{A}, namely $0 \leq D_{\mathrm{H}} \leq 1$ guarantees a fixed scale and so does $0 \leq D_{\mathrm{A}} \leq 2$, which follows trivially by considering two distributions with no overlap. D^*_{KL} and D_{B} on the other hand are not bounded from above and, in fact, diverge in the case of completely separated distributions. Weighing these two arguments against each other, we favor the larger contrast over the fixed scale, since the example discussed here features a rather artificial case and we do not expect the differences in real data to be as large. A propensity towards a higher contrast is therefore desirable.

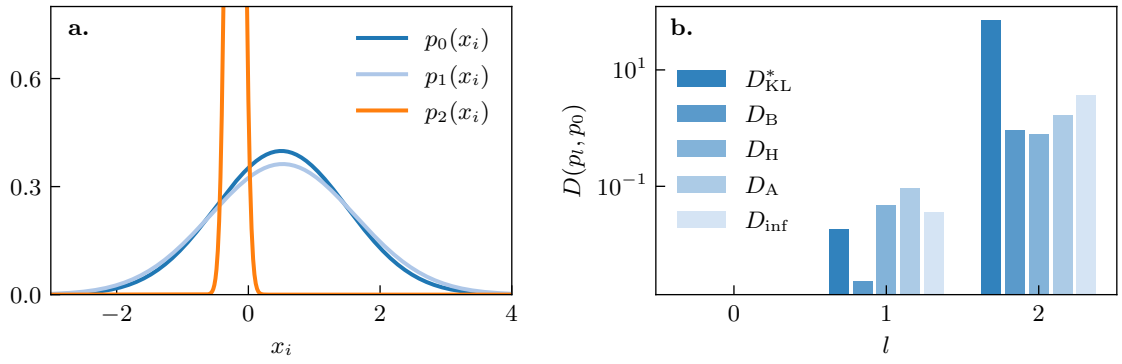

Figure 7.22: a. Distribution functions p_0, p_1, p_2. Here, we use Gaussian distributions with the parameters provided in Table 7.4. p_0 and p_1 are rather similar, while p_2 deviates strongly. **b.** Distances obtained with all five symmetric distance functions from Table 7.3. $D(p_0, p_0)$ vanishes as expected and all measures find $D(p_2, p_0)$ to be much larger than $D(p_1, p_0)$ (note the logarithmic scale).

We now return to the data set that we obtained for the Haldane model and apply the various distance measures in order to determine the most important features. While the importance is simply defined by the distance between two distribution functions, the reference has to be chosen carefully. Since we are mainly interested in the distinction between topologically non-trivial and trivial data points, it makes sense to compare for a given non-trivial class label the marginal distributions of all features with those corresponding to the trivial label. The importance of a feature x_i (or feature-importance of index i) for the description of a particular label l is therefore defined as

$$\mathrm{FI}_l(i) = D(p(x_i|y=l), p(x_i|y=0)), \tag{7.82}$$

where D can be any distance measure from Table 7.3.

In order to compute the feature-importance score, we have to estimate the marginal distributions from the data set. This is done simply by performing a binning analysis on a fixed grid. Suppose the data is bounded by $x^i_{\min}$ and $x^i_{\max}$ for feature x_i, i.e., $x^i_{\min} \leq x_i \leq x^i_{\max}$. Then, we define an equidistant grid through $x_{i,m} = x^i_{\min} + hm$ with $h = (x^i_{\max} - x^i_{\min})/n$ for some $n \in \mathbb{N}$ and $0 \leq m \leq n$. The value of the marginal distribution can then be evaluated at n points

$$b_{i,m} = \frac{x_{i,m} + x_{i,m-1}}{2}, \quad 1 \leq m \leq n \tag{7.83}$$

in terms of

$$\begin{aligned} p(b_{i,m}) &= \mathcal{N} \left|\{\mathbf{x}|\mathbf{x} \in X, x_{i,m-1} \leq x_i \leq x_{i,m}\}\right| \\ &\propto P(x_{i,m-1} \leq x_i \leq x_{i,m}), \end{aligned} \tag{7.84}$$

where $\mathcal{N}$ is a normalization constant. p can be normalized either as

$$\sum_m p(b_{i,m}) = 1 \tag{7.85}$$

or

$$\int_{x^i_{\min}}^{x^i_{\max}} p(x)\,\mathrm{d}x = 1, \tag{7.86}$$

where the implementation of the latter makes use of more robust integration routines such as the trapezoidal or Simpson's rules. We choose to use the Simpson's rule for all integration purposes. The expression given in Eq. 7.84 corresponds to the actual distribution function in the limit of large n_{samples}, which is guaranteed by the law of large numbers. Taking a fixed number of bins n, we have

$$\lim_{n_{\text{samples}} \to \infty} p(b_{i,m}) = \frac{1}{h} \int_{x_{i,m-1}}^{x_{i,m}} p(x)\,\mathrm{d}x, \tag{7.87}$$

i.e., the discrete values correspond to averages of the true distribution over a finite interval. The single external parameter introduced here is the resolution or step size h, which depends on the number n of bins $b_{i,m}$ at which the distribution is estimated. In principle, the larger n the better, since the error of the integration method scales as $\mathcal{O}(1/n^4)$ [260]. However, larger n reduces the number of samples in each bin so that the law of large numbers is compromised. In order to obtain a good estimate at manageable computational effort, a delicate balance between n and n_{samples} is required.

In this case, we have the bounds for each feature given by

$$\begin{aligned} x_0 &= t_1 \in [-1.5, 1.5), \\ x_1 &= t_2 \in [-0.3, 0.3), \\ x_2 &= \phi \in [-3\pi/4, 3\pi/4), \\ x_4 &= m \in [-1.575, 1.575), \end{aligned} \tag{7.88}$$

cf. Eq. 7.43, which are all $\mathcal{O}(1)$. For the resolution we tried values in between $n = 40$ and $n = 80$ and did not observe much difference in the results obtained. A larger number of samples would certainly be preferable, however, for comparability with previous results we take here the exact same data set. The resulting estimates obtained via Eq. 7.84 of the marginal probability density functions for each of the four features is shown in Fig. 7.23**a-d**. For the functions shown here we used $n = 40$ to obtain somewhat smoother lines. In contrast to the artificial Gaussian distribution functions that we showed in Fig. 7.22, the functions we observe here are not necessarily localized, i.e., they are normalizable only because of the finite sample space Ω. This is a consequence of the topology of the phase diagram. In general we cannot assume that a specific phase is contained within some bounded region and in case of a finite number of distinct phases this is even guaranteed not to be the case for at least one phase that extends to infinity. This is not a problem at all, though, since we always operate on finite subspaces that allow for a description in terms of a normalizable distribution function.

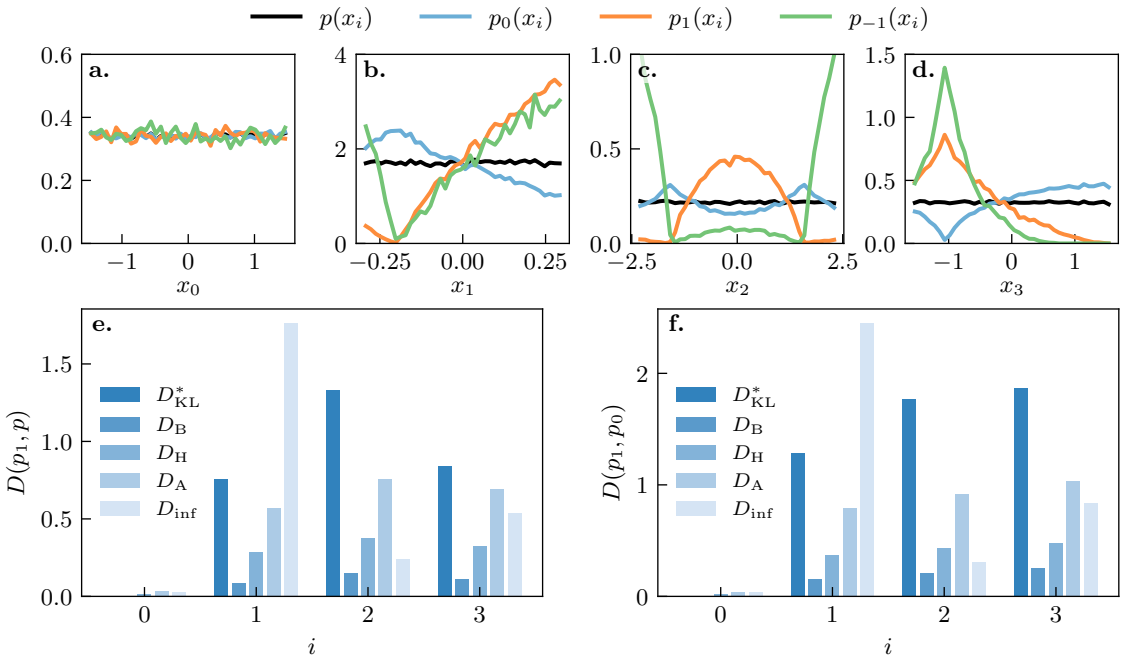

Figure 7.23: Marginal distribution functions $p_l(x_i) = p(x_i|y = l)$ for label l and bare distribution $p(x_i)$ for all four features (**a.-d.**). The marginal distributions of x_0 are indistinguishable from the bare distribution, for all other features we observe the characteristic domains for the different phases. **e.** Statistical distance of the $l = 1$ phase w.r.t. the bare distribution. We observe the same behavior as in Fig. 7.20 and Fig. 7.18. **f.** Statistical distance of the $l = 1$ phase w.r.t. the trivial phase. Again, features x_1, x_2, x_3 are most descriptive.

Comparing the different marginal distributions for our four features, we notice immediately that $p_l(x_0)$ are the same for all labels l, and therefore also correspond to the bare distribution $p(x_0)$ over the entire data set. Clearly, this is an indication of the fact that x_0 does not contain

any information about the topological classification and is therefore a candidate to be stripped from our model. For all other features we observe large differences between all distributions. In general, all are very different from the bare distribution and, in particular, for x_1 and x_3 $p_1 = p(\cdot|y = 1)$ and $p_{-1} = p(\cdot|y = -1)$ are rather similar, while for x_2 all p_l differ significantly from one another. Of course, we understand these findings through our knowledge of the phase diagram. Larger values of x_1 drive the system deeper into the topological phase (with the label l depending on the sign of the phase $\phi = x_2$). For x_2 we observe clearly that $l = 1$ appears only for $-\pi/2 < x_2 < \pi/2$, which is exactly the extent of the lobe in the phase diagram, cf. Fig. 6.4. Beyond that we enter the lobe of the $l = -1$ phase, which is also reflected in the distributions. Note that all features are shifted by $\mathbf{x}_{\text{ref}} = (1., 0.2, \pi/2, 1.05)$ w.r.t. the usual model parameters from Eq. 6.35.

We now investigate the statistical distance, i.e., the information contained within the marginal distributions. There are, in principle, two different viewpoints on what information to describe that we want to discuss here: i) the structure imposed on the data set through each label and ii) the amount of distinguishability between labels offered by each feature. Apparently, the former is what, e.g., the permutation importance of our tree model in Sec. 7.6 and the entropy reduction, see Sec. 7.7.1, describe. We can extract this information by computing the statistical distance between the marginal distribution of any feature and label and the corresponding bare distribution for that feature. Apparently, this does nothing other than measure the structure of the labeled data set in terms of the distribution functions. We find in Fig. 7.23**e** the same qualitative result as obtained by the entropy reduction in Fig. 7.20, where x_0 imposes no structure while x_1, x_3 and x_2 do so in increasing order. This result is obtained with all distance measures except D_{inf}, where the order is reversed. We can explain this agreement by noting that the entropy reduction of Eq. 7.49 is given by

$$
\begin{aligned}
I(X, Y = l) &= -\int p(x|y = l) \log\left(\frac{p(x|y = l)}{p(x)p(y = l)}\right) \mathrm{d}x && (7.89)\\
&= -\int p(x|y = l) \log\left(\frac{p(x|y = l)}{p(x)}\right) \mathrm{d}x + \int p(x|y = l) \log\left(p(y = l)\right) \mathrm{d}x && (7.90)\\
&= D_{\text{KL}}(p(x|y = l), p(x)) + p(y = l) \log\left(p(y = l)\right) && (7.91)\\
&= D_{\text{KL}}(p(x|y = l), p(x)) - H(Y = l), && (7.92)
\end{aligned}
$$

i.e., the entropy reduction is equal to the KL-divergence up to a constant that corresponds to the Shannon information content of the event $Y = l$.

The type ii) of information, i.e., the descriptiveness of a given feature w.r.t. a given topological phase compared to the trivial phase, is shown in Fig. 7.23**f**. Here, we focus on the $l = 1$ phase. Again, x_0 is found to be unimportant, while the other three features show a strong signal. The importance scores are all rather high so that we need all parameters to describe the topological phase. We focus instead again on the comparison between the different distance measures. All but D_{inf} show the same general behavior, which ranks the three features similarly and is compatible with our impression of the distributions. We therefore dismiss D_{inf} here as it does not accurately describe the information we are interested in. All other methods are, in principle, equivalently suited for this particular data set, however, rescaling all quantities reveals that again D^*_{KL} and D_{B} show the highest contrast.

7.7.3 Increased Dimension – Benchmark

We now increase the dimensionality of the problem by changing the model from the original description in terms of the four Haldane parameters of Eq. 6.35 to the more general notion described in Sec. 7.5 and, in particular, Eq. 7.16 and Eq. 7.19, where the features correspond

to hopping matrix elements $t_{ij}(\mathbf{R})$. We note again that we generally deviate from Eq. 7.19 by examining instead the deviation from the reference point, which is given by

$$\mathbf{x}_{\text{ref}} = (t_1, t_2, \phi, m) = (1, 0.2, \pi/2, 1.05), \tag{7.93}$$

and therefore, translate the entire phase diagram by $\mathbf{x}_{\text{ref}}$. This only translates the distributions of real features and creates better contrast in the phase of complex features due to the removal of the bias of $\mathbf{x}_{\text{ref}}$ as we will demonstrate.

Instead of the four-parameter model of Eq. 6.35, we now define

$$H = \sum_{\langle i,j\rangle} t_1^{ij} c_j^\dagger c_i + \sum_{\langle\langle i,j\rangle\rangle} t_2^{ij} c_j^\dagger c_i + \sum_i \varepsilon_i c_i^\dagger c_i, \tag{7.94}$$

which represents a general tight-binding model with up to next-nearest neighbor hopping. For the honeycomb lattice we distinguish two sublattices A, B that give rise to onsite terms ε_i with $\varepsilon_A - \varepsilon_B = 2m$, three nearest neighbor terms t_1 and six next-nearest neighbor terms t_2. Naturally, H must be hermitian, and therefore, ε_i must be real, since $(c_i^\dagger c_i)^\dagger = c_i^\dagger c_i$. There exist no additional restrictions for any other parameters, though, which leaves us with a total of nine complex and two real features. We choose here to remove a source of redundancy, namely the energy scale, by setting one of the onsite terms to be dependent on the other ($\varepsilon_A = -\varepsilon_B$). Consequently, $\varepsilon_B = -m$. This leaves us with the following feature vector[5]

$$\begin{aligned}\mathbf{x} &= \left(\varepsilon_B, \varepsilon_B, t_1^1, t_1^2, t_1^3, t_2^1, t_2^2, t_2^3, t_2^4, t_2^5, t_2^6\right) \\ &= \left(-\varepsilon_B, \varepsilon_B, t_1^1, t_1^2, t_1^3, t_{2A}^1, t_{2B}^1, t_{2A}^2, t_{2B}^2, t_{2B}^3, t_{2A}^3\right),\end{aligned} \tag{7.95}$$

where $t_{2A/B}$ connect the A/B sublattice, respectively. The order of the next-nearest neighbor terms has historical reasons and arose from a specific ordering of the displacement vectors in real space. The feature vector $\mathbf{x}$, introduced in Eq. 7.95, contains 9+1 (complex+real) degrees of freedom, which correspond to at least 19 real features depending on the real→complex mapping. In general, we will consider both mappings described in Eq. 7.22, which generates a total of $1 + 4 \times 9 = 37$ features. Clearly, both 19 and 37 are much larger than 4, the threshold that we defined for conventionally comprehensible data sets in Sec. 7.2, and therefore, this system already provides a good testbed for our methods.

We proceed by sampling all 9+1 features from a uniform distribution on Ω, cf. Eq. 7.24, with spread $\alpha = 2$. The sample size is $n_{\text{samples}} = 10^7$. In Fig. 7.24, we give a brief overview over the data set. The reference point is shown in Fig. 7.24**a** in relation to the t_2-vs.-m phase diagram. In Fig. 7.24**b**, we show the nine distinct hopping paths taken into account. Since the Hamiltonian is hermitian, the opposite direction is always given by the complex conjugate, which we have not drawn. In Fig. 7.24**c**, we show the composition of the data set in terms of the labels. Given our choice of the sample space, we only found the three Haldane phases given by $y = 0, 1, -1$. Out of these, the trivial phase makes up the majority of the data set with over 80%, while $y = 1$ weighs in at 12% and $y = -1$ only at about 4%. The much larger abundance of $y = 1$ samples is a consequence of the choice of the reference point. As indicated in Fig. 7.24**a**, the reference point is much closer to the $C = 1$ phase, which implies that the overlap of the sample space with the domain of the corresponding phase is also larger compared to that of the more distant $C = -1$ phase. The dominance of the trivial phase albeit the initial closeness to a non-trivial phase is a consequence of the generality of the feature space. By allowing all

[5] Note that we could have also set $\varepsilon_A = 0$. We will generally follow this strategy further down the line, however, in this case the current definition of Eq. 7.95 makes sense, since it allows for easier comparison with the Haldane model. The sampled values of m differ between the two options. With $\alpha = 2$ we would sample $m \in [-1.05, 3.15)$ for $\varepsilon_A = 0$ instead of $[0, 4.2)$, i.e., this choice shifts the range of values.

symmetries between parameters to be broken, the model hosts also a large number of samples that bear only little if any similarity to the Haldane model. In this enlarged space we expect the trivial phase to be most dominant as this is considered the generic case. In addition, we find that not all samples are insulators, which we define through a finite band gap $\gtrsim 4 \times 10^{-2}$. The threshold value has been chosen consistently by taking into account the k-grid resolution used and is explained later in more detail. The same argument that we used to explain the abundance of samples with $y = 0$ applies here, since with a more general model, the band structure becomes rather complicated, and therefore, more possibilities for (accidental) gap closings arise. We note that a parameterization $H(k) = a(k)\mathrm{Id} + \mathbf{h}(k) \cdot \boldsymbol{\sigma}$ similar to Eq. 6.37 in terms of Pauli matrices still applies and the band gap can only close for $|\mathbf{h}(k)| = 0$. Due to the increased complexity, however, this can happen at arbitrary k-points, which makes it much more likely than for the symmetric parameterization in the Haldane model. In this case, most samples are insulating, though, so that we do not have to worry about the integrity of our data in general.

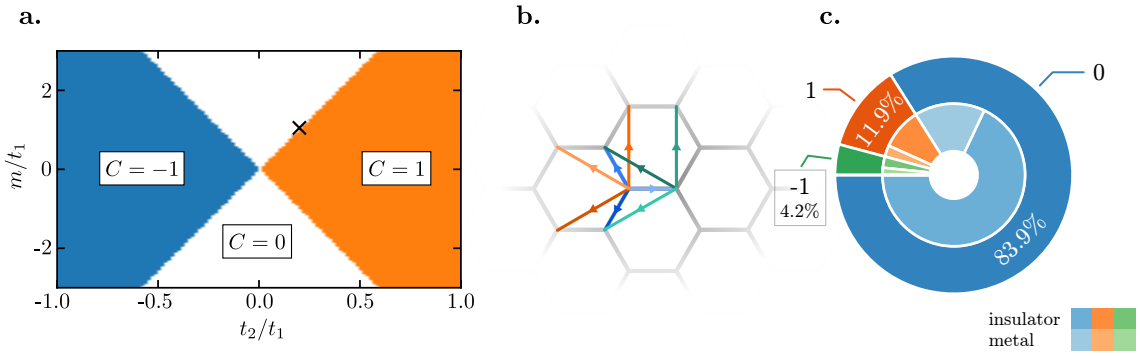

Figure 7.24: **a.** Haldane phase diagram with $\times$ marking the reference point $\mathbf{x}_{\mathrm{ref}}$. **b.** Honeycomb lattice with distinct hopping links drawn in different colors (blue=t_1, orange=t_{2A}, teal=t_{2B}) and the direction that corresponds to the chosen features indicated by arrows. Opposite links that are related to the links shown by complex conjugation are not drawn. **c.** Statistical composition of the data set. Most of the samples are trivial, with only roughly 16% corresponding to non-trivial samples (outer ring). Most samples are, in fact, insulators (inner ring). [Figure adapted from Ref. [173] based on different data]

It is already apparent at this point that we cannot perform the same visual analysis as in Fig. 7.23**a-d**, at least not in an illustrative way, since the resulting plot would barely fit onto a page. It is therefore necessary to first extract the amount of information carried by each feature to pick out the most relevant specimen. We do this by computing the Bhattacharyya distance between the marginal distribution of the topological $y = 1$ phase and the trivial phase. A comparison between all suitable distance measures indicated that the ranking of D_{B} is most informative based on the larger contrast. The overall result does not depend on the choice of D up to the scale. Due to a slightly reduced propensity for very large values we decided to use D_{B} rather than D^*_{KL}, which has a similar contrast.

For the estimation of the marginal probability density functions we use a grid of $n = 80$ points, which results in rather smooth functions. The importance score for all parameters is shown in Fig. 7.25**a**, where we assigned each group corresponding to an equivalent hopping path a single label, e.g., $\varphi(t_1)$ for the phase of t_1^1, t_1^2, t_1^3. We notice immediately that equivalent features have similar rankings. This is not entirely surprising, since although the global C_6 rotation symmetry is already broken for the Haldane model, the individual terms are still interchangeable. This can be understood by considering the nearest neighbor terms. All t_1 paths go from either the A sublattice to the B sublattice or vice versa. Interchanging the t_1^i among each other therefore merely corresponds to a redefinition of the lattice vectors. Of course, in principle, all

features would have to be interchanged according to this new set of lattice vectors, however, since we are considering integrated quantities this does not matter. As a consequence, the marginal probability distribution function is always the same for each set of equivalent features, however, this applies only to the marginal distributions.[6]

Starting from the top of Fig. 7.25**a**, we find that the local onsite term is the most important descriptor for the topological phase, which is compatible with the result obtained for the four-parameter model. Surprisingly, it is followed by the real part of the nearest-neighbor term and its phase. Within the Haldane parameterization, where it is assumed that all t_1^i are the same, we found that they are entirely non-descriptive and contained no relevant information as indicated by a vanishing statistical distance, cf. Fig. 7.23. This sudden gain in importance is not necessarily a result of only the variability in t_1, but of the generally increased number of free parameters, since different features are likely to influence one another. In principle, these correlations are not accessible through marginal distributions, however, we can apparently observe some of their consequences. Other important features are the phases of the next-nearest neighbor hoppings. All remaining features are separated from the aforementioned by a jump in the importance score by one order of magnitude.

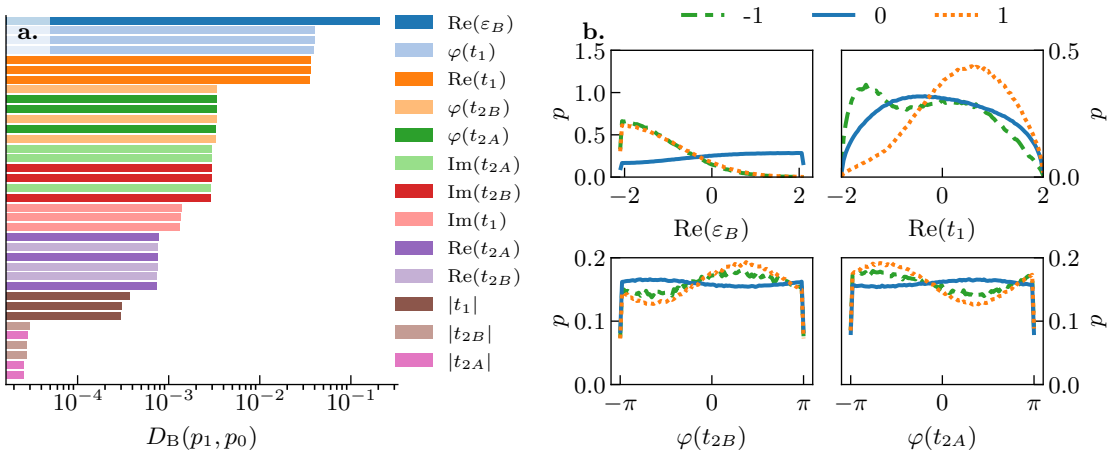

Figure 7.25: **a.** Importance score (Bhattacharyya distance) for all features. Features are grouped by equivalence and shown in the same color. The mass (onsite term) has the highest score followed by the phase and real part of the nearest neighbor terms and the phases of the next-nearest neighbor terms. Real parts are less important than imaginary parts and $|t_i|$ contain very little information. **b.** Selected estimated distribution functions for representatives of the most important features. There is indeed a clear signal for all important features. The mass term ($m = -\varepsilon_B$) is reduced for the topological phases as expected and the next-nearest neighbor hoppings show a tilt towards the characteristic Haldane values. Here, also the t_1 feature is very important, in contrast to the simple four-parameter model, and reveals a contrast between all three phases. [Figure adapted from Ref. [173] based on different data]

We show in Fig. 7.25**b** the estimated marginal distribution functions for a selection of important features. As expected, there is a recognizable amount of information in these distributions. In particular, the local term ε_B shows a significant bias towards negative values for the topological phases, in contrast to the trivial phase. This is in agreement with our knowledge of the phase diagram, since $\varepsilon_B = -m$, and therefore, the mass term is reduced w.r.t. the reference value. The distribution for the nearest-neighbor terms is most interesting, as it shows a contrast not only between the $y = 1$ and $y = 0$ phases, but also between $y = 1$ and $y = -1$. While the $y = 1$ phase

[6] "Equivalent features" here means not only features that connect the same sites and are therefore related by some symmetry operation. In order for the marginal distributions to be equal, the corresponding components of the reference point must be the same. This is expected not to be the case for reference points that are obtained from realistic (imperfect) systems.

leans towards positive t_1, the $y = -1$ phase has a pronounced peak at negative effective t_1, i.e., even after shifting by $\mathbf{x}_{\text{ref}}$. This is interesting, since we know that the value of the topological invariant does not depend on the sign of t_1 in the ordinary Haldane model. Therefore, this peak must carry the signature of another model, where the correlations between the different features and, in particular, between equivalent features play an important role. Incidentally, we have already seen this phase in Chapter 6, where we found that a negative hopping amplitude that links A and B sublattices within the unit cell leads to a non-trivial phase with negative Chern number, cf. Eq. 6.7. Astonishingly, we can extract this information from a large data set by using information theoretic tools only. We will explore this phase further below.

For the next-nearest neighbor terms we find the characteristic pattern of the phase factors, where $t_{2B(A)}$ is more likely to have positive (negative) values.[7] In contrast to the Haldane phase diagram, where the distribution would vanish in the respective opposite case, we here have positive values throughout all phases. This has two reasons. First, we have to keep in mind that the features are to be understood as displacements relative to $\mathbf{x}_{\text{ref}}$. Given that

$$\varphi(x + x_{\text{ref}}) = -i \log\left(\frac{x + x_{\text{ref}}}{|x + x_{\text{ref}}|}\right), \tag{7.96}$$

the value of the total phase depends strongly on x_{ref}. In particular, for $|x| \ll |x_{\text{ref}}|$, the phase exhibits only a small shift, which indicates that many samples, regardless of $\varphi(x)$, probably have a phase similar to $\varphi(x_{\text{ref}})$. Nevertheless, we can observe a tendency towards a particular direction in $\mathbb{C}$ when starting from x_{ref} in the data.

In order to clarify this point a bit more, we now compare the marginal distributions for our features with those of the actual hopping parameters, i.e., $\mathbf{t}_i = \mathbf{x}_i + \mathbf{x}_{\text{ref}}$. A plot of the resulting estimator for the same features as before is shown in Fig. 7.26. Apparently, the distributions of the real parts of ε_B and t_1 do not change at all (apart from a shift). This is a consequence of the linearity of the transformation

$$\begin{aligned}(\mathrm{Re}\left[\mathbf{x}_i\right], \mathrm{Im}\left[\mathbf{x}_i\right]) \;&\mapsto (\mathrm{Re}\left[\mathbf{x}_i + \mathbf{x}_{\text{ref}}\right], \mathrm{Im}\left[\mathbf{x}_i + \mathbf{x}_{\text{ref}}\right]) \\ &= (\mathrm{Re}\left[\mathbf{x}_i\right] + \mathrm{Re}\left[\mathbf{x}_{\text{ref}}\right], \mathrm{Im}\left[\mathbf{x}_i\right] + \mathrm{Im}\left[\mathbf{x}_{\text{ref}}\right]),\end{aligned} \tag{7.97}$$

which implies that the energy scales of the features are simply shifted. This leaves the shape of the distribution invariant, and therefore, also the importance score does not change. For the phases this is entirely different, as is evident from the non-linear relationship in Eq. 7.96. Consequently, unless samples of features are very large, they do not change the overall phase significantly, which leads to the distributions shown in Fig. 7.26, which for all three labels are very similar to each other.

For $\varphi(t_{2A/B})$, we observe a peak at $\mp\pi/2$, which corresponds to the reference value, and a minimum at the respective opposite. The three lines are almost indistinguishable from one another around the minimum, since values like this would correspond to a perturbation in the direction of the trivial honeycomb lattice with real phase factors. This configuration is thus very unlikely for a topological phase. Around the peak, however, we can distinguish the three phases, since this configuration is highly favorable for a topological phase, and therefore, more relevant in the corresponding conditioned data set. Notably, both topological phases seem to have the same sign structure of the phase factors, which indicates the same flux. The inversion of the topological index must therefore be generated by something other than t_2. We attribute the fact that the distribution for the $y = -1$ phase is a bit shallower around the peak and marginally

[7] The sign depends on the choice of directions. Generally, links that have a clockwise orientation have positive sign, and negative for counter-clockwise links. We plot here t^1_{2A}, t^1_{2B}, which are oriented counter-clockwise and clockwise, respectively.

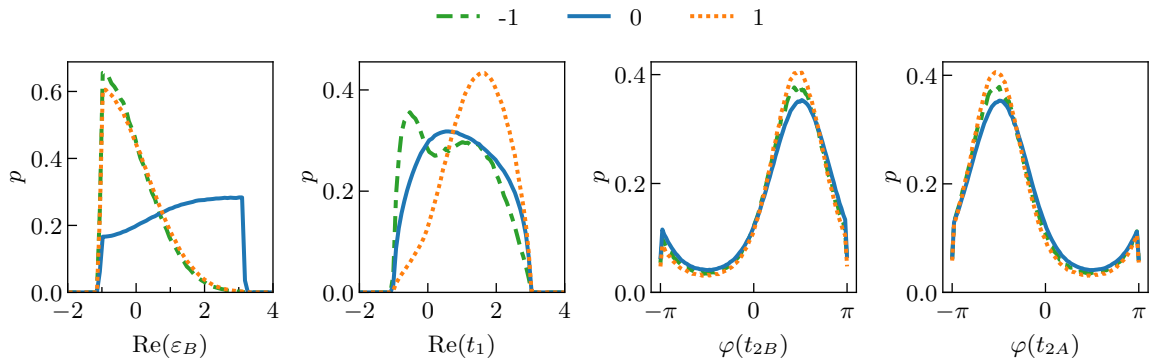

Figure 7.26: Probability density estimator $p(x_i|y=l)$ for $l=-1,0,1$, and the most important features $\mathrm{Re}\left[\varepsilon_B\right], \mathrm{Re}\left[t_1\right]$ and the complex phases of the next-nearest neighbor hoppings t_{2A}, t_{2B}. Here, we explicitly show the hopping parameters, not the displacement from the reference point. Real and imaginary parts do not change at all. The phases do change and we now clearly observe that the Haldane configuration $[\varphi(t_{2B}) > 0 > \varphi(t_{2A})]$ is clearly favored by all phases. The signal in the trivial phase is a remnant of the choice of $\mathbf{x}_{\mathrm{ref}}$.

larger at the minimum compared to that for $y=1$ to the presence of a small number of samples with reversed flux. The probability for such samples is given by[8]

$$P_{\text{reversed flux}} \sim \prod_{x_i \in \{t_{2A}^i\}} P(\varphi(x_i) > 0, |x_i| > |x_{\mathrm{ref}}^i|) \prod_{x_i \in \{t_{2B}^i\}} P(\varphi(x_i) < 0, |x_i| > |x_{\mathrm{ref}}^i|), \tag{7.98}$$

where $P(\varphi(x_i), |x_i|) = P(\varphi(x_i))P(|x_i|)$ due to the statistically independent sampling. With $P(\varphi(x_i) \lessgtr 0) = \frac{1}{2}$ and $P(|x_i| > |x_{\mathrm{ref}}^i|) = \frac{3}{4}$ for the uniform distribution of Eq. 7.24, we have

$$P_{\text{reversed flux}} \sim \left(\frac{1}{2}\right)^6 \left(\frac{3}{4}\right)^6 = \frac{3^6}{2^{18}} \approx 0.27\%, \tag{7.99}$$

which indicates that the reversed flux pattern that is characteristic for the $y=-1$ phase in the original Haldane model is extremely unlikely to occur in our data set in significant numbers. Overall, this implies that the PDFs of features that transform non-linearly under the perturbation $\mathbf{x}$, see Eq. 7.97, are immensely biased through the choice of the reference point, which hides the relevant information in the marginal distributions of the hopping parameters. The distributions of the features derived from the displacements w.r.t. the reference point on the other hand remove this bias, and therefore, provide a much clearer picture of the properties of individual classes.

We note that the distributions for the trivial phase have to be taken with a grain of salt, since due to the abundance of trivial samples they are to be understood as the weighted difference between the total distribution and those for the topological labels, and are therefore heavily biased by the design of the data generation procedure.

Finally, we discuss the second reason for the finite probability density of the phase terms in the "classically forbidden" regime (i.e., forbidden in the Haldane model) in Fig. 7.25**b**, where the effective value of ϕ is opposite to that in the phase diagram. This signal is still present in Fig. 7.26 and can therefore not be explained by the non-linearity of the corresponding features under the transformation $\mathbf{x} \mapsto \mathbf{x} + \mathbf{x}_{\mathrm{ref}}$. Based on our knowledge of the phase diagram, such configurations should be characteristic of the $y=-1$ phase, however, we find them with very similar weight

[8] Since the distributions for oppositely oriented links are simply mirrored we assume here that they are aligned with the distributions shown in Fig. 7.26 to simplify the notation.

regardless of the phase. The only explanation that remains is provided by correlations between different features that are integrated out in the marginal distributions. Clearly, if not all phase terms switch sign, but maybe just one, the total flux will still retain the same sign and so will the topological index. In order to find out more about the relations between class labels and features, we will therefore have to go beyond marginal distributions.

7.7.4 Correlations

We have focused so far only on marginal distributions of individual features that proved very useful in determining the information content of particular features. This establishes phase 2 (dimensional reduction) in Fig. 7.2. For phase 3 (model building) we need to extract more in-depth information that involves knowledge of the correlations between individual features.

Our description of relations between different features is based on two ingredients: statistical dependence and correlation. As introduced in Sec. 4.1.2, the joint PDFs of independent random variables can be factorized into a product of the marginal distributions. Applied to our case this means that $p(x_i, x_j) = p(x_i)p(x_j)$. Whenever this is not the case, the condition that one feature has a certain value leads to different distributions for the other feature. The features therefore depend on each other. It is quite instructive to use the definition of the conditional probability to write the condition for statistical independence as

$$p(x_i, x_j) = \left\{ \begin{array}{c} p(x_i|x_j)p(x_j) \\ p(x_i)p(x_j|x_i) \end{array} \right\} = p(x_i)p(x_j), \tag{7.100}$$

from which it follows immediately that $p(x_j|x_i) = p(x_j)$ and $p(x_i|x_j) = p(x_i)$, i.e., there is no information about x_i contained in x_j and vice versa. Statistical independence is a very strong statement and can certainly never be proven exactly in our finite data set. However, the degree of statistical (in)dependence can be quantified in terms of the mutual information, see Eq. 4.71,

$$I(x_i, x_j) = \int_{-\infty}^{\infty} p(x_i, x_j) \log \left[\frac{p(x_i, x_j)}{p(x_i)p(x_j)} \right] \mathrm{d}x_i \mathrm{d}x_j, \tag{7.101}$$

which vanishes for independent features and assumes a maximum value of

$$H(x_i, x_j) = -\int_{-\infty}^{\infty} p(x_i, x_j) \log \left[p(x_i, x_j) \right] \mathrm{d}x_i \mathrm{d}x_j \tag{7.102}$$

if one feature completely determines the other. It is therefore useful to define a measure of the information that one feature contains about the other that we call *redundancy*[9]

$$R(x_i, x_j) = \frac{I(x_i, x_j)}{H(x_i, x_j)}. \tag{7.103}$$

Apparently, given that $0 \leq I(x_i, x_j) \leq H(x_i, x_j)$ we have $0 \leq R(x_i, x_j) \leq 1$, i.e., the redundancy is normalized.

In addition to the redundancy, we use another measure of statistical correlations, which are usually defined in terms of a correlation function that is given by the covariance

$$\mathrm{Cov}[x_i, x_j] = \mathrm{E}[(x_i - \mathrm{E}[x_i])(x_j - \mathrm{E}[x_j])]. \tag{7.104}$$

[9]The term *redundancy* is usually used in information theory in the context of data compression, where it refers to the amount of surplus information that different random variables contain about one another. In other words, this information is unnecessary, and therefore, redundant. See also Ref. [132].

The covariance is bounded from above by $\sqrt{\text{Var}[x_i]\text{Var}[x_j]}$, since

$$\text{Var}[x_i + x_j] \geq 0, \tag{7.105}$$

and therefore, assuming that $\text{Var}[\tilde{x}_i] = \text{Var}[\tilde{x}_j] = 1$

$$0 \leq \text{Var}[\tilde{x}_i \pm \tilde{x}_j] = \text{Var}[\tilde{x}_i] \pm 2\text{Cov}[\tilde{x}_i, \tilde{x}_j] + \text{Var}[\tilde{x}_j] = 2 \pm 2\text{Cov}[\tilde{x}_i, \tilde{x}_j]. \tag{7.106}$$

As a result, we have $-1 \leq \text{Cov}[\tilde{x}_i, \tilde{x}_j] \leq 1$, and by defining $\tilde{x}_i = x_i/\sqrt{\text{Var}[x_i]}$ we have $-1 \leq \text{Cov}[x_i, x_j]/\sqrt{\text{Var}[x_i]\text{Var}[x_j]} \leq 1$. This leads us to the definition of the Pearson correlation coefficient [271]

$$r(x_i, x_j) = \frac{\text{Cov}[x_i, x_j]}{\sqrt{\text{Var}[x_i]\text{Var}[x_j]}}. \tag{7.107}$$

We now briefly explain what the two quantities R and r describe in terms of the distributions of features, i.e., which information we can extract through them. In Fig. 7.27, we illustrate typical values of both the redundancy and correlation coefficient with a number of different probability distributions.

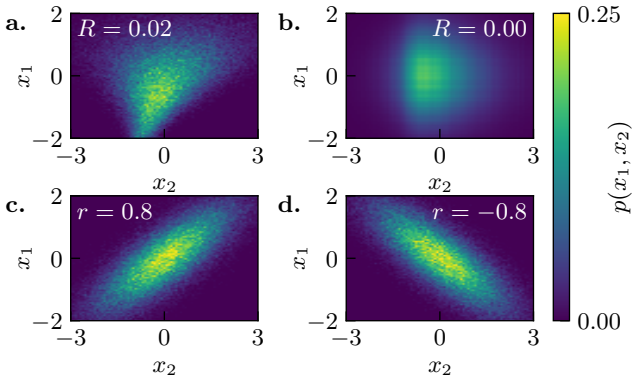

Figure 7.27: Illustration of the redundancy R and correlation coefficient r for different joint PDFs. In **a.**, we show a joint PDF for two dependent random variables, which have a finite redundancy. The product of marginals of this distribution is shown in **b.** as an example of two independent random variables. In **c.** and **d.**, we show two distributions with positive and negative correlation, respectively. [Figure adapted from Ref. [173]]

The distribution in Fig. 7.27**a** is given by

$$p_1(x_1, x_2) = N[0, 1]\,(x_1)\; N\left[\frac{3}{8}x_1, 2(1 - e^{-x_1})^{-1}\right](x_2), \tag{7.108}$$

where $N[\mu, \sigma]$ is the normal distribution with mean μ and standard deviation σ. Clearly this cannot be written as a product of two marginal distributions for x_1, x_2, respectively, and therefore, x_1, x_2 are not independent, which is indicated by the positive redundancy value $R = 0.02$ that is also shown in the figure. The distribution in subfigure **b** on the other hand is simply given by the product of marginal distributions obtained from the joint PDF in Eq. 7.108, which differs visibly from subfigure **a**, and the corresponding redundancy vanishes as expected. We note here that the values we can expect for the redundancy are rather small and while $R \in [0, 1]$ we will typically obtain values $\ll 1$, as $R = 1$ corresponds to the extreme case where knowledge of x_1 would completely determine x_2, which is not the case in this example.

The lower row of plots in Fig. 7.27 demonstrates the information contained in the correlation coefficient of Eq. 7.107. Apparently, we are looking at a distribution that extends along a

line with positive (negative) slope in the x_1-x_2 plane in **c**,**d**, respectively. The correlation coefficient reflects the sign of the slope, while the absolute value represents the degree to which the distribution is blurred out. While this is certainly useful information, only linear correlations are detected, which means that one either looks at a small enough region such that a linearization is valid or the result will only be useful accidentally, i.e., an accidental asymmetry leads to nonzero r. Despite this shortcoming, the correlation coefficient is an important tool in detecting whether a given feature is consistently larger than another.

Of course, we generally have more than two features, and therefore, the joint PDFs are still marginalized over a number of other features, which can lead to a loss of information if correlations are intrinsically based on a relation between larger subsets of features. For such a description, higher order correlation functions are required. Fortunately, the definition of the mutual information can be extended to [272]

$$I(x_1, x_2, \ldots, x_n) = \sum_i H(x_i) - H(x_1, x_2, \ldots, x_n), \tag{7.109}$$

which describes the total redundancy in a set $\{x_1, \ldots, x_n\}$ of n variables. However, $I(x_1, x_2, \ldots)$ is no longer strictly positive, and therefore, an interpretation of this quantity is not straightforward. For the correlation coefficient this generalization is also possible, however, not necessarily useful, since the sign of

$$r(x_1, x_2, \ldots, x_n) = \frac{\mathrm{Cov}[x_1, x_2, \ldots, x_n]}{\sqrt{\prod_i \mathrm{Var}[x_i]}} \tag{7.110}$$

no longer has a well-defined meaning. For instance, assuming we have three random variables, then

$$r(x_1, x_2, x_3) = \frac{\mathrm{Cov}[x_1, x_2, x_3]}{\sqrt{\mathrm{Var}[x_1]\mathrm{Var}[x_2]\mathrm{Var}[x_3]}} = \frac{\mathrm{E}[(x_1 - E[x_1])(x_2 - E[x_2])(x_3 - E[x_3])]}{\sqrt{\mathrm{Var}[x_1]\mathrm{Var}[x_2]\mathrm{Var}[x_3]}} \tag{7.111}$$

can be positive if negative signs are distributed pairwise between the three parentheses, which does not reveal a lot about the correlations. In general, multivariate correlations are considered very difficult to describe [272–275], and therefore, we do not go into any more details here and leave this issue open for future work.

Application to the Data

We now try to extract more information from the data set through an in-depth investigation of the correlations between individual features. This is done by computing our correlation measures for the features that we already found to contain most information. We lay particular emphasis on those features that represent equivalent hopping paths, i.e., the triplets (t_1^1, t_1^2, t_1^3), $(t_{2A}^1, t_{2A}^2, t_{2A}^3)$ and $(t_{2B}^1, t_{2B}^2, t_{2B}^3)$, since here we kept our data set artificially general which does not necessarily reflect the reality in many materials and leads to overly complicated models. In general, we can distinguish between these correlations between equivalent features and those between inequivalent features. The latter are best investigated in simplified models, where the other type of correlation is fixed by construction. This will be discussed later.

In Fig. 7.28, we show a selection of measures of correlations for both topologically non-trivial phases found in our data set. Due to our assumption that the trivial phase represents the generic case, which is also supported by the large abundance of trivial samples, we do not perform such an analysis for the trivial phase. The top row of panels is devoted to the class $y = 1$, the bottom row corresponds to $y = -1$. We start the discussion with the $y = 1$ phase. In Fig. 7.28**a**, we plot the values of the redundancy $R(\phi_i, \phi_j)$, cf. Eq. 7.103, for $i, j \in \{0, \ldots, 10\}$ on a grid. Naturally,

the redundancy is symmetric in i, j so that the upper half triangle is identical to the lower half. We set the diagonal to zero for better visibility of the interesting data. The diagonal redundancy is simply equal to 1 for any feature, since $p(x_i = x, x_i = y) = p(x_i = x)p(x_i = y|x_i = x) = p(x_i = x)\delta(x - y)$, and therefore

$$R(x_i, x_i) = \frac{\int_{-\infty}^{\infty} p(x_i = x, x_i = y) \log\left[\frac{p(x_i=x, x_i=y)}{p(x_i=x)p(x_i=y)}\right] \mathrm{d}x\mathrm{d}y}{\int_{-\infty}^{\infty} p(x_i = x, x_i = y) \log\left[p(x_i = x, x_i = y)\right] \mathrm{d}x\mathrm{d}y} \tag{7.112}$$

$$= \frac{\int_{-\infty}^{\infty} p(x_i = x) \log\left[\frac{p(x_i=x)}{p(x_i=x)^2}\right] \mathrm{d}x}{\int_{-\infty}^{\infty} p(x_i = x) \log\left[p(x_i = x)\right] \mathrm{d}x} \tag{7.113}$$

$$= \frac{H(x_i)}{H(x_i)} = 1. \tag{7.114}$$

Thus, the diagonal value is of no interest anyway. In the computed values we find a clear signal for the features $\varphi_2, \varphi_3, \varphi_4$ ($\varphi(t_1^1), \varphi(t_1^2), \varphi(t_1^3)$) between one another. The symmetry between all three distinct pairs is again a consequence of the fact that the joint distribution, from which the correlations are inferred, is also a marginal distribution, marginalized over the third variable among others, and therefore, $p(x_i, x_j) = \int p(x_i, x_j, x_k)\mathrm{d}x_k$ with $p(x_i, x_j, x_k)$ symmetric in i, j, k. This symmetry is a consequence of the degree of freedom of choosing the lattice basis, i.e., the coordinate system in which we represent the lattice that cannot change the physical properties.

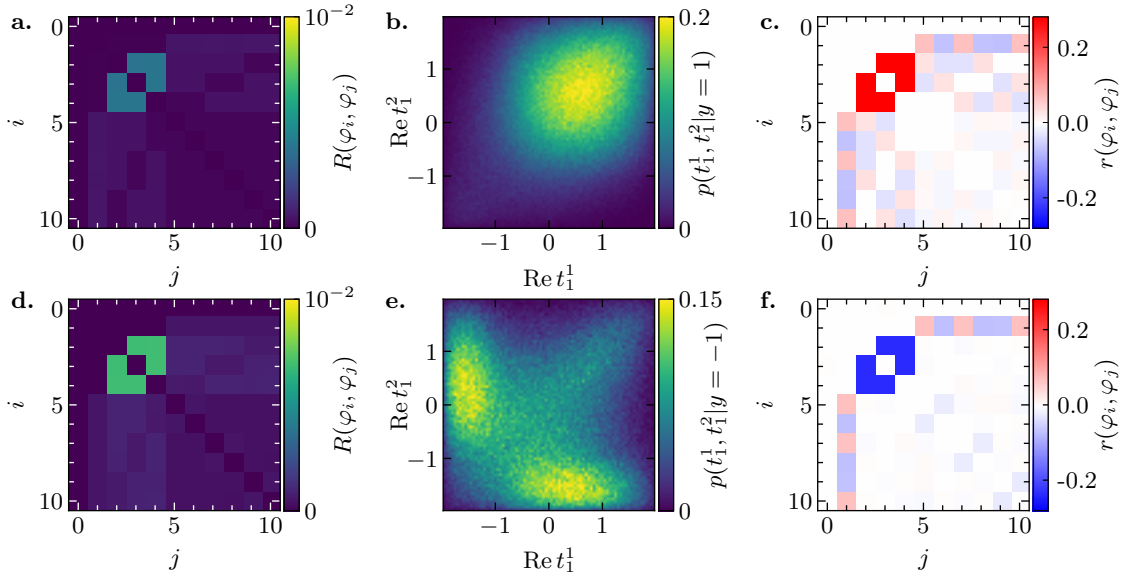

Figure 7.28: Correlation measurements. Top row: class $y = 1$, bottom row: class $y = -1$. Panels **a.**,**d.** show the redundancy R [Eq. 7.103] between phases ϕ_i, ϕ_j, where clear positive values are detected for features x_2, x_3, x_4, which correspond to t_1^1, t_1^2, t_1^3. **b.**,**e.** Joint PDF for the real parts of t_1^1, t_1^2. The correlations are clearly of a different kind for the two phases. While for $y = 1$ both features tend to have similar positive values, the opposite is the case for $y = -1$. **c.**,**f.** Correlation coefficient r [Eq. 7.107]. The sign of r clearly distinguishes the two phases.

We note that we previously decided to give the real part of t_1^i precedence over the phase, since both were found to have similar importance, cf. Fig. 7.25. Here, we use the phase so that we cover also the important phases of the next-nearest hopping terms, which show no notable

redundancy. We confirmed that the real part of the nearest neighbor hoppings shows a similar redundancy.

In Fig. 7.28**b**, we plot the joint distribution function of the real parts of two nearest neighbor hoppings, which sheds more light on the type of correlations present in our data set. We observe a clear concentration of weight at finite values of both parameters, while, in contrast to that, negative values are strongly suppressed. The fact that the probabilistic weight is localized in the upper right quadrant indicates positive correlations, which is also confirmed by the correlation coefficient shown in Fig. 7.28**c**, which assumes a clear positive value for pairs of nearest neighbor hoppings. Most other correlations we find are rather weak and can mostly be considered noise. Interesting is certainly the correlation between the sign of the local potential m (note that this is real, and therefore, the phase is binary) and the next-nearest neighbor phases, which indicates that an increase in m (negative values of $x_1 = \varepsilon_B$, and therefore $\varphi_1 = \pi$) is likely to be accompanied by smaller values of $\varphi(t_{2A})$ and larger values of $\varphi(t_{2B})$, which is consistent with the phase diagram that we know, where larger m can only be compensated by phases leaning towards $\mp\pi/2$.

We now investigate also the $y = -1$ phase, where the redundancy shows a similar signature, i.e., strong correlations are only present among the nearest neighbor hoppings, cf. Fig. 7.28**d**. The corresponding joint PDF of the real part is shown in Fig. 7.28**e**, where we observe a clearly different picture than for the $y = 1$ phase. Apparently, configurations that have one positive and one negative value are far more common than others. The finite weight along the diagonal seems to contradict this, however, we have to take into account the third t_1 feature that is integrated out. Here, we could imagine an application of a three-variable correlation function such as Eq. 7.111, however, the positive or negative sign of the outcome would not differentiate between different types of negative correlations, i.e., $(+,+,-)$ and $(-,-,+)$ or, in other words, what the majority signs of $x_i - \mathrm{E}[x_i]$ are. We therefore define another characteristic that provides us with this information:

$$r_+(x_1, x_2) = r(x_1, x_2 | x_3 > \mathrm{E}[x_3]), \qquad r_-(x_1, x_2) = r(x_1, x_2 | x_3 < \mathrm{E}[x_3]). \tag{7.115}$$

In the case that $(+,+,-)$ is more likely we will obtain $(r_+, r_-) = (-,+)$ and $(+,-)$ for $(-,-,+)$. The positively correlated cases $(+,+,+), (-,-,-)$ both yield $(+,+)$. Computed for both topological phases we find $(+,+)$ for the $y = 1$ and $(-,+)$ for $y = -1$, which indicates that the three features are positively correlated in the former and negatively correlated in the latter case. More precisely, this result proves what was not entirely obvious from the joint distribution of two features alone: the generic configuration of the $y = -1$ class is such that one of the three hopping parameters is reduced greatly such that the effective value is negative, while the other two are not reduced as much and stay positive.

7.7.5 Optimized Model – Information Leads to Improvement

We can now use the information we gathered from the statistics of the data set to reduce the dimensionality of the model once more. This is now straight-forwardly done by introducing symmetries between parameters based on their correlations between one another. For the nearest-neighbor hoppings, the positive correlations ($y = 1$) motivate us to make them symmetric, i.e., use only one random variable to describe all three hoppings: $t_1^i \to t_1$. The particular kind of negative correlations found for $y = -1$ on the other hand motivates us to define $t_1^1, t_1^2 \to t_1$ and $t_1^3 \to -1$, where we explicitly break the corresponding point group symmetry that one would expect for a perfect honeycomb lattice. Moving on to the next-nearest neighbor hoppings, we found no notable correlations among any of them and, in particular, the redundancy R assumes negligible values. Therefore, we can assume that these features are almost independent, which

gives us free choice. In order to retain a sufficiently general model, we keep two independent parameters for the two sublattices, i.e., $t^i_{2A} \to t_{2A}$ and $t^i_{2B} \to t_{2B}$, where the symmeries are chosen such that all marginals are the same. This implies that counter-clockwise hoppings are complex conjugates of clockwise hoppings. As representatives we chose the same $t^1_{2A/B}$ as before.

It is important to note here that these symmetries are not required to obtain a topological phase. However, due to the robustness of topological phases w.r.t. small changes to the Hamiltonian, there is a large feature space of complicated configurations that is extremely difficult to parameterize, and therefore, does not advance our understanding of the general characteristics of the individual phases. Therefore, we seek here to reduce the complexity of the effective description of each phase by making educated assumptions about certain symmetries that result in a minimal model capturing the characteristic configurations describing each particular phase. In a way, this corresponds to transitioning from the most general description in terms of a large number of hopping matrix elements $t_{ij}(\mathbf{R})$ to a small set of model parameters that each describe a whole class of hopping amplitudes. Of course, there need not be only one effective model for any phase and indeed, we will find more than one in the following.

We now have a new reduced feature space tailored to each topological class individually that allows us to study their characteristics with greater precision by removing some of the arbitrariness from the description. In particular, the symmetries introduced just now eliminate the superposition of configurations contributed by different permutations of equivalent features that effectively reduce the contrast in our signal. The new set of features is defined by

$$\mathbf{x} = (-\varepsilon_B, \varepsilon_B, t_1, t_{2A}, t_{2B}), \tag{7.116}$$

where the amplitude t_1 of the nearest neighbor hopping terms contributes with sign structures $(+,+,+)$ for $y = 1$ and $(+,+,-)$ for $y = -1$. Despite the previously reported importance scores, cf. Fig. 7.25, we do not restrict ourselves to real t_1 and only a phase term for $t_{2A/B}$, and instead keep all remaining parameters complex, so as not to remove too much complexity at once. Therefore, we are left with one real and three complex parameters, which results in a total of seven degrees of freedom. The initial $-\varepsilon_B$ in Eq. 7.116 is symbolic for the fixed zero of the energy scale and the corresponding value will not be sampled.

We now create another data set by sampling from the distribution of Eq. 7.24 for the new features defined in Eq. 7.116. Since during the sampling procedure the symmetries between individual hopping terms have to be respected, we need to create separate data sets for the two classes. In the following we do this for $y = 1$, i.e., positive correlations between nearest neighbor hoppings. Again, we choose a spread parameter of $\alpha = 2$, cf. Eq. 7.24, but reduce the number of samples to $n_{\text{samples}} = 10^6$, taking into account the compressed feature space. For the reference point we use again the same as before, cf. Eq. 7.93, that translates to

$$\mathbf{x}_{\text{ref}} = (-1.05, 1.05, 1, 0.2e^{-i\frac{\pi}{2}}, 0.2e^{i\frac{\pi}{2}}). \tag{7.117}$$

We illustrate the most basic information about the data set in Fig. 7.29, where subfigure **a** represents the improved model, that now considers all equivalent hopping paths with the same amplitudes, which we indicate through equal coloring, cf. Fig. 7.24 for the unsymmetric most general case. The links that have hoppings equal to $t_{2B(A)}$ are oriented in clockwise (counter-clockwise) direction. In Fig. 7.29**b**, we show the composition of the data set in terms of a doughnut chart. Instead of a majority of trivial samples, we now have roughly the same number of $y = 0$ and $y = 1$ samples, while the $y = -1$ phase is barely present at all. This is understandable, since the model has been tailored to the $y = 1$ phase using the information we extracted from the general model. A large number of $y = -1$ samples would therefore contradict the distinct correlations we found earlier. While also in the general model most samples were insulating, this is even more so now with the vast majority being insulating samples. We note that

we assign a class label to every sample, insulating or not. This is possible since gap closings are not protected by any symmetries, and therefore, require careful tuning of parameters, which is a rather rare case. Most metallic samples, in fact, simply lack an overall band gap but still feature two separated bands that never touch, which means that the Berry curvature and topological index for the lower band are well-defined.

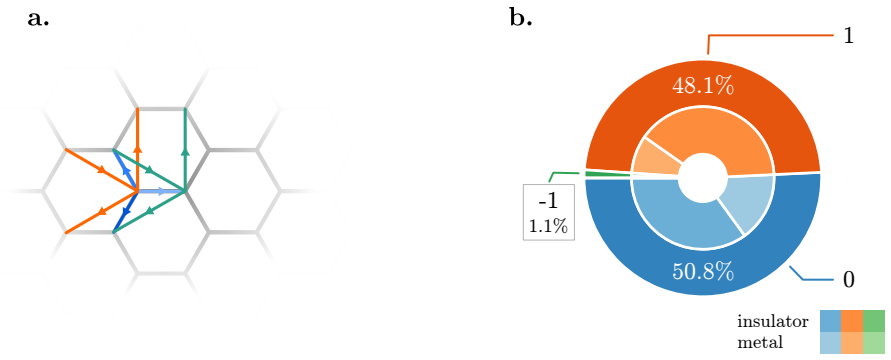

Figure 7.29: Data set generation for the compressed $y = 1$ model. **a.** Hopping paths taken into account. Same color indicates same random variable. The representative for t_{2A} (teal) has counter-clockwise direction, while t_{2B} (orange) is clockwise. Due to the symmetries imposed all equivalent links are aligned. **b.** Composition of the data set. We find roughly 50% trivial and 50% non-trivial samples ($y = 1$). The number of samples with $y = -1$ is negligible. Throughout all phases, the grand majority of samples is insulating. [Figure adapted from Ref. [173] based on different data]

For this new data set we again estimate the marginal probability density functions for all features and class labels via Eq. 7.84 and compute the importance score in terms of the Bhattacharyya distance [Eq. 7.62], while we confirmed that the score does not qualitatively depend on which qualified distance function from Table 7.3 is used. We show the results in Fig. 7.30, where subfigure 7.30**a** is again a bar chart of the importance scores for all features. In direct comparison to Fig. 7.25, we see the reduced complexity of the compressed model very clearly, since now all parameters appear only once. The local term is most important, followed closely by the imaginary part and phase of the next-nearest neighbor terms. The only other parameters with non-negligible importance scores are the real parts of $t_{2A/B}$. Clearly, the phase, real and imaginary parts are redundant and so we choose the phase as the most descriptive variant. All features relating to nearest neighbor hoppings are astoundingly unimportant now that we fixed the correlations among them.

In Fig. 7.30**b**, we show a selection of probability density functions. It is quite clear that the local energy ε_B and the phases of $t_{2A/B}$ show a good contrast. The marginal distributions of the real part of t_1, on the other hand, are all the same regardless of the topological class, which is the reason for the low importance score. This indicates that the value of t_1 contains no information about the topological phase at all and is therefore completely arbitrary. This is interesting, as it is in stark contrast to the previous more general model, where the same feature was one of the most descriptive. Clearly, this is a consequence of the enforced symmetry, i.e., by constructing this more refined model with symmetric t_1^i amplitudes we already infused this information into the model so that the resulting data is less informative and, in fact, contains no additional information (about the t_1 parameters) at all. The case is entirely different for the next-nearest neighbor terms, which show a greatly increased contrast after symmetrization compared to the most general model. The reason for this lies hidden in correlations that we were not able to observe in the general model. We will investigate this in more detail below. Again, we find that the $y = 1$ phase favors values of $\varphi(t_{2A/B}) \approx \mp\pi/2$, now with better contrast.

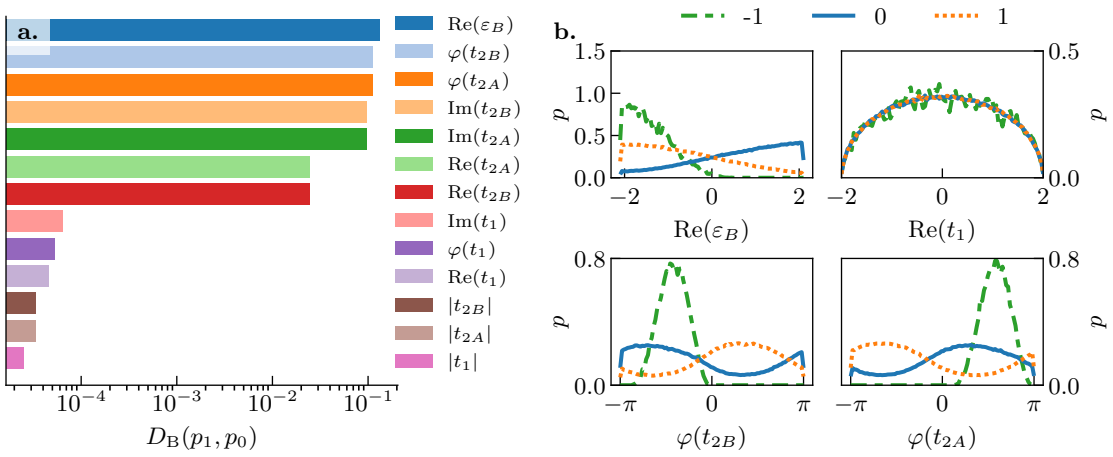

Figure 7.30: Evaluation of the improved model for the $y = 1$ class. **a.** Importance score in terms of the Bhattacharyya distance (for $y = 1$). The local term, as well as the imaginary part or phase of the next-nearest neighbor hoppings are most important. Nearest neighbor hoppings are altogether unimportant. **b.** Marginal PDFs for a selection of features. ε_B as well as $\varphi(t_{2A/B})$ show a clear discrimination between topological and trivial phases. For the unimportant parameter $\mathrm{Re}(t_1)$ we see that all marginal distributions are the same. [Subfigure **b.** adapted from Ref. [173] based on different data]

However, in stark contrast to the general model, we observe a different characterization of the $y = -1$ phase that was hidden before. The symmetrization chosen according to the positive correlations between t_1^i in order to better describe the $y = 1$ phase apparently uncovered a different realization of the $y = -1$ phase that is achieved not through negative correlations in the nearest neighbor hoppings but through a phase inversion in the next-nearest neighbor hoppings, which corresponds to a reversed flux and is, of course, known to us from the Haldane phase diagram. Hence, this corresponds to the generic $y = -1$ phase for the Haldane model. Clearly, our reference point was chosen with a bias towards the $y = 1$ phase so that it is not surprising that the description of the usual $y = -1$ phase remained hidden initially.

Having discussed the information directly accessible through the marginal PDFs of individual features, we now turn our attention to correlations again. Due to the way the symmetries between features were implemented we have a data set filled with redundant copies of the few actual features with the numbers corresponding to the original multiplicities of each hopping term. This is rather convenient, though, since it allows for a direct one-to-one comparison between the general and the symmetrized model.

We compute the redundancy [Eq. 7.103] and correlation coefficient [Eq. 7.107] for the new data and present the results in Fig. 7.31. Subfigure 7.31**a** shows a comparison between the redundancy in different phase features for the general and the symmetric model. With our symmetry assumption, clearly all t_1^i and t_{2A}^i, t_{2B}^i are perfectly correlated among themselves, i.e., $R = 1$, which is an immediate consequence of Eq. 7.114. In order to resolve more subtle correlations, we cut the scale at a suitable value. It turns out that a finite redundancy exists now between t_{2A} and t_{2B} as well as between all $t_{2A/B}$ and ε_B. This data has been computed for the abundant $y = 1$ phase. In subfigure 7.31**b**, we show a comparison between the joint PDFs of $t_{2A/B}$, again for the general and symmetric models. The PDF of the general model is extremely flat and featureless, which is a consequence of the high degree of marginalization, i.e., averaging

over equivalent features:

$$p(t_{2A}^1, t_{2B}^1) = \int_{-\infty}^{\infty} p(\{t_{2A}^i, t_{2B}^i | i = 1, 2, 3\}) \prod_{\substack{s\in\{A,B\} \\ i\in\{2,3\}}} \mathrm{d}t_{2s}^i. \tag{7.118}$$

This does not happen in the symmetric model, where we instead find a very clear structure that is barely visible also in the PDF for the general model. What we find is that configurations with negative $\varphi(t_{2A})$ and positive $\varphi(t_{2B})$ are most likely, those with both positive or both negative less so and the opposite, i.e., $\varphi(t_{2A}) > 0$ and $\varphi(t_{2B}) < 0$, is expressly unlikely. This observation agrees well with our understanding of the Haldane $y = 1$ phase for which the most likely configuration is characteristic. We note that the joint distribution, despite displaying a rather clear structure and relation between the two parameters, is not particularly highly correlated. In fact, the product of the two marginal distributions has a very similar structure, meaning that in this case we can deduce the most relevant information already from the marginal distributions.

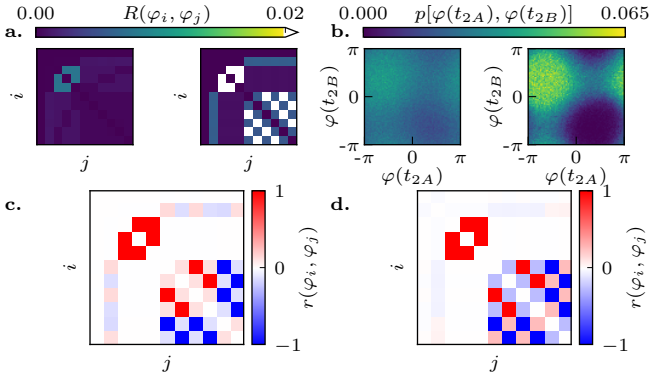

Figure 7.31: Correlation measurements for the symmetric $y = 1$ model. **a.** Redundancy R as a function of the phase features for the general model (left) and the symmetric model (right). Symmetric features are perfectly dependent, while next-nearest neighbors are correlated among each other and with the local potential. **b.** Joint PDF for the two distinct next-nearest neighbor phases for the general model (left) and the symmetric model (right). The underlying structure observed is the same, albeit the contrast being greatly improved by the symmetry. **c.** Correlation coefficient r for the phase features for the $y = 1$ phase. We observe positive correlations between $\varphi(t_{2A})$ and $\varphi(t_{2B})$ and correlations of opposite sign between the sign of ε_B and the two phases. The perfect correlations among $\{t_{2A}^i\}, \{t_{2B}^i\}$ are a consequence of the symmetry and the alternating sign is related to the original directions of hopping links that we have not changed. **d.** Correlation coefficient for the $y = -1$ phase computed for the symmetric model. Here, the correlations between $\varphi(t_{2A})$ and $\varphi(t_{2B})$ are negative. [Figure adapted from Ref. [173] based on different data]

In Fig. 7.31**c**,**d**, we show the correlation coefficient computed for the $y = 1$ and $y = -1$ phases, respectively. In addition to the trivial perfect correlation between symmetrized features, whose sign reflects the original orientation of the hopping links, we find positive correlations between $\varphi(t_{2A})$ and $\varphi(t_{2B})$ for the $y = 1$ phase and negative correlations for $y = -1$. The latter is expected, however, the former is not. We have already seen in both marginal and joint PDFs that equal values of the two phases are extremely unlikely compared to opposite. Clearly, this is related to the measurement performed by the correlation coefficient. Since $r(x_i, x_j)$ represents a correlation measure that describes how the two random variables change w.r.t. their respective mean, it does not make any statement whatsoever about the actual values of x_i and x_j. In many cases where the exact values are not necessarily obvious from the marginal distributions

we would be interested also in another correlation function of the type

$$c(x_i, x_j) = \frac{\mathrm{E}[x_i x_j]}{\sqrt{\mathrm{E}[\mathrm{x_i^2}]\mathrm{E}[\mathrm{x_j^2}]}}, \tag{7.119}$$

that does not include the subtraction of the mean, and therefore also respects the actual values of the variables. Since it arises from the definition of the correlation coefficient by setting $\mathrm{E}[x_i] = \mathrm{E}[x_j] = 0$, most properties are the same. However, for statistically independent variables we now have $c(x_i, x_j) \propto \mathrm{E}[x_i]\mathrm{E}[x_j]$, which is generally nonzero. By computing Eq. 7.119 for our features we obtain as expected strong negative correlations for both topological phases, which reflects the sign structure expected from the marginal distributions. We note that the definition of Eq. 7.119 with implied zero-mean makes sense considering the fact that all features subject to no further conditions have zero mean by construction and are uncorrelated. Only upon projecting onto the subset corresponding to a particular label do we shift the mean away from zero. The correlation function $c(x_i, x_j)$ measures if this shift happens in the same or opposite directions. This information is in a sense orthogonal to that revealed by the correlation coefficient $r(x_i, x_j)$ and thus it is only reasonable to consider both quantities for a maximum of information recovered from the data.

Effective model

Since we are already working on a streamlined model with a smaller number of parameters, it is now time to use the information revealed by the data analysis to define an effective model for the topological phase. To this end, the actual parameters are required, i.e., independently of the reference point. These are computed simply by adding the reference point to the feature values according to $\mathbf{t} = \mathbf{x} + \mathbf{x}_{\mathrm{ref}}$. The marginal distributions we obtain are shown in Fig. 7.32 and we use these to extract most likely values.

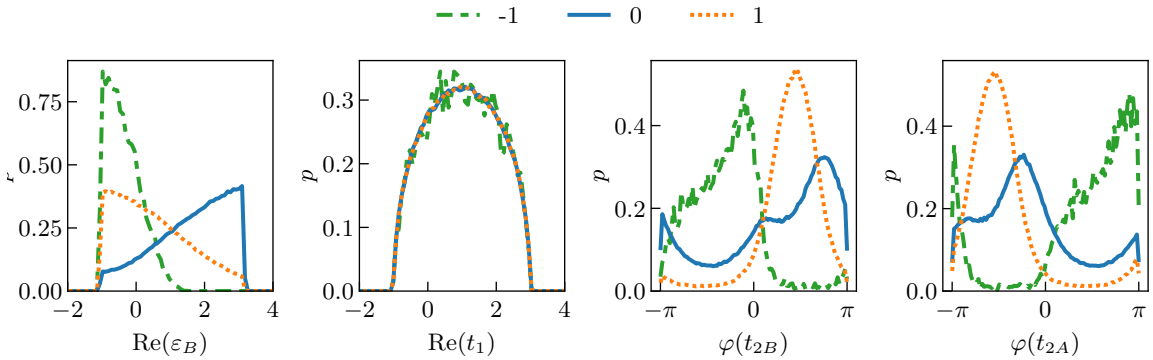

Figure 7.32: Marginal distributions $p(x_i|y = l)$ of the shifted features $\mathbf{x}_i + \mathbf{x}_{\mathrm{ref}}$ that correspond to the actual hopping parameters for $l \in \{-1, 0, 1\}$. Real parts are only shifted in energy w.r.t. Fig. 7.30. The two phases show a very clear preference for a particular (opposite) sign and the reversed configuration is extremely unlikely. The estimates of the distributions for $y = -1$ are not fully converged due to the relatively small number of samples. Nevertheless, it is clear that these features perfectly distinguish the two topological classes from each other.

The distributions for $\mathrm{Re}(\varepsilon_B)$ and $\mathrm{Re}(t_1)$ are the same as before up to a shift due to the linearity of the transformation. Apparently, small values $m \to 0$, i.e., $\varepsilon_B \to -1.05$ are preferred for topological phases. The peaks observed here lie at the boundary due to the limits of the sampling space Ω. For the phases we now observe a clear preference of the Haldane sign structure with almost no samples lying in the forbidden regime (opposite flux) for the $y = 1$ phase. The

statistics for $y = -1$ are rather poor due to the low number of samples, however, even in this case the data allows for an understanding of the typical configuration. Our effective model is comprised of the most relevant features, i.e., $\varepsilon_B, \varphi(t_{2A}), \varphi(t_{2B})$, with the values given in Table 7.5. The $y = 1$ model describes the essentials of the phase rather well, for $y = -1$ we would prefer more data. Although the effective model is a perfectly accurate description of the characteristic configurations for $y = -1$, a model with symmetric phase values would be obtained by using this estimate as a reference point and reiterating the procedure, since the asymmetry is merely a consequence of the sparseness of the data for that phase.

y	ε_B	$\varphi(t_{2A})$	$\varphi(t_{2B})$
1	≈ 0	$+\frac{\pi}{2}$	$-\frac{\pi}{2}$
-1	≈ 0	$-\frac{\pi}{4}$	$+\frac{3\pi}{4}$

Table 7.5: Effective description of typical parameters for the two topological phases obtained from the symmetrized data tailored to the $y = 1$ phase, cf. Fig. 7.32. Given no residual correlations, we can extract the characteristic values from the peaks of the marginal distribution. For ε_B the values are limited by the finite sampling interval, and therefore, the peaks lie at the boundary. Sampling around this new point or using a larger interval reveals that $\varepsilon_B = 0$ is, indeed, a maximum.

We have not performed the same detailed analysis in terms of a symmetrized model for the $y = -1$ phase, where we had found negative correlations between the nearest neighbor hopping parameters. Everything we have done so far is applicable to this case as well and we expect a resulting description with the same number of parameters, where only the sign of one t_1 is reversed.

7.7.6 Removing the Initial Bias – Towards Predictive Power

Our analysis so far has led us to a rather good understand of the Haldane phases with all information being extracted from the data. Our previous knowledge was used only for verification, except for the choice of the reference point. Clearly, most of the structure of the data must be a consequence of the fact that the initial point, around which we looked for traces of topological phases, was already corresponding to a Haldane configuration. In a way, this means that all of our results up to this point were obtained based on an informed bias that we infused our algorithm with. We will now demonstrate that this reference point can, in fact, be obtained by virtue of another analysis that follows our previous procedure, and therefore, we will finally obtain a completely unbiased algorithm.

We implement this minimal bias by constructing our reference point rather arbitrarily and motivated predominantly by convenience in terms of a purely real configuration of hopping amplitudes that satisfy

$$t_{ij}(\mathbf{R}) = \begin{cases} \frac{1}{\|\mathbf{R_{ij}}\|_2} & \text{if} \quad \| \mathbf{R}_{ij} \|_2 \neq 0 \\ 1 & \text{else,} \end{cases} \tag{7.120}$$

where $\mathbf{R}_{ij} = \mathbf{R} + \mathbf{r}_i - \mathbf{r}_j$, i.e., the distance between the real-space coordinates of two site-orbitals i, j. While this is very reasonable for single-orbital materials, it is not necessarily satisfied for multi-orbital systems, where the overlap of orbital wave-functions can lead to a lowering of hopping amplitudes at small distances due to the shape and orientation of the orbitals involved. However, by setting the spread parameter α appropriately, i.e., $\alpha > 1$, these cases are covered by the sampling procedure. Hence, while certainly being physically motivated, the configuration defined by Eq. 7.120 is still very general. A bias towards any particular topological phase on the other hand is entirely absent.

For the honeycomb lattice we fix the lengths of nearest neighbor links to 1 and obtain for next-nearest neighbors $\sqrt{3}$, 2 for next-next nearest neighbors and $\sqrt{7}$ for neighbors of fourth degree. With multiplicities of 3, 6, 3 and 6, respectively, we arrive at 18 complex features plus the one real feature for the local potential after fixing the zero of the energy scale, which means that we are sampling in a 37-dimensional feature space of vectors (in 19-dimensional complex space for brevity)

$$\mathbf{x}_i = (0, \varepsilon_B, t_1^1, t_1^2, t_1^3, t_2^1, t_2^2, t_2^3, t_2^4, t_2^5, t_2^6, t_3^1, t_3^2, t_3^3, t_4^1, t_4^2, t_4^3, t_4^4, t_4^5, t_4^6) \tag{7.121}$$

with a reference point

$$\mathbf{x}_{\text{ref}} = \Big(0, 1, 1, 1, 1, \underbrace{\frac{1}{\sqrt{3}}, \ldots, \frac{1}{\sqrt{3}}}_{\times 6}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \underbrace{\frac{1}{\sqrt{7}}, \ldots, \frac{1}{\sqrt{7}}}_{\times 6}\Big). \tag{7.122}$$

We use $\alpha = 1.5$ and generate a data set as before using the uniform distribution of Eq. 7.24 that creates a cloud of $n_{\text{samples}} = 10^7$ data points around the reference point. The composition of the data set is illustrated in Fig. 7.33**a**, where we immediately notice a drastic difference compared to the previous approach: the majority of samples is now metallic. In fact, only $\approx 2.4\%$ of our samples are insulating and only $\approx 44\%$ of these are topologically non-trivial. This is primarily related to the fact that the generic reference point that we chose does not automatically open a large band gap. Instead, the reference point is metallic with $E_g \approx -0.51$a.u. At the same time, the "direct gap", i.e., the gap at each k-point

$$\Delta E = \min_k \{\epsilon_2(k) - \epsilon_1(k)\}, \tag{7.123}$$

where $\epsilon_\alpha(k)$ is the dispersion of band α, is around $\Delta E \approx 1$a.u. Assuming for now that this is also true for our samples, we can assign a Chern index to each one of them and find in total seven distinct topological classes: $y \in \{-3, -2, -1, 0, 1, 2, 3\}$. The number of samples with $y \notin \{-1, 0, 1\}$ is so small, however, that we cannot obtain useful statistics for these. These phases that are not related to the Haldane phases correspond in total to only 1.6% of our data set, which amounts to roughly 10^5 samples. The majority of the samples is distributed over the three phases known from the Haldane model, where the trivial phase takes up the majority of the data set and the two topological phases each contribute about 20.6%, i.e., a little over 2×10^6 samples each. The fraction of metallic states is $\approx 97\%$ for both the trivial and the two topological phases. Since the Chern number is only guaranteed integer-valued for systems without degeneracies between conduction and valence bands, this is problematic. We therefore investigate the band separation defined in Eq. 7.123 for all samples in addition to the band gap. The resulting distribution is shown in Fig. 7.33**b**, where we find a distribution that is mostly concentrated around $\Delta E \approx 0.5$. We defined the threshold for metallic samples at $E_g \approx \sqrt{2}\frac{\pi}{82} \approx 0.054$, where we took into account the finite resolution in momentum space of $\Delta k = \frac{2\pi}{82} \approx 0.08$. Since the value of the gap is least certain at k-points that are farthest away from one of our grid points we investigate exactly these center points that lie between points of our k-grid. These points have a distance $d = |0.5(\Delta k, \Delta k)^T| \approx \Delta k/\sqrt{2}$ to the nearest grid points. With $\Delta\epsilon/d = v$ we obtain for $v = 1$, $\Delta\epsilon = \Delta k/\sqrt{2} = \sqrt{2}\pi/82$. The value $v = 1$ indicates that we only miss metals whose bands have a slope steeper than 1a.u., however, based on Fig. 7.33**b** we could also choose $v = 4$ without changing the qualitative result. A band separation smaller than this threshold is only found for ≈ 80000 samples, which are statistically insignificant. For comparison, we also show the distribution of band gaps, which are mostly negative due to the overlapping energies of the bands.

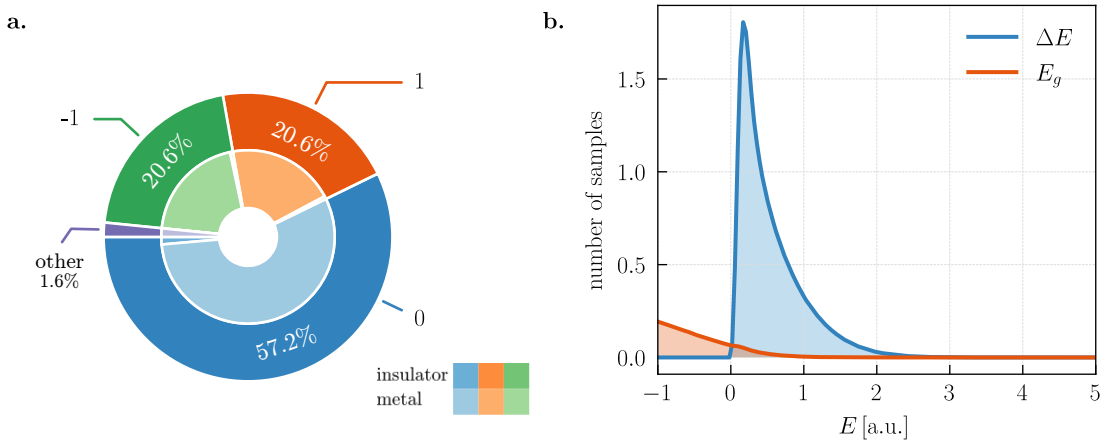

Figure 7.33: Overview statistics of the generic honeycomb data set. **a.** Doughnut chart illustrating the fraction of samples per label. Of the total seven labels only three appear in significant numbers. Notably, almost all of the trivial ($y = 0$) and non-trivial samples are metallic. **b.** Distribution of the band separation ΔE [Eq. 7.123] and band gap E_g of all samples. The majority of our data set is found to have bands separated by energies ≥ 0.2a.u. and almost all ≥ 0.05. [Subfigure **a.** adapted from Ref. [173] based on different data. Compared to Ref. [173], we are using a narrowed sampling interval for ε_B that excludes, in particular, trivial insulators found at large ε_B.]

This analysis shows that most of the metallic samples feature separated bands that allow for the assignment of a well-defined topological index. Since no band crossings exist, these configurations are smoothly connected to the topologically insulating phase that should share many of the same properties. We will confirm this later by explicitly comparing the distributions for topological insulators and "topological metals".

Since we have now confirmed the accuracy of the assigned labels and made sure that the data set does contain a reasonable number of interesting samples that is large enough to extract a useful description out of the statistics, we can now move on towards phase 2, i.e., the dimensional reduction. This step is even more important now, since by taking into account also neighbors of third and fourth degree we blew up the feature space to 73 dimensions (1 real + 18 complex à 4 reals each). We estimate the distribution functions from the samples and compute the Bhattacharyya distance as our importance score. The result is shown in Fig. 7.34. In subfigures **a.** and **b.** we explicitly differentiate between the statistics obtained from all samples (Fig. 7.34**a**) and from the subset of insulating samples alone (Fig. 7.34**b**). The calculations shown are for the $y = 1$ phase only. Independent of the subset of the data taken into account, the local term shows again the highest contrast. Below this, however, we are faced with a predicament. Apparently, the exact order of importance scores depends on whether we take into account only insulating or also metallic phases, which seems to invalidate our earlier assumption that we can infer information from the metallic samples as well. We choose to investigate this more in-depth, and by comparing the features in the upper and lower halves of the importance spectrum, respectively, we find that changes happen mostly locally, i.e., important features do not suddenly become unimportant and vice versa if we exclude metallic samples. Examples for unimportant features are, e.g., all features related to t_3 and t_4, which are found at the lower end in both rankings. Important features that change ranks are, in particular, $\varphi(t_{2A})$ and $\mathrm{Re}(t_{2A})$, which still remain in the higher ranked regime.

Having established that the importance ranking is not affected significantly by the metallic samples we now consider the second difference between Fig. 7.34**a**,**b**—the existence of a well-defined cutoff value. For the complete data set shown in Fig. 7.34**a**, we find several smaller

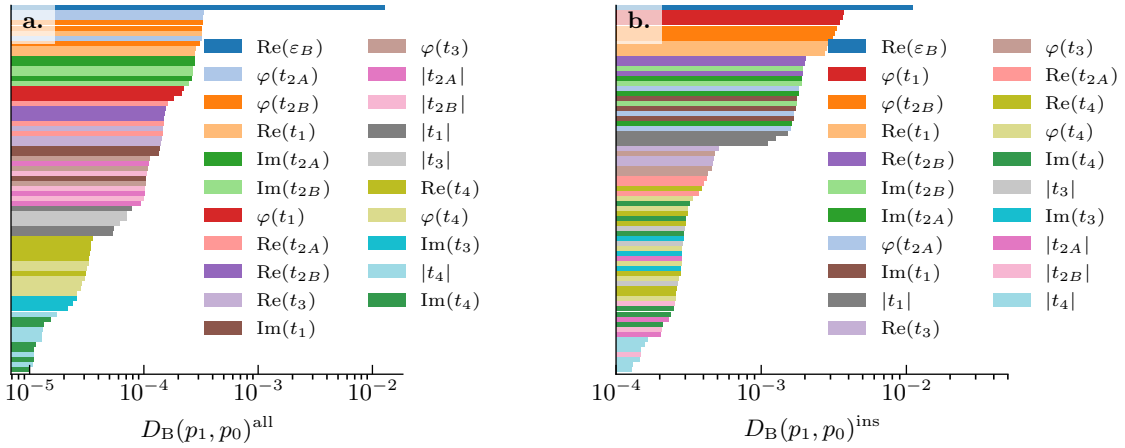

Figure 7.34: Importance score in terms of the Bhattacharyya distance [Eq. 7.62] taking into account **a.** all samples and **b.** only insulating samples. The colors are the same to aid comparability. Not only the ranking changes but also the location of discontinuities. Both plots show a drop in importance below $|t_1|$, where we find the features related to t_3, t_4. The ordering among the highest-rated features changes, however, due to the small number of insulating samples, we expect errors in the corresponding importance score.

jumps in the spectrum, the most useful one between $|\varphi(t_1)|$ and $\mathrm{Re}(t_{2A/B})$. Labeling $\varphi(t_{2B})$ as unimportant, however, contradicts the reality of the insulating phases, since as can be seen in subfigure 7.34**b**, $\varphi(t_{2B})$ belongs to the most descriptive set of features for those phases that are of actual interest to us. The statistics obtained for insulators alone show only one jump in the importance scores that effectively separates features related to neighbors of first and second degree from those related to longer-range hopping (plus $|t_{2A/B}|$).

Clearly, simply taking into account all features with no regard as to the conducting behavior cannot yield an accurate description. We show in the following, however, that we need not throw away the abundance of metallic samples. To this end, we take a look at the marginal distributions of some of the important features in Fig. 7.35. In order to avoid confusion, we plot the distributions related to insulating phases in separate panels. The top row considers only insulating samples, where we used the threshold $E_g \geq 0.04$ to separate metallic configurations from those that are insulating. We confirmed that the distributions are insensitive to increasing the threshold to, e.g., $E_g \geq 0.08$. The bottom row corresponds to the statistics over the entire data set regardless of the conduction behavior. Inspecting first the overall case (both insulating and metallic, bottom row) that was also used to compute the importance scores in Fig. 7.34**a**, we find a clear contrast in $\mathrm{Re}(\varepsilon_B)$, a minor contrast in $\varphi(t_{2A/B})$ and barely a difference in $\mathrm{Re}(t_1)$, which reflects the order in which these features appear in the ranking. Regarding the ranges of values, a negative local potential is apparently more favorable for topological phases (note that according to Eq. 7.122, a perturbation of $\varepsilon_B = -1.05$ corresponds to the symmetric case with no mass), while t_1 largely does not matter. The two phases of the next-nearest neighbor hoppings on the other hand do play a role and we find an ever so weak tendency for the t_{2A} phases towards assuming negative values and vice versa for t_{2B}. The total distribution is also shown here and corresponds simply to the uniform distribution that we used in the sampling algorithm.

Comparing the distributions from the overall data set with those of insulating samples we find several differences. For one, the distributions of topological samples seem to change, which is evident, e.g., for ε_B, where the tail of the distribution towards positive values is raised slightly. $\mathrm{Re}(t_1)$ reveals a pronounced tendency towards positive values, while the two phase terms favor

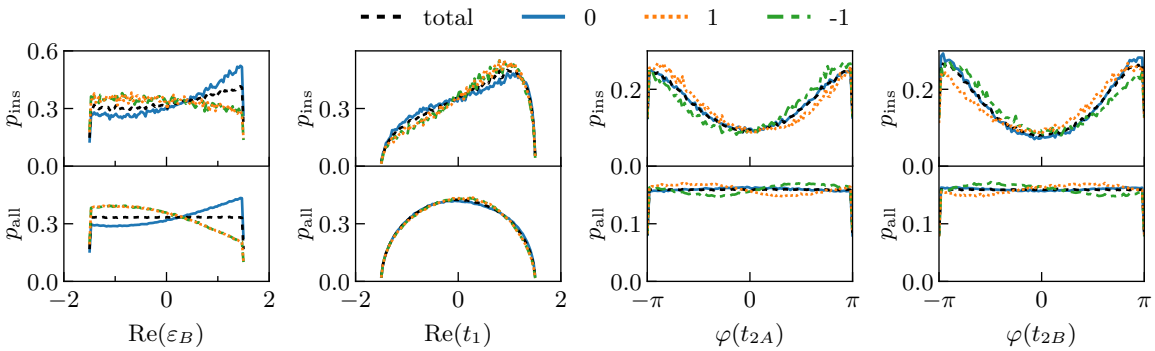

Figure 7.35: Marginal distributions of selected features for the most general honeycomb lattice model. Top row: only insulating samples ($E_g \geq 0.04$) are taken into account, bottom row: distributions estimated using all samples. By comparison between the two sets of distributions we find mostly qualitative agreement. The total distribution is shown as well, which is the uniform distribution (or a projection of it) for all samples, but something completely different for the insulating phases alone. The marginal distributions of insulating topological samples are approximately products of overall topological and overall insulating distributions. [Figure in parts adapted from Ref. [173] based on different data]

values away from zero with a preference for negative (t_{2A}) or positive (t_{2B}) values. While this is qualitatively similar to what the complete data set shows, the quantitative differences are too large to neglect. We can understand the relation between the two sets of distributions much better by recalling that the samples corresponding to non-trivial topological insulators are described by the compound condition $y \neq 0 \wedge E_g \geq 0$. By making use of the definition of the joint probability we have

$$p_{\mathrm{ins}}(y = l) = p(y = l, E_g \geq 0) = p(y = l | E_g \geq 0) p(E_g \geq 0), \tag{7.124}$$

where $p(y = l | E_g \geq 0)$ is the sought-after description of topological labels assuming that we need not worry about the existence of a band gap, which itself is described by $p(E_g \geq 0)$.

Taking as an ansatz $p(y = l | E_g \geq 0) \approx p_{\mathrm{all}}(y = l)$ we confirm that this indeed produces distributions that look similar to the measured distributions of topological insulators. The difference in the two sets of topological distributions can therefore be explained through the probability distribution of the generic insulator $p(E_g \geq 0)$ (total distribution in the upper row in Fig. 7.35). Consequently, although the importance of individual features is not reproduced particularly well when including metallic samples, the information about topological phases can still be inferred, which confirms our earlier assumption that also samples labeled as "topological metals" can be used for constructing topological insulators. This hinges, of course, on the fact that the bands are separated, since otherwise the Chern number itself is not well-defined. This result is particularly important in light of the very small fraction of insulating samples.

The ranges of values that we observed show that large ε_B is undesirable and should be avoided in a topological model, where we would fix a value below the initial $\varepsilon_B = 1$. In addition, we observe that $\varphi(t_{2A/B}) \approx 0$ is unfavorable, however, we have to keep in mind that the distributions shown in Fig. 7.35 do not include the reference point, and therefore, do not represent the parameters that enter the Hamiltonian.

Apparently, the signal in the phase features is extremely weak, which is a consequence of the large number of degrees of freedom in the arrangement of values on equivalent hopping links. In order to improve this, we need to introduce symmetries that we choose based on the correlations between features. We computed the redundancy between all pairs of variables, but did not find any notable values. The largest redundancies found were of the order of $R \sim \mathcal{O}(10^{-4})$, which

we consider to be negligible. Given no discernible correlations between pairs of features, we can perform a reduction of the number of degrees of freedom. This is done rather carefully, assuming symmetries between all equivalent hoppings, but taking into account also features that were determined unimportant previously. The resulting (5+1)-dimensional feature space contains one real feature and five complex features:

$$\mathbf{x} = (0, \varepsilon_B, t_1, t_{2A}, t_{2B}, t_3, t_4), \tag{7.125}$$

where we use the phase information obtained from the marginal distributions of next-nearest neighbor hoppings to align the hopping links according to their preferred (anti-) clockwise order. Now, we repeat basically the same steps as before to obtain a new data set that hopefully produces a better contrast and allows for an easier definition of a more refined reference point.

Generating $n_{\text{samples}} = 10^7$ samples with these imposed symmetries on the hopping parameters, but starting again from the most generic reference point, i.e., Eq. 7.120, we obtain the data set summarized in Fig. 7.36**a**. The fraction of non-trivial samples has increased slightly, and we have now more than 6% of additional topological samples that have mostly $y = 2$. The fraction of insulating states has increased as well, which means that we can now expect decent statistics for insulating samples. Again, 99+% of samples have separated bands. We compute again the importance score for all features, which is shown in Fig. 7.36**b**. After having established that the topological information is contained also in the metallic samples we make use of the better statistics guaranteed by the larger size of the complete data set compared to insulating samples only. In fact, using a smaller data set to compute the feature importance can easily introduce numerical errors due to the reduced accuracy of the estimated probability distributions. The resulting importance ranking is consistent with what we obtained for the more general model, indicating that we did not remove any particular component that was captured before from our model by introducing symmetries.

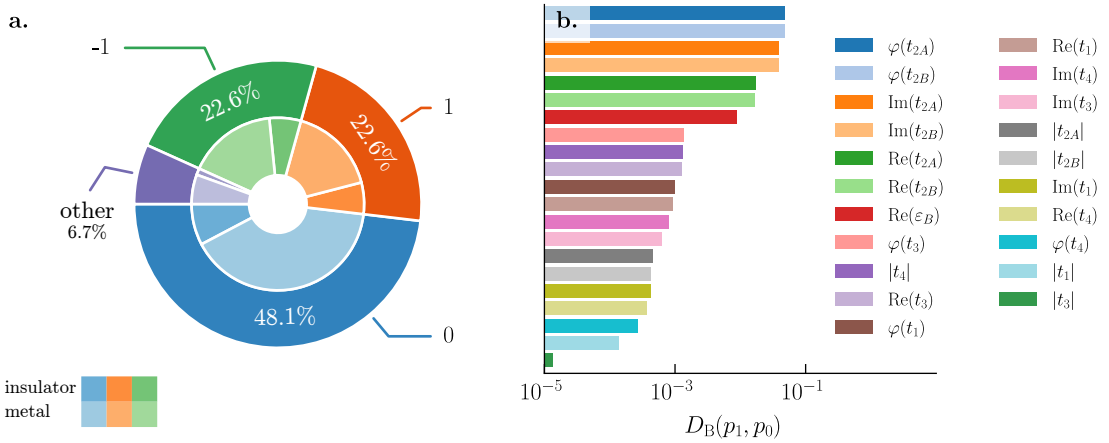

Figure 7.36: **a.** Composition of the symmetrized honeycomb data set. The number of topological samples has increased compared to Fig. 7.33 and so has the number of insulating samples. **b.** Importance score for the $y = 1$ phase computed using all samples. The ranking does not change qualitatively compared to the completely unsymmetric case, i.e., next-nearest neighbor hoppings are still most important and third- and fourth-nearest neighbors do not play a role. [Subfigure **a.** adapted from Ref. [173] based on different data]

The next task is to define an improved reference point that is informed by what we learned about the hopping parameters of typical topological samples. Motivated by the importance ranking, we decide to remove the 3rd and 4th nearest neighbor hoppings from our model since

they do not seem to contribute significantly to the understanding of the topological phase diagram. In order to extract a good effective model we then investigate the marginal PDFs of the remaining features more closely.

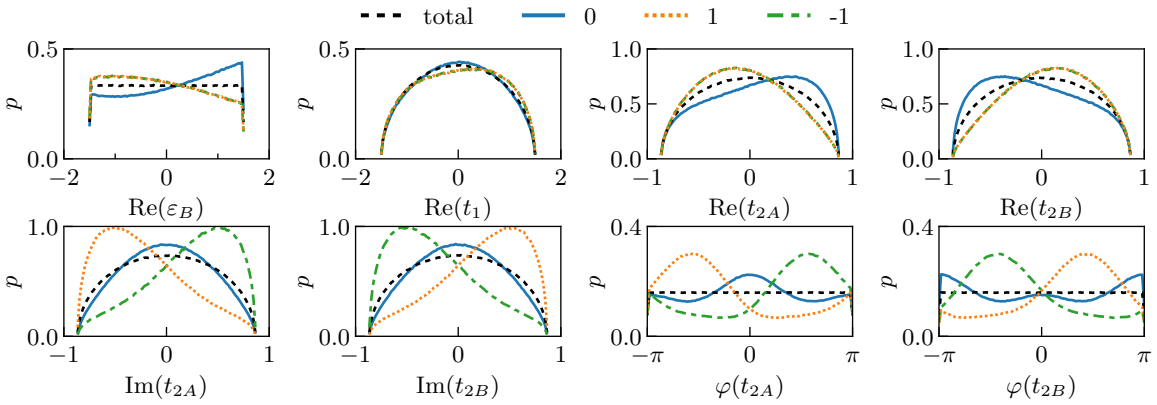

Figure 7.37: Marginal distributions for the most relevant features. We show the real parts of the local potential ε_B and the nearest-neighbor hopping t_1 and real and imaginary parts and phase of next-nearest neighbor terms $t_{2A/B}$. The real parts of the hopping features are generally less relevant than their imaginary parts, which is consistent with the importance scores shown in Fig. 7.36**b**. Topological phases are favored by negative ε_B and negative (positive) $\mathrm{Im}(t_{2A})$ and positive (negative) $\mathrm{Im}(t_{2B})$ for $y = 1$ ($y = -1$) compared to the original $\mathbf{x}_{\mathrm{ref}}$. [Figure in parts adapted from Ref. [173] based on different data]

The marginal PDFs estimated from the data set for a selection of features are shown in Fig. 7.37. The lower importance of $\mathrm{Re}(t_1)$ and $\mathrm{Re}(t_{2A/B})$ is clearly reflected here. For the most descriptive effective model we are looking for properties that distinguish the topological phases from the trivial phase. One such property is apparently the local potential ε_B, which was originally taken to be equal to 1 in our arbitrary units of energy. Trivial samples are obtained for values that are significantly larger than 1, while predominantly non-trivial samples are obtained for values around $\varepsilon_B = 0$. This can be understood in terms of limiting cases, where $\varepsilon_B = -m \to \pm\infty$ leads to a trivial band insulator. From this small peek at the general honeycomb model, any value $\varepsilon_B \in [-1, 1]$ would be a good choice. Interestingly, though, the mass term does not distinguish between the two topological phases. In contrast to this, the imaginary part of the next-nearest neighbor hoppings performs exactly this distinction. For the $y = 1$ phase we observe a predominantly negative shift of $\mathrm{Im}(t_{2A})$ w.r.t. the real reference point. The opposite sign is observed for the $y = -1$ phase, i.e., the sign of the imaginary part of $t_{2A/B}$ is a strong descriptor of the topological phase. This does not work without fail, though, as is apparent from the finite probability density for the opposite sign. This is an indication towards additional information hidden in the correlations between individual parameters. The phases of $t_{2A/B}$ are shown only for completeness and display the same characteristics as $\mathrm{Im}(t_{2A/B})$. In fact, it turns out that for the small number of samples that show the opposite sign, the third and fourth nearest neighbors do play a role. This type of dependency is rather difficult to observe in the data immediately. One way to go about this would be to first focus on the majority cases as we have done here by determining, e.g., the position of the maxima of the marginal distributions. Then, one can proceed with the same analysis for the remainder of the samples, i.e., outliers, those that are not compatible with the majority. We will explain a general algorithm applicable to generic systems in the next section.

Having established particular ranges of values of the important parameters in our model for the two topologically non-trivial phases, we can now infer a new reference point that lies

closer to the respective topological phase and is therefore likely to produce a larger number of useful samples. We pick a local potential $\varepsilon_B \in [-1, 1]$, and imaginary next-nearest neighbor hoppings $\varphi(t_{2A/B}) = \mp\frac{\pi}{2}$ for the $y = \pm 1$ phase. The nearest neighbor hopping t_1 was found to be rather non-descriptive, so we are inclined to keep it at a generic value. These choices correspond essentially to the reference point chosen earlier that was motivated by the Haldane model. Therefore, we have shown that no prior information is needed to extract the Haldane configurations as representatives for the topological phase from the data set. In a way, this means that the Haldane model can be thought of as the prototypical topological model on the honeycomb lattice, since the corresponding configurations are most likely to produce topological phases. On the other hand, we also find configurations that cannot be realized within the Haldane model by introducing certain correlations between hopping parameters. Starting from an improved reference point is usually more informative, since noise generated by metallic phases is removed. This facilitates, in particular, the analysis of correlations, which are usually blurred out in noisy data.

7.7.7 General Algorithm

We chose to present a "historical" introduction of the method in the preceding sections that reflects how it was developed. From this alone it might be rather unclear how an application to any other generic system is best approached. We cover such general applications here by discussing a rather generic algorithm that collects all the pieces and inserts them into a grander scheme.

Clearly, the intent is to facilitate the development of an understanding without any previous knowledge. Therefore, any most general approach should begin with a completely unbiased data set that can be generated from a reference point in analogy to Eq. 7.120. This first data usually contains a lot of noise and is therefore only used to define a new reference point that is likely to produce a higher quality data set in the sense that a larger fraction of interesting phases other than the generic trivial phase is contained.

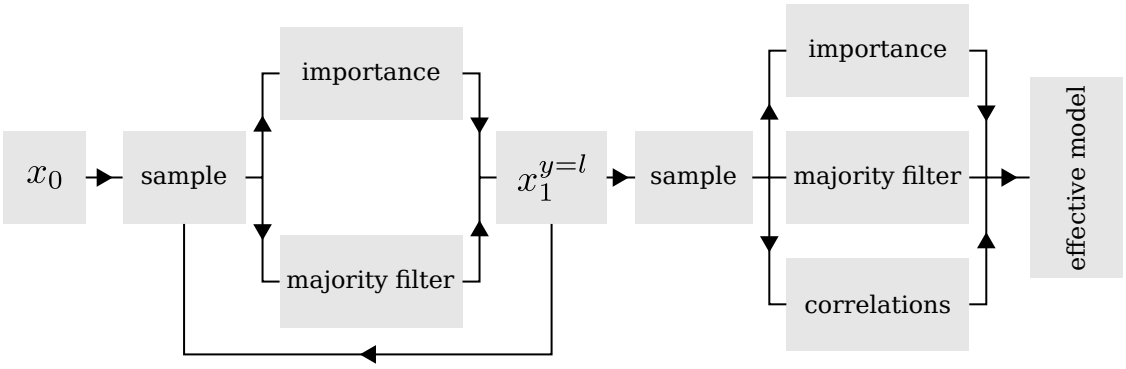

Figure 7.38: Illustration of a typical algorithm applying the statistical method. Starting from a generic x_0, a data set is generated that is then analyzed in terms of feature importance and typical values thereof to arrive at a more informed parameter set $x_1^{y=l}$ for class label l. This step can be performed iteratively until phases are reasonably separated. The last data set is then analyzed to obtain an effective model for each phase.

We illustrate the typical program flow of the proposed algorithm in Fig. 7.38. Starting from the maximally unbiased initial reference point x_0, samples are drawn from a random distribution. The resulting data set is then analyzed with the methods introduced in our discussion earlier, i.e., importance scores to reduce the number of features required for a description of the topological

phase and a majority filter that projects onto the maximum of the distribution. Via this analysis one then arrives at an updated reference point $x_1^{y=l}$ that generally differs between nonequivalent phases. Instead of a single shot update $x_0 \to x_1$ it is reasonable to define

$$x_i = x_{i-1} + \delta x_i, \tag{7.126}$$

where δx_i is obtained from the maxima of the marginal feature distributions. More generally, via a learning rate $\alpha \in (0, 1]$, we can define the iterative update as

$$x_i = (1 - \alpha)x_{i-1} + \alpha \delta x_i. \tag{7.127}$$

With this choice we allow the algorithm to discover more phases along the way. The aim of this iterative part is to increase the number of samples corresponding to a particular phase by improving the underlying reference configuration. At the same time the feature importance will increase for those features that are most relevant for the respective phase. Having arrived at a reference configuration that shows a clear distinction between different phases, the data is then analyzed further by taking into account also correlations between the reduced number of features. This information can subsequently be used to impose symmetries on the parameters and arrive at a minimal effective model.

We note that the examples discussed in this section, i.e., the Haldane model and honeycomb lattice, are rather simple cases that we used to deliver a proof of concept. The simplicity of the model together with our previous knowledge of the phase diagram for the Haldane model allowed us to validate many aspects during the analysis. The benefits of automation become much clearer when studying lattices with increased unit cells, where the definition of equivalent features is no longer straight-forward, and therefore, the importance score plays a much more prominent role in sorting through an abundance of features. At the time of writing, active research is being done where the statistical method introduced here is applied to a kagome lattice, which represents a step up in terms of complexity with a unit cell of three sites and six nearest neighbors that allow for a much higher complexity already at the level of first degree neighbors. We will discuss the three-site unit cell and few results from the ongoing work on the increased 12-site unit cell in Sec. 7.8.

While we expect that a program following this general algorithm can enable a large degree of automation, and therefore, make accessible systems with a large number of parameters that include large unit cells and long-range hopping, this type of autonomy is expected to require more intensive development of the methods introduced here. We will mention a few possible improvements below.

Feature Engineering

During the discussion of the sampling algorithm in Sec. 7.5, we had already hinted at the possibility of feature engineering to improve the effectiveness of the method. Clearly, we are limited by the choice of the features through the simple transformation to real features that is given in Eq. 7.22. It cannot be expected that, in general, the topological phase can be characterized exactly and in an understandable manner through just the values of these generic features. In fact, we observed that this is possible with rather high precision for the Haldane case when reducing the degrees of freedom through symmetrization, cf. Fig. 7.32. For an arbitrary system, the phase transition lines will likely not be aligned with these parameter axes so that the corresponding marginal distributions will have finite weight across a large region, which then reduces the overall contrast between different phases. By constructing additional features out of the important candidates, it is, in principle, possible to increase the contrast. Assuming that we use the Bhattacharyya distance as the importance/contrast measure, the aim of the

iterative algorithm would be the maximization of the importance over all possible combinations of available features, where we can use the previous importance measurements as a guide to reduce the number of combinations. For ideal contrast, i.e., completely separated marginal distributions of a feature x_i for labels l_1, l_2, the Bhattacharyya distance diverges. This is typically not achieved due to the complexity of possible shapes of transition lines, however, large values would be enough to further the understanding of the local behavior (around the reference point).

We note that this task would be an ideal candidate for neural network applications, where $1/D_{\mathrm{B}}(p(x_i|y = l_1), p(x_i|y = l_2))$ is used as a loss function, however, the resulting dependency would likely be extremely complicated and not necessarily helpful.[10] Instead, we suggest to restrict to simple polynomial combinations of features and ratios. A linear transition line that can be a rather good approximation over a small region is already described perfectly by the sum of features $x_{n_{\mathrm{features}}+1} = x_i + x_j + \ldots$. We could instead try all suitable combinations together with few relevant powers as

$$x_{n_{\mathrm{features}}+1} = P_m(\{x_i \mid D_{\mathrm{B}}(p(x_i|y = l_1), p(x_i|y = l_2)) > \epsilon\}), \tag{7.128}$$

where P_m is a polynomial of maximal order m, and ϵ describes the importance cutoff that can be chosen dynamically to only include the most important features. This should be feasible computationally and would allow for an extraction of valuable information about multivariate correlations that is otherwise extremely difficult to obtain. We note that in order to interpret the polynomial feature it would be helpful to look only at those terms with the largest coefficients.

Working With Complex Features Entirely

During the entire discussion we used a specific mapping that relates the initially complex features to real features as is required by our methodology. This mapping was given by Eq. 7.22, where we basically use a redundant representation of the data in order to determine the best choice. Working with polynomial features that depend on several of the original features becomes very complicated if we insist on this mapping, since we have to decide on the order of operations: transform to real features first and then compute polynomial features or the other way around?

Thankfully, there is no necessity for a mapping to real features. In fact, we can work entirely in the framework of complex features while only sacrificing the ability to easily illustrate the origin of the statistical distances, i.e., the relation between marginal distributions. Keeping the complex description, however, will remedy the problem of ill-defined order of operations and will also simplify the interpretation of the importance score, since for each hopping parameter there is only one value that contains all information.

The generalization to complex-valued features is rather straight-forward, since we only have to redefine the marginal distributions. Instead of the real function $p(x)$ ($\mathbb{R} \to \mathbb{R}$) we now define

$$\rho : \mathbb{C} \to \mathbb{R}, \rho(x) = p(\mathrm{Re}\,[x]\,, \mathrm{Im}\,[x]), \tag{7.129}$$

where $p(\mathrm{Re}\,[x]\,, \mathrm{Im}\,[x])$ is the joint distribution function of the real and imaginary part of x. With Eq. 7.129, we can also generalize the statistical distances, in particular, the Bhattacharyya and

[10] We do not recommend to use these statistical distances as a cost function in a training algorithm. Not only for the reason mentioned, but also because of the high computational cost of determining the Bhattacharyya distance that would have to be performed in each training iteration. Our remark was only meant to point out the equivalence between the approaches.

Hellinger distances

$$D_{\mathrm{B}}(\rho_1, \rho_2) = -\log\left[\int_{\mathbb{C}} \sqrt{\rho_1(z)\rho_2(z)}\mathrm{d}z\right], \tag{7.130}$$

$$D_{\mathrm{H}}(\rho_1, \rho_2) = \sqrt{1 - \int_{\mathbb{C}} \sqrt{\rho_1(z)\rho_2(z)}\mathrm{d}z}. \tag{7.131}$$

Given this generalized definition of distributions and statistical distances, the general algorithm can be applied independently of a particular choice of mapping to real features, which reduces arbitrariness and improves the accuracy, since despite the redundancy in the definition of the real features we do not necessarily include the optimal mapping. Due to the increased dimensionality of the domain of ρ compared to p, we can no longer easily inspect the marginal distributions, since plotting distributions for separate labels on top of each other would require making use of the third dimension, which comes at a loss of readability. One would instead have to plot distributions in separate panels which makes comparisons slightly more difficult. This disadvantage is entirely compensated by the importance score that encodes the comparison between two distributions in a single number.

We will demonstrate the use of this generalized formulation of statistical distributions in combination with the method of feature engineering in Sec. 7.8.

7.7.8 Further Comments

For completeness, we also want to comment on a few attempts that were made along the way that did not work out or did not provide any benefit over what was presented earlier.

The attentive reader might have wondered why we introduced the Gaussian distribution in Sec. 7.5, since we ended up using the uniform distribution all the time. The reason for this is partly historical, since the initial idea was to use a Gaussian around the reference point, and therefore assure some sort of closeness of the data points to the reference point. In order to be less biased, since the reference point does not really have any physical meaning for the study of phase diagrams that we presented here, we instead used the totally unbiased uniform distribution. The Gaussian distribution on the other hand—due to its builtin concentration around $\mathbf{x}_{\mathrm{ref}}$—is expected to be better suited for the material application that is outlined in Sec. 7.9.

In addition to the uniform and Gaussian distributions we also explored other possibilities. In order to simplify the interpretation of the data set, we defined what we call "ring distributions"

$$\rho_{\mathrm{ring}}(x) = \rho_{r,\alpha}(|x|)\rho_{\mathrm{uniform}}(\varphi(x)), \tag{7.132}$$

where $\rho_{r,\alpha}$ can be either the uniform or Gaussian distribution with mean r and spread α. This corresponds, in principle, to setting the reference point to zero and sampling data points around a spherical surface (in one complex dimension) at distance r from the origin. Given this construction of the data points, we automatically have a certain ratio $|x_i|/|x_j|$ fixed on average, so that the number of samples with longer-ranged hoppings larger than shorter range can be controlled more effectively. While the removal of the reference point suggests a more unbiased approach, the restriction to a small region of magnitudes for each hopping actually results in a less general ansatz that will expose the dependence of the label on the phase degrees of freedom, but not the magnitude.

In this context we also want to stress the importance of the reference point regardless of the distribution. In Sec. 7.5 we argued that a specific but generic choice of the reference point guarantees our samples to conserve a certain kind of physicality on average. While this is

certainly true, there is also a strictly technical necessity for introducing a finite reference point. We show this in terms of an example. Let us assume that $\mathbf{x}_{\text{ref}} = 0$. Since all our distributions are uniform in the complex phase of the samples, this means that the total marginal distribution for all samples is just a uniform distribution on $[0, 2\pi)$. With all parameters assuming allowed values $|x| \in [0, b]$ with some finite b, it is then evident that given any value for x_i we can choose the remaining parameters such that they compensate the tilt toward whatever topological phase the chosen value of x_i would favor. Therefore, once we average out (marginalize) all parameters except x_i, we do not expect to observe much, if any, contrast in the phase degree of freedom, since the information is hidden in the correlations that are much harder to extract. This is exactly what we observed in our tests. Thus, by omitting the reference point we cannot extract any information about the importance of the phase degree of freedom (that we found to be very important for topological phases) from the marginals. Choosing, however, a finite real reference point like we did removes this global symmetry, since our data set covers only an asymmetric subset of points around $\mathbf{x}_{\text{ref}}$ by construction, thereby assuring that the phase dependence is readily observable in the data without the need for any additional transformations.

Finally, we also experimented with using an iterative scheme, where we use knowledge about the marginal distributions for particular labels gained in one step for the generation of the data in the next. Given that we sample magnitude and phase independently, we have

$$p_{i+1}(|x_j|) = p_i(|x_j||y = l), \qquad p_{i+1}(\varphi(x_j)) = p_i(\varphi(x_j)|y = l), \tag{7.133}$$

i.e., the distribution used for sampling in step $i + 1$ is the measured distribution from step i for some selected label l that we want to optimize for. The general idea is that sampling from these distributions will lead to a machine that can generate only samples belonging to that particular class l.

Algorithmically, this works as follows. Given a distribution $p : \Omega \to \mathbb{R}^+$ with $\Omega = [a, b]$ for a random variable X we can compute the cumulative distribution function

$$F(x) = \int_a^x p(y)\ \mathrm{d}y \tag{7.134}$$

for $x \leq b$. By definition, F is bounded to the interval $[0, 1]$ and monotonic. If p does not have any roots on Ω then F is even strictly monotonic and therefore invertible with $F^{-1} : [0, 1] \to \Omega$. According to Sec. 4.1.4, the random variable $Y = F^{-1}(Z)$, with Z uniformly distributed on $[0, 1]$, yields a distribution $q_Y : \Omega \to \mathbb{R}^+$ that has the same distribution function as X. This is also apparent from the equality of the cumulative distribution functions

$$\begin{aligned} F_Y(y) &= P(Y < y) = P(F_X^{-1}(Z) < y) \\ &= P(Z < F_X(y)) = F_X(y). \end{aligned} \tag{7.135}$$

Therefore, evaluating F^{-1} with a uniform random variable yields samples that are distributed according to p. F and F^{-1} can be computed numerically from the measured data, so that we are able to produce samples from a measured distribution.

The problem with this approach is that the marginal distributions do not contain any information about the correlations between samples, and therefore, we were not able to significantly increase the fraction of samples from the target class. In order to apply this technique successfully one would instead have to sample from joint distributions that are much more difficult to measure due to the large numbers of samples required.

This brings us to the last point: using a joint distribution function for sampling. With the independent sampling approach employed here, we reach the desired goal of covering the

sampling space uniformly. A concentration around the reference point can be achieved by using instead a Gaussian distribution as noted above. However, the probability of finding samples in the vicinity of the reference point is given by

$$P(|\mathbf{x}_{\text{ref}} - \mathbf{x}| < \delta) \leq \prod_i P(|x^i_{\text{ref}} - x_i| < \delta), \tag{7.136}$$

which for the uniform distribution evaluates to $\prod_i \delta/|\Omega_i|$. This upper limit is in general very small for small δ due to the scaling $P \sim \mathcal{O}(\delta^{n_{\text{features}}})$, and therefore, the probability for obtaining a sample close to the reference point is actually vanishingly small. This is a consequence of the size of the sample space, which leads to a large distance between individual samples. For the Gaussian distribution we generally obtain a larger value depending on the standard deviation that is used. Even more control is possible through the use of a distribution $p(|\mathbf{x}|) = \mathrm{N}_{\mu=\mathbf{x}_{\text{ref}},\sigma}(|\mathbf{x}|)$, from which we can sample the distance from the reference point. Rescaling the uniform samples with the value obtained from this distribution will then assure that the majority of samples will lie close to the reference point.

7.8 Kagome Systems

Motivated by a recent surge in research activity on kagome metals that was fueled by the observation of symmetry-breaking charge order in KV_3Sb_5 [276–281] and an anomalous Hall effect in CsV_3Sb_5 [282], we started an investigation into the general topological phase diagram of the kagome lattice that should ultimately predict under which circumstances this class of materials can host topological insulators. This work is still ongoing and we present some of the current results in this section. First, we will discuss the system with the full translational symmetry of the kagome lattice and then comment with some results on the current state of the work on the symmetry-broken regime. In general, we define the Hamiltonian as

$$H = t \sum_{\langle i,j \rangle} c_i^\dagger c_j + t' \sum_{\langle\langle i,j \rangle\rangle} c_i^\dagger c_j + t'' \sum_{\langle\langle\langle i,j \rangle\rangle\rangle} c_i^\dagger c_j + \sum_i \varepsilon_i c_i^\dagger c_i, \tag{7.137}$$

where t, t' and t'' correspond to nearest-neighbor, next-nearest neighbor and next-next-nearest neighbor hoppings, respectively, and ε_i are the onsite energies. All hoppings can, in principle, be complex numbers. Note that the hopping terms are chosen according to convention with a negative sign, i.e., $t, t', t'' < 0$ for the generic reference point.

Like for the honeycomb lattice, the Bravais lattice of the kagome lattice is the triangular lattice, however, the size of the unit cell is increased to three. Due to this effective increase of the degrees of freedom, kagome systems can be considered one step up in terms of complexity. We illustrate the lattice and our choice of basis in Fig. 7.39, where in subfigure **a.** we show the lattice vectors $\mathbf{a}_1, \mathbf{a}_2$ and the corresponding unit cell that contains exactly three sites which are labeled 1,2,3 for future reference. The hoppings we take into account are shown in Fig. 7.39**b**, where we differentiate between links that connect sites within the cell and those that connect different cells in addition to the order of the neighbors (here, up to third neighbors). Coordinates of the neighboring cells are shown in terms of the lattice vectors. We choose here a specific convention for the choice of hoppings that can be described as most generic and is accomplished by including links based on an ordering of their coordinates. This means favoring, e.g., $(1, 0)$ over $(-1, 0)$ and with lower priority site $1 \to$ site 2 over $2 \to 1$, which correspond to the complex conjugate, respectively. With this generic choice that is easily automated and requires no laborious construction of lattice geometries[11] we minimize the initial bias and show in the

[11] The indeed very laborious illustration was completely unnecessary and serves only the presentation.

following that the data we obtain will naturally point us towards a more convenient choice of basis.

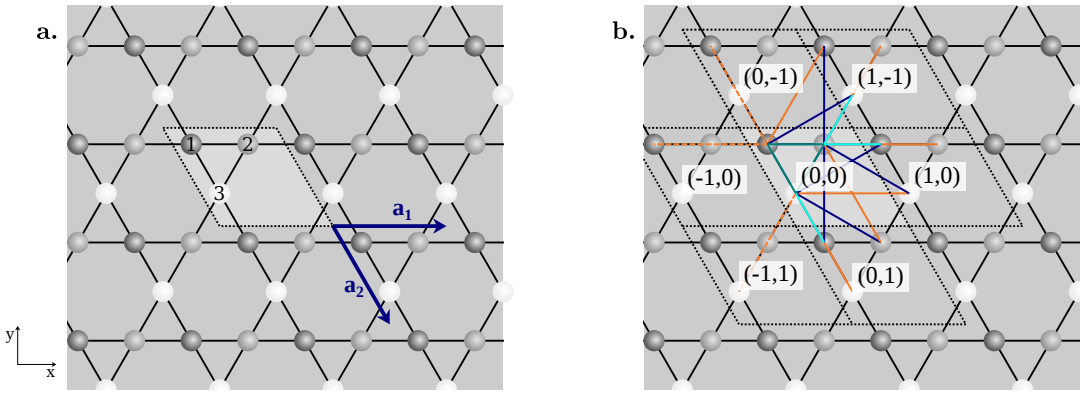

Figure 7.39: Illustration of the kagome lattice. **a.** We show the unit vectors and corresponding smallest unit cell that contains three sites. The three inequivalent sites are labeled for reference and shown in different shades of gray. **b.** Nearest, next-nearest and next-next-nearest neighbor hoppings are shown along with the corresponding cell coordinates in terms of lattice vectors $\mathbf{a}_1, \mathbf{a}_2$. Nearest neighbor links are colored teal (hoppings within the cell) or turquoise (outside of the cell). Next-nearest neighbor links are colored dark blue and next-next-nearest neighbors orange. All of them connect different cells. Our parameters correspond to those links illustrated by solid lines with directions out of the cell or towards larger site index. Since some of them overlap with nearest neighbor links we show in these cases the opposite directions as dashed lines.

Assuming no symmetries at all, this construction leaves us with a total of 24 independent hoppings, out of which three are the real onsite potentials $\varepsilon_1, \varepsilon_2, \varepsilon_3$, six are nearest neighbors, six next-nearest neighbors and nine neighbors of third degree. Before we can run a calculation we have to determine which bands we want to take into account. For this purpose we take a look at the band structure for the simple case where $t = -1$ and $t' = t'' = 0$, i.e., only nearest neighbor hopping on the perfect kagome lattice. The resulting bands are easily computed numerically and shown in Fig. 7.40. Clearly, the model with three sites per unit cell gives three bands, which also increases the number of possible gaps by one, w.r.t. to the honeycomb lattice, to two. Hence, we have another degree of freedom, namely the gap that we want to look at, which basically controls the bands for which the Chern number is computed. This approach is comparable to fixing a specific chemical potential, however, by instead choosing to sum over a fixed number of bands we guarantee that the chemical potential lies within the respective band gap if it exists.

The bare model shown in Fig. 7.40 apparently does not have a band gap. There are three notable features in the band structure, namely the existence of a Dirac point at K (and K′) at energy $-t$, a van Hove singularity at M at energies $0, -2t$, and a flat band at energy $2t$. We choose here to investigate the possibility of opening a gap at the Dirac points, which corresponds to taking into account only the Chern number of the lowest band, and note that the other possible choice can be investigated analogously.

Given this preliminary information, we define again our general model by defining a generic reference point in (21+3)-dimensional feature space (21 complex and 3 real), denoted by a feature vector

$$\mathbf{x} = (\boldsymbol{\varepsilon}, \boldsymbol{t}, \boldsymbol{t}', \boldsymbol{t}''), \tag{7.138}$$

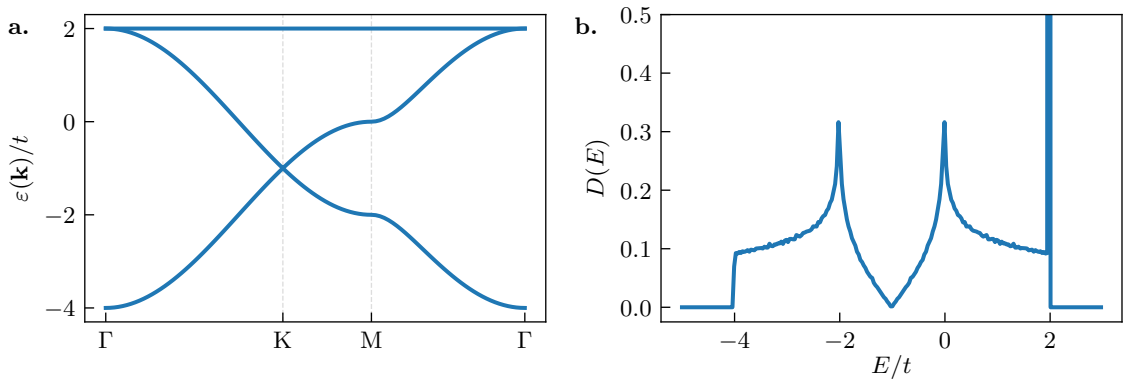

Figure 7.40: a. Band structure of the perfect kagome model with only nearest neighbor hopping $t < 0$ and no local potential, i.e., $\varepsilon_i = 0$ for $i = 1, 2, 3$. We observe a Dirac point at the K point and van Hove singularities at $\varepsilon = 0, -2t$, respectively. In addition, there exists a flat band at $\varepsilon = 2t$. The Fermi level at half filling, which corresponds here to $E_\mathrm{F} = 0.5$, is above the van Hove singularity located at 5/12 filling. **b.** The corresponding density of states. Compared to the graphene model, the additional flat band shifts half filling away from the Dirac point.

where

$$\begin{aligned} [\boldsymbol{\varepsilon}]_i &= \varepsilon_i = \frac{1}{4}, & i &= 1 \dots 3, \\ [\boldsymbol{t}]_i &= -1, & i &= 1 \dots 6, \\ [\boldsymbol{t}']_i &= -\frac{1}{\sqrt{3}}, & i &= 1 \dots 6, \\ [\boldsymbol{t}'']_i &= -\frac{1}{2}, & i &= 1 \dots 9, \end{aligned} \tag{7.139}$$

which are motivated as before by $t \sim 1/d$, where d is the length of the corresponding hopping link, and the sign is chosen negative by convention. The local energies are chosen finite such that the program produces a finite sample domain of width $\alpha/4$, where $\alpha = 1.5$ is a control parameter. In order to fix the zero of the energy scale, we enable sampling only for parameters $x_{i\neq 0}$, i.e., $x_0 = 0.25$ remains fixed. Throughout this entire discussion we will work again in arbitrary units of energy.

We now generate a data set of size $n_\mathrm{samples} = 10^7$ using the uniform distribution of Eq. 7.24 for our model, and compute the marginal distributions and importance scores for all real features (Re, Im, $|\cdot|$, $\varphi(\cdot)$). We find considerable numbers of topological samples with Chern numbers 1 and -1, respectively, each accounting for around 22% of the total number of samples, while $C = 2, -2$ account for $\approx 1\%$. In the following, we focus entirely on $C = 1$. We detect rather low importances throughout for this very general model, the largest values are found for phases and real parts of nearest and next-nearest neighbor hoppings. The importance scores of 3rd-nearest neighbors are very small indicating that these hopping parameters are not relevant for the generation of a topological phase.

In Fig. 7.41, we show the corresponding distributions for different Chern labels 0, 1 and -1, where we selected two representatives for each class of parameter ε, t, t', t''. Note that $\varepsilon_1 = x_0 = 0.25$ (no deviation from the reference point). Apparently, $\mathrm{Re}(\varepsilon_{2,3})$ and $\varphi(t''_{1,2})$ do not differentiate topological from trivial phases at all, as we have already learned from the importance scores (not shown). The interesting information is thus contained in t, t', of which there are six each. In this very general model we are obviously dealing with a very low contrast due to the large number of degrees of freedom. In order to obtain a simple model, we therefore have to extract constraints on the parameters first. However, there is already a rather clear distinction between the trivial

and non-trivial phases and also between the two distinct Chern labels. By comparing the distributions of t_1, t_2 and t_1', t_2', respectively, for labels $y = 1, -1$, we find that there are different classes of hoppings that can be differentiated by the sign of the phase, i.e., the distributions are the mirror images of one another.

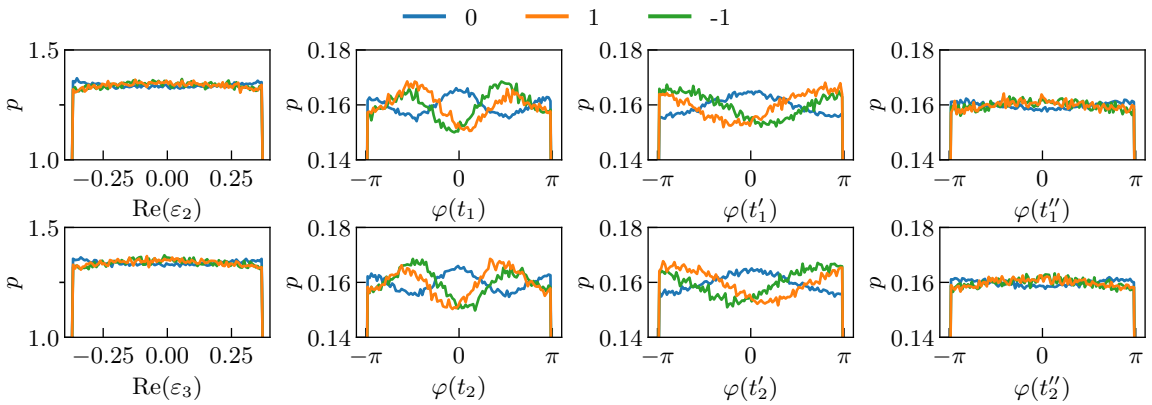

Figure 7.41: Marginal distributions of a selection of features (reference point subtracted) for the most general kagome model with a minimal unit cell. We show two different features for each class. Re($\varepsilon_{2,3}$) and $\varphi(t''_{1,2})$ both show no discernible contrast and can therefore be removed from the model. The topological information is mostly contained in t and t', of which the phase is the most important real descriptor. Here, all labels $y = 0, 1, -1$ are discriminated by different distributions. Note that p_0 is always the same for different $t_i^{(\prime)}$ from the same class, while some p_1 and p_{-1} are mirrored (here, we show one mirrored example).

We now use this information to group hopping parameters from each class (nearest/next-nearest neighbors). Apparently, there is a subset of samples that are aware of two different types of hoppings in each class that differ by the sign of the phase or equivalently of the imaginary part. By choosing as a reference the phase value of the first hopping term $t_1^{(\prime)}$, we assign all hoppings with the same marginal distributions to a set $A_1^{(\prime)}$ and all of those with the mirrored marginals to another set $A_2^{(\prime)}$. By connecting this information back to a lattice picture, we arrive at an induced phase order that is shown in Fig. 7.42. Subfigure **a.** shows the nearest and next-nearest neighbor hoppings that we take into account, where we now also explicitly indicate the direction of the associated process. Then, each link belonging to the set $A_2^{(\prime)}$ is reversed, while links in $A_1^{(\prime)}$ remain the same, since they are equivalent to the reference link. In Fig. 7.42**b**, we then show on the left the bond-phase order that is obtained by this procedure. For nearest and next-nearest neighbors we have apparently obtained a well-defined chirality, i.e., all hopping phases are aligned in a way that the arrows point along the direction of the path around the region that encloses the smallest unit. The nearest neighbor terms therefore point in the same direction when moving around a triangle, while next-nearest neighbors can be represented by a hexagon with unidirectional phase. This image immediately reminds us of the phase pattern inherent to the Haldane model, where the same order appears. In contrast to the honeycomb lattice, though, there is no distinction between two sets of next-nearest neighbor links, since they all represent hopping between different sublattices. Although we draw a particular direction of the arrows in Fig. 7.42**b**, this is only indicative of the relations between hopping phases and not the overall sign of this phase. The order we obtain apparently respects inversion symmetry at each lattice point. This is only one possible solution, and since the signal in the marginal distributions is rather weak we can assume that there exists also an opposite solution, which we would construct by explicitly breaking inversion symmetry at each point. This configuration

can be represented by a set of marginals that are invariant under mirror at $\varphi(t) = 0$. Clearly, an asymmetric function can always be represented in terms of a superposition of a symmetric and asymmetric part, which is why we generally obtain two possibilities here. We show also this configuration in Fig. 7.42**b** (right).

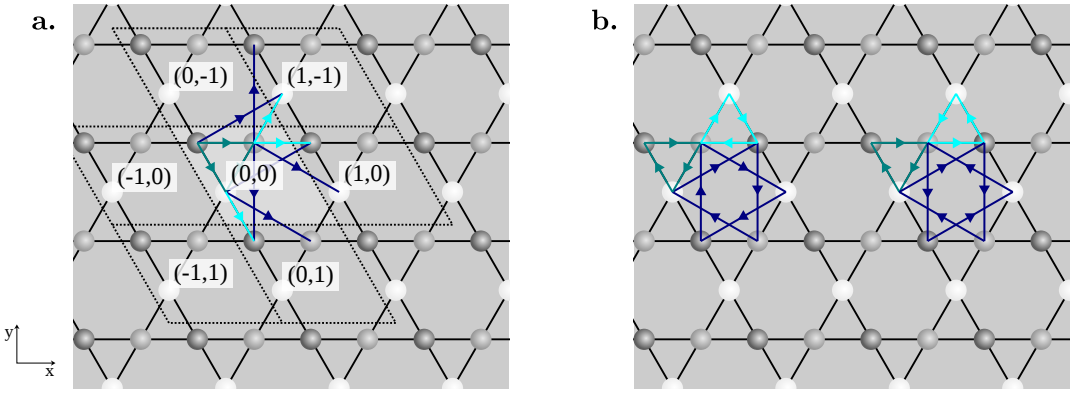

Figure 7.42: Phase order induced by topology. **a.** The original links that we take into account are shown, for clarity this time only for nearest and next-nearest neighbors. We add arrows denoting the direction of each link. By convention all links point outside of the (0, 0) cell or from smaller to larger site index. **b.** We show the optimized bond choices that are obtained from the information about the symmetry of the marginal distributions of Fig. 7.41. For the inversion symmeric case on the left nearest-neighbor vectors wind around the triangle obtained by connecting the three sites in the unit cell. For next-nearest neighbors we find the same. In particular, the combination of both triangles forms a Haldane-like hexagon with a unidirectional phase winding. Note that the global direction of arrows is arbitrary. The opposite configuration shown on the right breaks inversion symmetry at every point.

With this symmetry that is motivated by the observed preference in the topological data set we now construct a more streamlined model. Due to their low importance scores, we remove in this reduced model the third-nearest neighbor hoppings entirely and set all onsite terms to be the same. The remaining set of free parameters is composed of six nearest and six next-nearest neighbor links that we subdivide into three sets each depending on which sites they connect:

$$\mathbf{x} = (t_1, t_2, t_3, t_1', t_2', t_3'), \tag{7.140}$$

with

$$\begin{aligned} t_{12} = t_{21} = t_1, \quad t_{13} = t_{31} = t_2, \quad t_{23} = t_{32} = t_3, \\ t_{13}' = t_{31}' = t_1', \quad t_{23}' = t_{32}' = t_2', \quad t_{12}' = t_{21}' = t_3'. \end{aligned}$$

Assuming equal values within these sets we enforce inversion symmetry locally and arrive at a model with six independent complex parameters, which realizes a system with 180° rotational symmetry. While the configuration in Fig. 7.42**b** could, in principle, also host a six-fold rotational symmetry, we choose here not to impose this for now so that we do not oversimplify the model.

The resulting data set obtained in this (6+0)-dimensional feature space (12 real dimensions) is astonishing. As shown in Fig. 7.43, almost all data points belong to the two topological classes 1 and −1, which account for ≈98%, while only ≈1% of samples are trivial. This represents a complete change compared to the initial model, where trivial samples had a weight of about 50%, and proves that we are on the right track to engineer a topological model. Moreover, most samples are insulating as follows from the presence of a finite band gap at the K point (we check this in the whole Brillouin zone). If we relax the requirement of the existence of a finite band gap and look instead at the separation of the two lowest bands, we obtain the striking result that >99% of samples have separated bands, and therefore, have a well-defined topological label.

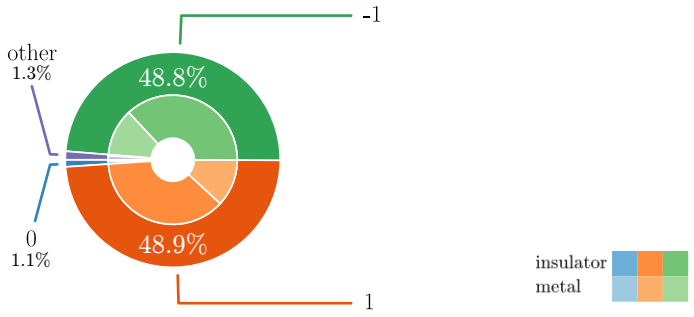

Figure 7.43: Composition of the data set for the inversion-symmetric six-parameter model of Eq. 7.140. Almost all samples belong to the 1 and -1 topological classes as is revealed by the respective labels. Only roughly 1% of samples are trivial, a similar number belong to other topological classes such as $2, -2$. The two signs of topological labels appear symmetrically, which is an indication that the reference point lies centered between the two phase regions. The vast majority of samples are insulating and we confirmed that >99% of samples feature separated bands.

This result has a couple of implications. Clearly, the model is very likely to describe a topological phase even with the completely general reference point that we chose. This is good, since it indicates that non-trivial topological properties should be ubiquitous in kagome materials that can be described by this model. On the other hand we still have six (complex) parameters left that describe nearest and next-nearest neighbor hoppings. As such the model is still a bit cumbersome. Due to the large number of topological samples, however, we are faced with the reality that almost every perturbation that we add onto the reference point generates a topological phase. As a consequence, the data that describes the topological phase is rather unstructured as follows from the absence of structure in the complete data set. We can therefore not expect to find correlations between our features. What we do find, however, is a more refined signal in the data that is shown in Fig. 7.44. Apparently, there is a clear indication that the phases associated with the hopping parameters are correlated with the topological classification. In particular, in Fig. 7.44**a**, we find that samples with Chern number 1 favor a positive phase of t_1, while the -1 topological insulator favors a negative phase, and vice versa for t_1'. With the convention that we chose for the parameters, cf. Fig. 7.42**b** (left), this means that the 1 (-1) phase has (counter-) clockwise phase winding around unit triangles. The next-nearest neighbor hoppings shown in Fig. 7.44**b** show the opposite preference, i.e., negative and positive phases for $1, -1$ classes, respectively, which result in counterclockwise (clockwise) winding. We make one more attempt at investigating the correlations between parameters and compute the Pearson correlation coefficient of Eq. 7.107 for the parameters t_1, t_2 for the complex features directly. With

$$\mathrm{Cov}[x, y] = \mathrm{E}[(x - \mathrm{E}[x])^*(y - \mathrm{E}[y])] \tag{7.141}$$

for general complex variables $x, y \in \mathbb{C}$, the definition of the correlation coefficient easily generalizes to

$$r(x, y) = \frac{\mathrm{Cov}[\mathrm{Re}\,[x]\,, \mathrm{Re}\,[y]] + \mathrm{Cov}[\mathrm{Im}\,[x]\,, \mathrm{Im}\,[y]] + i(\mathrm{Cov}[\mathrm{Re}\,[x]\,, \mathrm{Im}\,[y]] - \mathrm{Cov}[\mathrm{Im}\,[x]\,, \mathrm{Re}\,[y]])}{\sqrt{\langle |x - \langle x\rangle|^2\rangle \langle |y - \langle y\rangle|^2\rangle}}, \tag{7.142}$$

where $\langle\cdot\rangle$ is shorthand for $\mathrm{E}[\cdot]$. We have still $|r(x, y)| \leq 1$, $-1 \leq \mathrm{Re}\,[r(x, y)] \leq 1$, $-1 \leq \mathrm{Im}\,[r(x, y)] \leq 1$ and $r(x, y) = 0$ for independent variables, since independent x, y implies independent $\mathrm{Re}\,[x]\,, \mathrm{Re}\,[y]\,, \mathrm{Im}\,[x]\,, \mathrm{Im}\,[y]$. The real part of $r(x, y)$ is therefore a measure of correlations between real-real or imaginary-imaginary, while the imaginary part measures cross correlations between the two. In our case we find very small correlations $\mathcal{O}(10^{-2})$ for the real part and even

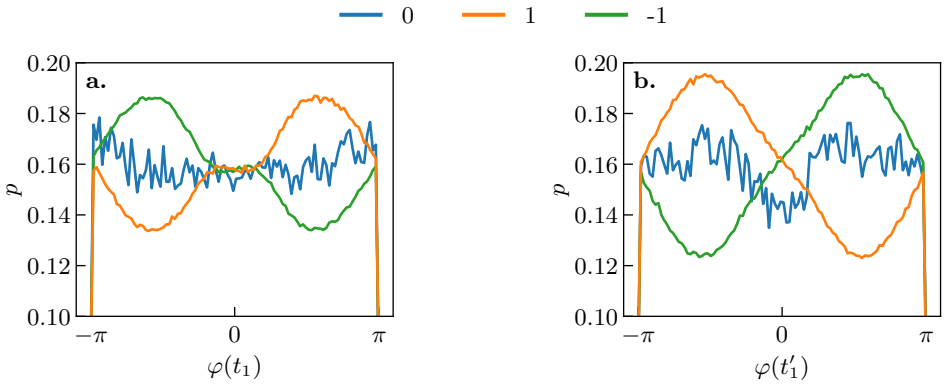

Figure 7.44: Marginal distributions obtained from the symmetric six-parameter model. **a.** Phase of the nearest neighbor hopping. The non-zero topological indices 1, -1 favor different signs, while the trivial class is undiscriminating. **b.** Phase of the next-nearest neighbor hopping. Also here, the phase carries information about the topological class label, however, the sign is reversed. Note that the jaggedness of the distribution for the 0 class is a consequence of the low sample count, cf. Fig. 7.43.

smaller $\mathcal{O}(10^{-4})$ for the imaginary part. Compared to the full data set that is uncorrelated by construction, and therefore serves as a base value to judge the noise that results from the finite data set, we find that the cross correlations are on the same order as the noise level, while the direct correlations of real and imaginary parts are increased by two orders of magnitude for all three topological classes. Digging a bit deeper into this by computing separate correlations between real features, we find that these correlations occur mainly between imaginary parts. An additional computation of the three-variable correlators from Eq. 7.115 for the imaginary parts yields no further information, presumably due to the correlations with the t'_i variables. We repeat this analysis also for the pair t'_1, t'_2 and find also there very small correlations that do not lend themselves to justify a solid conclusion. The general analysis is therefore inconclusive due to the correlations between all six parameters that cannot be unraveled realistically. A more detailed analysis of the correlations therefore has to be performed in a reduced model.

From the scarce data for Chern labels $2, -2$ that is not shown here due to the strong noise we can additionally construct a characteristic model for that case with $|t'| > |t|$, where the sign of the Chern number is determined by the sign of $\mathrm{Im}[t]$, which is motivated by corresponding shifts in the marginal distributions that show significant contrast despite the noise.

We also generated a data set for the second possible phase order illustrated in Fig. 7.42**b** (right), where the hopping links break inversion symmetry at every lattice site. For this case, we find the opposite distribution of labels: ≈94% trivial and the remaining 6% $l = 1, -1$, however, almost all samples are metallic. Therefore, the inversion-symmetric phase winding is the preferred topological configuration.

At this point we have gained a basic understanding of what drives the topological phase, however, the complexity of the model poses a severe challenge to our data-analytic tools, since there are simply too many parameters that are seemingly correlated with one another. There are now two avenues along which we proceed: 1. reduce the model to two parameters (t, t') and try to optimize phase patterns, 2. look at a nearest-neighbor model only, but relax symmetry constraints.

We start with the two-parameter model that is rather simple as it has only four real degrees of freedom. We use essentially the model from Eq. 7.140, where we set $t_i = t, t'_i = t' \ \forall i$ and enforce the inversion-symmetric phase pattern from Fig. 7.42**b**. The resulting data set is again peculiar in that topologically non-trivial phases are again much more abundant than the trivial

phase. The composition of the data set in terms of the numbers of samples for each Chern label that we observed is shown in Fig. 7.45. Here, we distinguish two different classes of data points based on the characteristics of the band gap. We compute both the band gap E_g and the band separation ΔE, cf. Eq. 7.123, and use these markers to classify the data as *insulators* and general samples that can be either insulators or metals, i.e., *all*, where we assume that the bands are separated by finite ΔE. In both cases, the Chern number for the lowest band, i.e., the Chern number consistent with a chemical potential at the Dirac point in the reference model, is well-defined. The label "insulators" is more restrictive, and therefore, all insulating samples are also included in the statistics for separated bands.

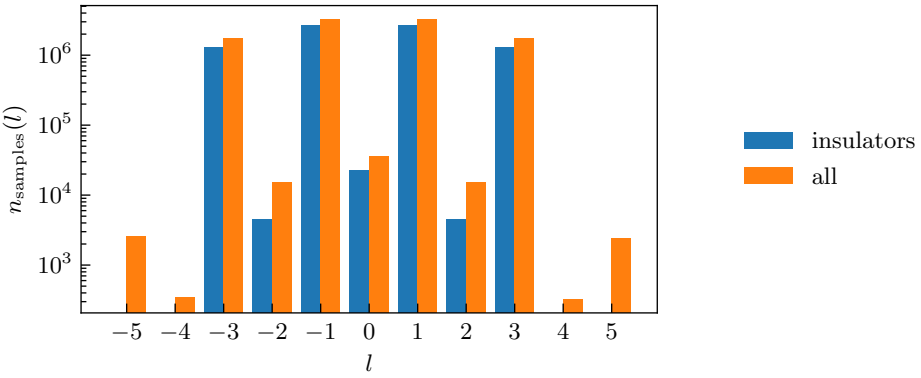

Figure 7.45: Composition of the data set for the simplified two-parameter model with t, t' [Eq. 7.140 with $t_i = t, t_i' = t'$] and the inversion-symmetric phase pattern of Fig. 7.42**b**. We distinguish insulating phases with finite band gap and general phases that could also be metallic. Without restrictions on the band gap we find Chern labels l in the range $-5 \ldots 5$ with odd numbers notably more abundant and decreasing numbers towards larger l. Especially samples with $|l| > 3$ are much rarer. For insulators, we find roughly the same, although here, $|l| > 3$ is not observed at all.

In total, we find that with a threshold of $\Delta E > 0.05$, >99% of our samples have separated bands and most of them are also insulating with the exception of samples with Chern label $|l| > 3$ which are all metallic. Overall, we find Chern labels $l = -5, \ldots, 5$ with $|l| > 3$ being much less frequent than the rest. The data shown in Fig. 7.45 reveals an interesting pattern: the sample counts for even l are strongly suppressed compared to those for odd l. Given that the trivial phase with $l = 0$ counts towards even l, this is an indication that the low number of trivial samples is a consequence of a more general property of this model that extends also to non-trivial topological phases. Phases with $|l| = 1, 3$ contribute similarly and are almost all insulating. The fraction of insulators for $|l| = 2$ is slightly lower.

In order to understand more about the properties of the individual phases in terms of their realization as a function of the parameters of our model, we turn our attention to the marginal distributions. Since we are interested in all phases, we have to compute the importance score for all labels and find that all real features are important for at least one label, which requires us to investigate all of them. Due to the sheer number of different labels, we split the data up into two groups depending on the value of $|l|$.

All marginals are shown in Fig. 7.46, where we plot not as before only the perturbation to the reference point, but the actual hopping parameters. There is already a lot of information accessible through the marginal distributions alone and we find that while the trivial (0) and 1, -1 classes are both characterized by large $|t|$ and small t', $l = 0$ favors strong negative $\mathrm{Re}\,[t]$, while $l = \pm 1$ favors positive/negative $\varphi(t)$, and negative/positive $\varphi(t')$. We investigate this further by computing the correlation coefficient $r(\mathrm{Im}\,[t], \mathrm{Im}\,[t'])$ for the labels 0, 1, -1 and find $r_0 \approx 0.8$, $r_{\pm 1} \approx 0.2$, i.e., trivial samples have strong positive correlations between the imaginary

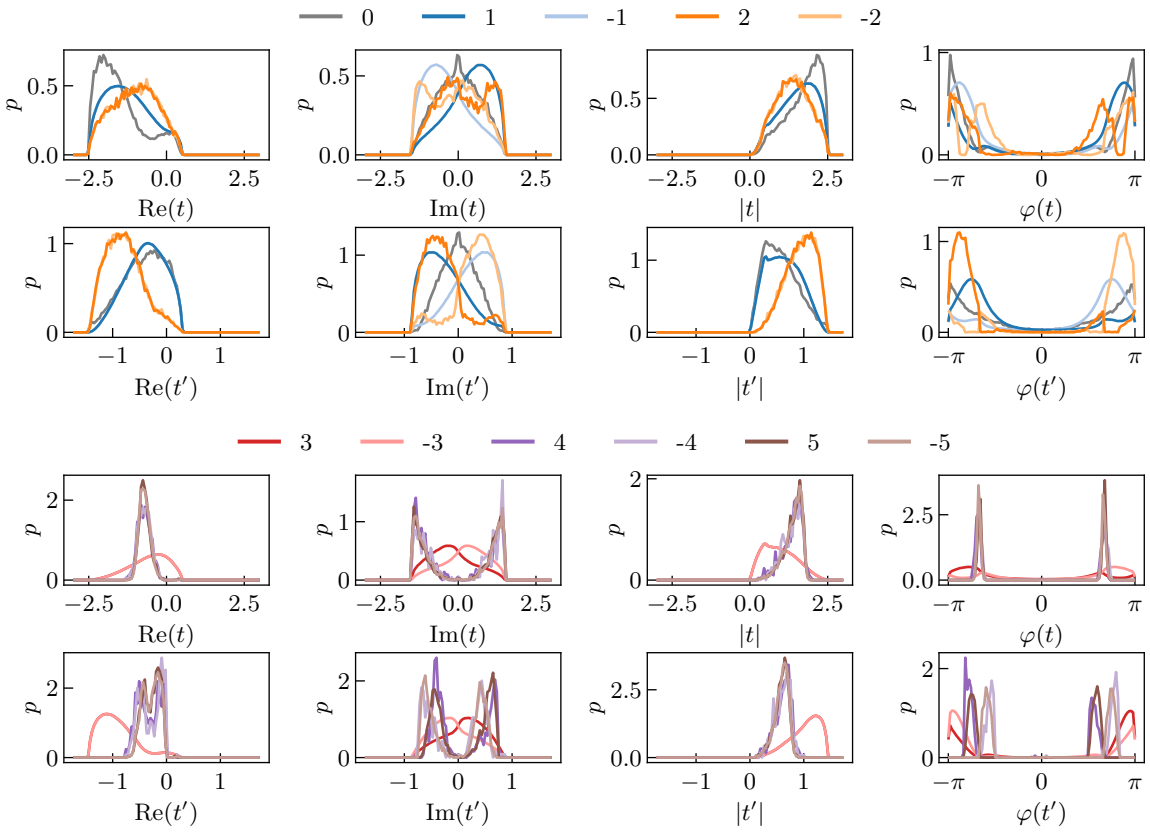

Figure 7.46: Marginal distributions for all real parameters for the two-parameter inversion symmetric model. In order to improve legibility we split the data into two parts. The topological phases with $l = 1, -1$ are rather difficult to distinguish from $l = 0$, since the discrimination is comparatively low. In contrast to $l = 0$ the $l = 1$ phase favors negative (positive) imaginary part of t (t') and vice versa for $l = -1$. $l = 2, -2$ are mainly distinguished from other phases by a strong negative real part of t' that is also reflected in larger $|t'|$ and positive (negative) $\mathrm{Im}(t')$ for $l = 2$ (-2). The two labels $3, -3$ show similar preference for opposite signs of $\mathrm{Im}(t^{(\prime)})$ compared to $1, -1$, where the sign is generally reversed. The characteristic property is a small real part of t and large negative real part of t' together with $|t| < |t'|$. Chern labels $|l| > 3$ require very specific configurations, i.e. $\varphi(t) \approx \pm\frac{2\pi}{3}$, and the sign of l is apparently correlated with specific values of $\varphi(t')$. The presence of multiple peaks for the same distribution is an indication of strong correlations.

parts which implies equal signs as opposed to non-trivial samples that are only weakly correlated and favor opposite signs. The two classes 2,-2 are mainly distinguished from the previous three by the comparatively stronger modulus of the next-nearest neighbor hopping t', that is caused primarily by strong negative $\mathrm{Re}\,[t']$. The sign of the Chern label is determined almost entirely by the sign of $\mathrm{Im}\,[t']$, where the statistical distance between the two marginals is especially large. In addition, we find correlations $r(\mathrm{Im}\,[t]\,, \mathrm{Im}\,[t']) \approx 0.46$ that imply that, e.g., for $l = 2$ unfavorable positive $\mathrm{Im}\,[t']$ can be compensated by positive $\mathrm{Im}\,[t]$ and vice versa for $l = -2$. Regarding the $l = 3, -3$ classes we find that they in many ways appear to be the opposite of $l = 1, -1$. For instance, moderate $\mathrm{Re}\,[t]$ and large $\mathrm{Re}\,[t'] < 0$ is characteristic here, in addition to a reversed preferred sign of the corresponding imaginary parts.

We now turn our attention towards the most interesting classes $|l| > 3$ that are rare, metallic and do not seem to have an insulating analog. Immediately, it becomes clear from the sharpness of the marginal distributions shown in Fig. 7.46 that these phases require a rather specific set of parameters. Some of this information can be read off immediately, and we have $\mathrm{Re}\,[t]\,, \mathrm{Re}\,[t'] < 0$.

In addition, $\varphi(t) \approx \pm\frac{2\pi}{3}$, while $\varphi(t') = \frac{2\pi}{3} \mp \eta$ with a small $\eta > 0$ and $\mp$ for $l = \pm 4, \pm 5$. The occurrence of multiple peaks in these distributions indicates that there are underlying correlations. We therefore compute the correlation coefficient and find perfect positive correlations $r(\varphi(t), \varphi(t')) \approx 1$ for all four classes. This leaves us in a predicament. Even after having exhausted all available data we cannot distinguish between labels 4 (-4) and 5 (-5). This is a clear warning sign that the computation of the Chern label might be unstable. We therefore inspect band structures of several samples and find that there seems to be a band crossing that was not detected during our earlier computation of the band separation. Even though the threshold of 0.05 was chosen based on the resolution of the k-grid in our computations ($\Delta k \approx 0.07$), these band crossings remained undetected, i.e., the slope is rather steep. We note that the requirement of distinguishability of any two phases constitutes a check that serves to identify wrongly labeled data and should always be taken into account in order to avoid false claims. After increasing the threshold, our machinery identifies the $l = \pm 4$ samples as metals with band crossings and only a small number of the samples from the ± 5 class remain. We verified numerically that the bands are separated, i.e., the band separation converges to a finite value with increasing momentum-resolution, and found that the computation of the Chern number is stable. Nevertheless, the bands approach each other at a variety of points in the Brillouin zone at closely avoided crossings. For a full degeneracy, the randomness in our data set is apparently too large.

What remains is to analyze these metallic phases a bit further. Although the Chern number does not necessarily have a meaning, apparently, the pairs $(4, 5)$ and $(-4, -5)$ each label the same class since their marginal distributions are the same across all degrees of freedom and there are no correlations present that differentiate between the two. What is left to investigate is if there is a physical argument for the sign change of the Chern label across $\varphi(t') = \frac{2\pi}{3}$ or if all labels correspond to the same metallic phase. Due to the very low sample count, this requires a more refined sampling approach to procure more relevant statistics. At the time of writing this issue could not yet be clarified.

We have now produced a model that hosts a number of topological phases by taking into account nearest and next-nearest neighbor hoppings. Now we turn our attention to the second approach that we introduced earlier, where we relax symmetry requirements and instead neglect the next-nearest neighbors. In particular, we use the nearest-neighbor model

$$\mathbf{x} = (t_1, t_2, t_3, t_4, t_5, t_6), \tag{7.143}$$

where all onsite terms have been set to 0 and all nearest neighbor terms are independent of one another, i.e., we impose no constraint on the individual hoppings. This means that also the phase order we discovered earlier [cf. Fig. 7.42] is not enforced, although all hoppings are chosen such that positive phases correspond to this order. Eq. 7.143 represents a model in a 6-dimensional complex space. We proceed again in the same way as before by generating a data set with $n_{\text{samples}} = 10^7$. This time around, we decide to stay in the framework of complex features that was introduced in Sec. 7.7.7 entirely, i.e., we do not make use of the mapping from Eq. 7.22. Our data set contains statistically significant numbers of samples categorized into three classes: $y = 0, 1, -1$ with $y = 0$ accounting for $\approx$70% and $y = \pm 1$ for $\approx$15% each. The majority of the samples of all classes are insulating.

In Fig. 7.47, we show the marginal distributions $\rho_l(t_i)$ for $l \in \{0, 1, -1\}$ and $x_i = t_1 = t$. Note that due to the rotational symmetry inherent to our unconstrained feature choice, the marginal distributions for different features t_i are necessarily equal. It is therefore sufficient to investigate one example. Since the distribution function of complex features takes two real arguments, we clearly cannot easily draw all distributions in the same panel unless we resort to 3D plots which makes comparisons more difficult. For this reason, the automation of this comparison in terms of the importance score is very useful even for small numbers of features. Nevertheless, the

differences between the three distributions are rather clear in this case and we find for the trivial phase a rather homogeneous distribution with a shallow peak around $t = 0$. The two non-trivial phases are essentially conjugate to each other, i.e., $\rho_{l=1}(t) \approx \rho_{l=-1}(t^*)$. Here, we find a clear minimum at $t = 0$ for both phases that reaches $\rho_{l=\pm1}(0) = 0$, indicating that in the non-trivial phase no hopping is equal to 0. Also here, the distributions are rather blurry and essentially all values other than 0 are allowed. However, the maxima located around $t \approx 1.5e^{\pm 3\pi/8}$ for $l = \pm 1$ are more pronounced than that in the distribution of $l = 0$. For comparison, we note that our uniform base distribution is $\rho_{\text{uniform}}(t) = 4/9\pi \approx 0.14$.

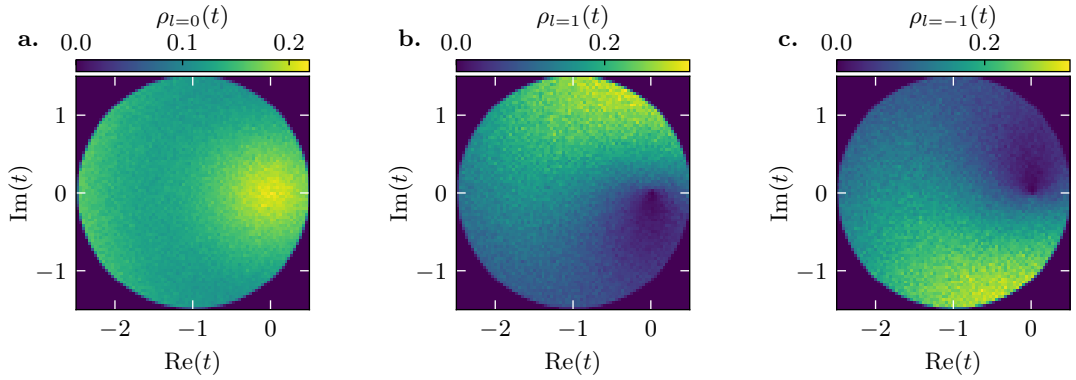

Figure 7.47: Marginal distributions obtained from the nearest-neighbor model of Eq. 7.143. Here, we plot the complex distribution function of Eq. 7.129 for a single $t_i = t$ only, since all marginal distributions are the same due to symmetry. In the three panels, we show $\rho_l(t)$ for labels **a.** $l = 0$, **b.** $l = 1$, **c.** $l = -1$. The reference point is located in the center of the circular region. We observe a notable difference between the three, in particular, the Hellinger distance between non-trivial and trivial distributions is $D_{\text{H}}(p_{1/-1}, p_0) \approx 0.2$. The distribution of the trivial phase has a maximum around the origin, i.e., $t = 0$, where the topological phases have a minimum. The 1 (-1) phase prefers positive (negative) imaginary part.

Since we still have six independent parameters that are highly correlated with one another, it is very difficult to measure the exact relationships. As described in Sec. 7.7.7, we try to ameliorate this constraint by introducing new features that allow us to look beyond the bare marginal distributions. We construct these as

$$x_{j>6} = \prod_{k=1}^{n} t_{i_k}, \quad n \in \{2, \ldots, 6\}, \quad i_k \in \{1, \ldots, 6\}, \tag{7.144}$$

where we choose random subsets of $\{t_i \mid i = 1, \ldots, 6\}$ without replacement. There are therefore $\binom{6}{n}$ possibilities for fixed n and in total

$$N = \sum_{n=2}^{6} \binom{6}{n} = (1+1)^6 - 7 = 57 \tag{7.145}$$

such choices. This number is small enough so that we can construct them all and compute the statistical distances, i.e., importance scores, between topological and trivial, but also between different topological phases in order to find the most descriptive features that reveal information that is not visible in the marginal distributions of the plain features.

In Fig. 7.48, we plot the importance score, here defined via the Hellinger distance [Eq. 7.69], for all 63 features, i.e., the original 6 plus all 57 engineered features. We chose the Hellinger distance here due to the property $0 \leq D_{\text{H}} \leq 1$ as opposed to $0 \leq D_{\text{B}} \leq \infty$. Since we are here only looking for features with large importance as opposed to earlier where we tried to filter out

unimportant features, the improved contrast of the Bhattacharyya distance is less relevant. We compute the importance for discrimination between topological and trivial for both topological classes $l = \pm 1$, the result, however, is identical.

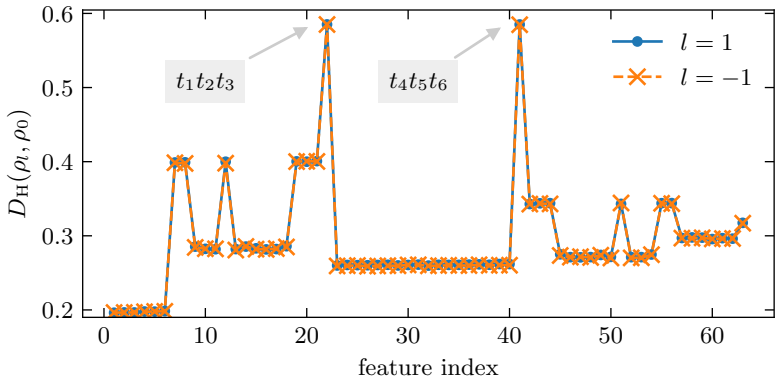

Figure 7.48: Statistical distances or importance scores $D_{\mathrm{H}}(\rho_l, \rho_0)$ for the discrimination of topological and trivial phase for all bare and engineered features. We show values for both topological classes $l = \pm 1$ and find that they are identical. The original features (index 1-6) have the lowest importance, i.e., feature engineering in this case always increases the importance score. Several artificial features stand out with large importance scores. Above all, two triple products, $t_1t_2t_3$ and $t_4t_5t_6$, reach almost three times the importance of the original features. Combinations of two and four features that minimize the mixing between $\{t_1, t_2, t_3\}$ and $\{t_4, t_5, t_6\}$ are also significantly more important than those that mix between the two sets.

The data shown in Fig. 7.48 reveals that all engineered features have a larger importance score than the original features, however, the variation among them is large. The vast majority display only a minor increase in importance, and therefore do not lend themselves as particularly powerful descriptors of the system—at least not more so than the original features. Notable exceptions are the products of pairs of features $\{t_1t_2, t_1t_3, t_2t_3, t_4t_5, t_4t_6, t_5t_6\}$ with an importance score of roughly 0.4 each, and products of quadruplets $\{t_1t_2t_3t_4, t_1t_2t_3t_5, t_1t_2t_3t_6, t_1t_4t_5t_6, t_2t_4t_5t_6, t_3t_4t_5t_6\}$ with an importance of $\approx$0.35. Dominating above all else are, however, the products of triplets $\{t_1t_2t_3, t_4t_5t_6\}$ with an importance score of almost 0.6 that is about three times as large as that of the bare features. We notice a pattern here: the important pairs are exactly those pairs taken from one of the subsets $\{t_1, t_2, t_3\}, \{t_4, t_5, t_6\}$, important quadruplets are those that take all features from one subset and add one from the other, i.e., minimal mixing. The overall most important are the triplets that are taken from either subset. This indicates that there is a separation between subsets and that hoppings from each subset satisfy a particular relationship depending on the topological class.

This can, of course, be investigated by analyzing the probability distributions of the engineered features more closely. We do this for the triple products in Fig. 7.49, where we observe a spectacular result: While the distribution for samples with $y = 0$ is approximately symmetric around 0, the distributions for $y = l$ ($l \in \pm 1$) are not. Instead, they are again conjugate to each another and, moreover, completely localized to the upper (lower) half of the complex plane for $l = 1$ (-1). This stunning result implies that the importance for discrimination between topological phases assumes the maximal possible value $D_{\mathrm{H}}(\rho_1, \rho_{-1}) = 1 = \max$ or $D_{\mathrm{B}}(\rho_1, \rho_{-1}) = \infty$, which is a clear sign that these artificial features contain the complete information about the sign of the Chern number. Note that all distributions decay rather rapidly away from 0 since $|t_i| < 1$ on average, and therefore, distributions of products of higher order decay increasingly fast. The localization to either half of the complex plane is best described in terms of the phase

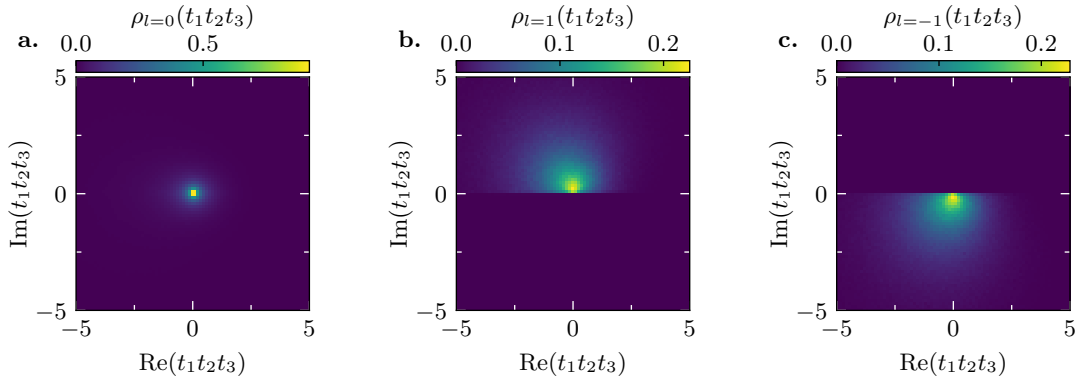

Figure 7.49: Marginal distributions of the engineered feature $x_{23} = t_1t_2t_3$ obtained from the nearest-neighbor model of Eq. 7.143. Due to symmetry, this is equal to $\rho(t_4t_5t_6)$. In the three panels, we show $\rho_l(t)$ for labels **a.** $l = 0$, **b.** $l = 1$, **c.** $l = -1$. For $l = 0$ the distribution is localized around 0, which indicates that $|t_i|$ are mostly smaller than 1. For the topological classes we find a less strong localization to a particular point, however, the distribution is restricted to the upper (lower) half plane for the 1 (-1) phase.

$\varphi(t_1t_2t_3) = \varphi(t_1) + \varphi(t_2) + \varphi(t_3)$, and by comparing with the marginal distributions we obtain

$$\varphi(t_1) + \varphi(t_2) + \varphi(t_3) \left\{ \begin{array}{c} > \\ < \end{array} \right\} 0 \quad \text{for} \left\{ \begin{array}{c} y = 1 \\ y = -1 \end{array} \right\} \tag{7.146}$$

and

$$\varphi(t_4) + \varphi(t_5) + \varphi(t_6) \left\{ \begin{array}{c} > \\ < \end{array} \right\} 0 \quad \text{for} \left\{ \begin{array}{c} y = 1 \\ y = -1 \end{array} \right\}. \tag{7.147}$$

In order to understand this, we note that the hoppings t_1, t_2, t_3 correspond to those links drawn in teal in Fig. 7.42**a** and t_4, t_5, t_6 to those colored turquoise. Clearly, the requirements of Eqs. 7.146, 7.147 are satisfied if the combined phases of the three hoppings around the two individual triangles are each positive (negative) for the topological class $l = 1$ (-1). We had already learned about particular phase patterns that favor a topological phase before. This analysis, however, revealed a deeper understanding of the relationship between the Chern number and tight-binding parameters. Note that during this entire discussion we did not make any assumptions that would require prior knowledge about topology. The result has therefore been obtained entirely through our data-analytical methods and demonstrates the power of the general methodology. Moreover, we have not only generated knowledge from data, but also provided a selection of powerful features that could be employed by other machine learning approaches to improve their performance.

What is not entirely understood yet is how exactly one discriminates between the trivial and non-trivial phases given only the tight-binding parameters. We have seen that the engineered features provide a significant improvement in contrast, however, the separation between trivial and non-trivial is not perfect. From the inverse of the requirements of Eqs. 7.146, 7.147 we can extract a property of the trivial phase, i.e., the cases where the phases wind in opposite directions around the triangles are topologically trivial. This is supported by the significant negative correlations of the imaginary parts of the product features $t_1t_2t_3$ and $t_4t_5t_6$ of about $r \approx -0.28$, however, the value $r > -1$ also clearly means that this is not the only way to construct a trivial phase. Thus, if it is not only the phase that yields the separation, then the magnitude must play a role, too. And indeed, plotting the real PDF $p_l(|t_1t_2t_3|)$ shows contrast between topological and trivial phases—only so much, though, to give a hint that $|t_1t_2t_3|$ has a tendency to be smaller in the trivial phase. Here, we note that while speaking about magnitudes

we also have to be aware of the fact that in our data set there is still a scale degree of freedom, i.e., a scale transformation $\mathbf{x} \mapsto a\mathbf{x}$ for $a \in \mathbb{R}^+$ leaves all physical properties of the system entirely invariant, and therefore constitutes a redundancy (noise) in our data. A classification in terms of magnitude without first fixing this scale degree of freedom is therefore not well-defined.

We fix the scale by setting $|t_1| = 1$, which is achieved after the fact by simply renormalizing the data, and find that this increases the contrast between topological and trivial distributions, however, not to an extent that sufficiently describes the mechanism behind the distinction between topological and trivial phase. In order to investigate this in more detail, we once again simplify the system. Since the role of the phases has already been settled, we now impose a symmetry between the two sets of nearest neighbor hoppings that are defined in Fig. 7.42**a** such that inversion symmetry is preserved w.r.t. inversion centers at every lattice site. In particular, this means that $t_1 = t_4, t_2 = t_5, t_3 = t_6$, and therefore, $\mathbf{x} = (t_1, t_2, t_3)$. Thus, the requirements for topological phases given by Eqs. 7.146, 7.147 reduce to just Eq. 7.146. In addition, we fix the scale degree of freedom from the start this time around and thus require $|t_1| = 1$. There is one more unnecessary complication in our model. Remember that only the total phase around the triangle counts, which is at present described by the sum of three numbers. We can fix a specific gauge by choosing arbitrarily $t_1 = 1$, i.e., $\varphi(t_1) = 0$. With this choice, our streamlined model has complex dimension 2 and is described by

$$\mathbf{x} = (t_1 = 1, t_2, t_3). \tag{7.148}$$

We proceed with the generation of a new data set and obtain a striking 98% topological phases indicating that our model is already tailored to producing topological phases due to both the imposed phase order of the hoppings and the symmetries. The remaining ≈1.2% of trivial samples are too few in number to provide good estimates for the PDFs and we observe jagged lines that nevertheless show a considerable but insufficient amount of contrast w.r.t. the topological samples. We consider again engineered features and this time there is really only one choice if we use the same strategy: $t_2 t_3$. We investigate the distributions of $t_2 t_3$ and find that the combined phase around a triangle that is now given by $\varphi(t_2 t_3) = \varphi(t_2) + \varphi(t_3)$ necessarily has to vanish (up to some numerical accuracy) in order to obtain a trivial phase as a result of the imposed symmetries. This is shown explicitly in Fig. 7.50, where we compare the distributions of the artificially engineered feature $t_2 t_3$ evaluated with samples belonging to the $y = 0$ and $y \neq 0$ classes, respectively. Note that we do not differentiate the two distinct topological phases $y = \pm 1$ here, since we have already identified the characteristic difference between those phases and are now only interested in discriminating them from the trivial phase.

In Fig. 7.50**a**, we show the distribution for the trivial phase and observe the remarkable feature that possible values of $t_2 t_3$ are entirely confined to the real axis. This is, of course, never exactly satisfied numerically, since it is highly unlikely that our sampling procedure produces a completely real number. Nevertheless, the preceding statement is found true up to a remarkable accuracy of $\approx 0.04 t_1$. In Fig. 7.50**b**, we show for comparison also the distribution of the same feature for the combined non-trivial classes. Here, a plethora of complex values are possible, reflecting the fact that our model is streamlined to produce topological samples. In both cases, small negative real values are strongly suppressed. This is not a feature of the individual topological phases but rather an artifact of the sampling method, since the reference point was chosen with negative signs for all t_i, cf. Eq. 7.137. Therefore, the product of two hoppings is generally positive, and negative values require both perturbations δt_i to the initial t_i to be rather large, which together with the phase restriction is very unlikely. Finally, we note that the Hellinger distance between the two distributions shown evaluates to $D_{\mathrm{H}}(\rho_{l=0}, \rho_{l\neq 0}) \approx 0.93$, which means that our engineered feature is highly descriptive of the trivial phase.

We have now gained a lot of information about the relationships between the configuration of the system in terms of tight-binding parameters and the topological phase. Before we assemble

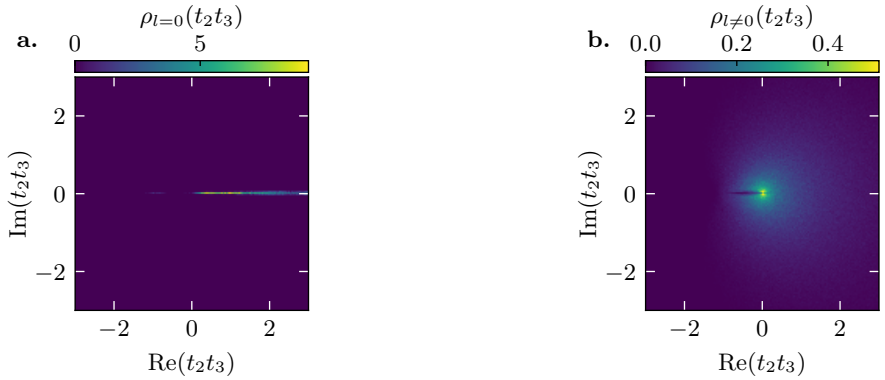

Figure 7.50: Marginal distributions of the complex artificial feature t_2t_3 for labels $l = 0$ and $l \neq 0$. Since we are only interested in the trivial phase, we compare it here against all non-trivial phases (here, only $1, -1$) combined. **a.** The distribution of the artificial feature for the trivial phase is confined entirely to the real axis (within reasonable accuracy, since in manageable time we will never observe a truly real sample). We note here that the product of t_2, t_3 is mostly positive. This is a consequence of the negative reference point that places a considerable bias on the probable outcomes of the product. **b.** The same distribution function evaluated for non-trivial samples. Here, also complex values are possible. The suppression of small negative real values is an artifact of the sampling method. The Hellinger distance between the two distributions is ≈0.93.

this information into a phase diagram type overview there is only one case left to discuss. Namely, we have established that the conditions of Eqs. 7.146,7.147 represent necessary conditions that must be fulfilled in order to obtain a topological phase. They are, however, not sufficient conditions, i.e., we are still lacking an understanding of when exactly a trivial phase occurs if Eqs. 7.146,7.147 are satisfied. In order to answer this question, we return to our earlier data set for the model of Eq. 7.143, where this case is generally allowed to occur. This time, we have to engineer features that are most descriptive for the distinction $l = 1$ vs. $l \neq 1$. Testing our earlier product features of Eq. 7.144 against the new importance score $D_{\mathrm{H}}(\rho_{l=0}, \rho_{l\neq0})$ we obtain only very weak improvements w.r.t. the bare hoppings. We try instead also the other obvious schemes for constructing new features

$$x_{k>6} = \begin{cases} \sum_{k=1}^{n} t_{i_k} \\ \sum_{k=1}^{n} |t_{i_k}| \end{cases} \qquad n \in \{2, \ldots, 6\}, \quad i_k \in \{1, \ldots, 6\}, \tag{7.149}$$

however, none of these features produce an importance score beyond ≈ 0.1. At this point it is clear that the actual dependence of the phase boundary between the trivial and either non-trivial phase is more complicated and we would be better served by a more complicated model that could, e.g., be given by a polynomial of the original features, the artificial choices above and the product features from Eq. 7.144. This more general approach would, however, completely defeat the entire purpose of this analysis as it tries to extract details that are intrinsically difficult to understand. Instead, we decide to get a more qualitative idea by investigating correlations.

Restricting the set of trivial samples to those that satisfy Eqs. 7.146,7.147 we compute the correlation coefficient $r(t_1t_2t_3, t_4t_5t_6)$, i.e., we measure the correlations between the most descriptive features. With to the restriction to only a part of the data set we make sure that we do not observe the same information that we already know, i.e., the conditions of Eqs. 7.146,7.147. Indeed, we find two very different results for the two cases: with $r_{l=0}(|t_1t_2t_3|, |t_4t_5t_6|) \approx -0.18$ we see negative correlations between the magnitudes. In contrast to this, we find for the non-trivial samples $r_{l\neq0}(|t_1t_2t_3|, |t_4t_5t_6|) \approx 0.34$, i.e., positive correlations for the topological classes. This indicates that a strong imbalance of magnitudes of the hopping parameters between the two

sets $\{t_1, t_2, t_3\}, \{t_4, t_5, t_6\}$ destroys the topological phase. The same information is also accessible through $r_{l=0}(|t_1|+|t_2|+|t_3|, |t_4|+|t_5|+|t_6|) \approx -0.22$ and $r_{l\neq 0}(|t_1|+|t_2|+|t_3|, |t_4|+|t_5|+|t_6|) \approx 0.44$. With this last piece of information, we now have a pretty good understanding of where in the original (21+3)-dimensional feature space we can expect to find which phases.

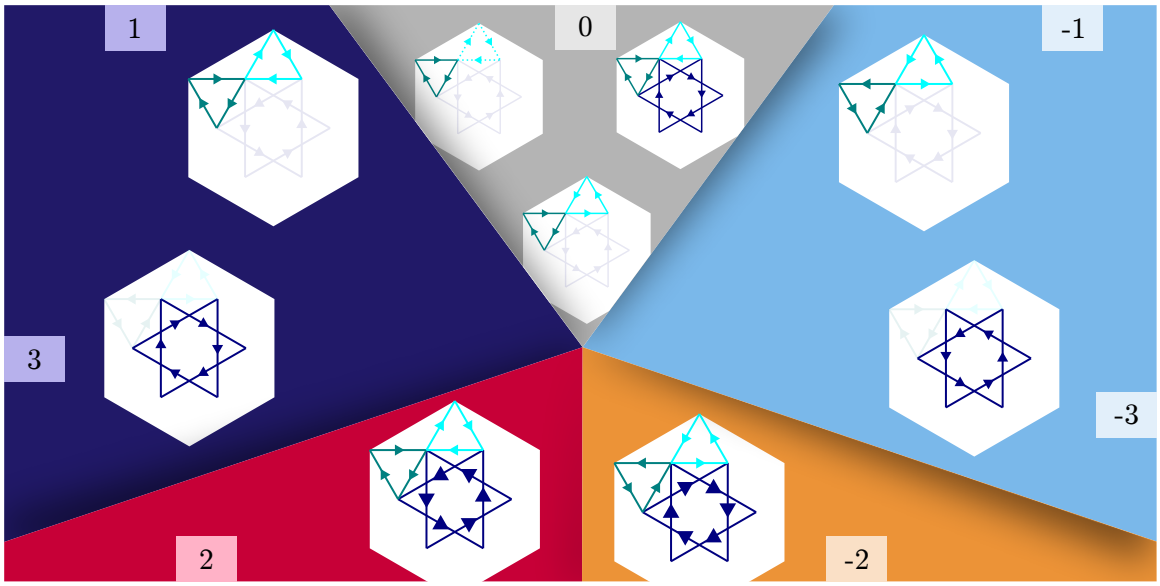

Figure 7.51: Phase diagram for the kagome lattice with a three-site unit cell summarizing the information learned with our statistical analysis. There are in total seven topologically distinct phases, one of which is the trivial insulator. Together with each phase we illustrate typical configurations that clarify how to construct that particular phase. We observe a symmetry w.r.t. 0, i.e., both signs of the Chern label appear symmetrically. The qualitative difference between positive and negative signs is a different orientation of phase windings of the complex hopping parameters. The trivial phase can be realized through anti-parallel winding or an imbalance of magnitudes within different triangles. We group the phases 1, 3 and $-1, -3$ together since they are physically equivalent.

In Fig. 7.51 we gather all of the information that we obtained through our analysis. The resulting figure is a type of phase diagram that illustrates rather well the type of understanding one can expect to achieve with our method, cf. the discussion in Sec. 7.2. We find in total seven topologically distinct phases that are characterized by different Chern numbers. One of them is, of course, the trivial insulator with Chern label $y = 0$. In addition, we find a variety of topological insulators with Chern labels symmetric around 0, i.e., for each positive label there is also a phase with the corresponding negative label. The qualitative difference between phases with positive and negative label is found to be an opposite winding of the phase of complex hopping parameters around unit triangles. Specific required orientations are denoted by arrows in the figure. In contrast to the non-trivial phases, the trivial insulator can be realized with anti-parallel winding in different triangles or an imbalance in the magnitudes of hoppings for different triangles.

The two topologically distinct phases with labels ± 1 and ± 3, respectively, are grouped together. While their topological indices are distinct, and therefore, the two configurations constitute different phases, there is a physical equivalence between the two. Namely, the nearest neighbor only configuration that realizes $y = 1$ is equivalent to the next-nearest neighbor only configuration that realizes $y = 3$ in a lattice with a larger lattice constant of $a' = \sqrt{3}a$. Through this analogy we learn on top of everything else that the Chern number is proportional to the surface area enclosed by the phase winding. The two phases with labels ± 2 appeared less frequently in our data. Nevertheless, we can assign typical configurations. Here, the nearest

neighbor hoppings are generally weaker and the predominant characteristic feature is the large real part of the next-nearest neighbor hoppings that results in a stronger phase vortex. Here, we learn that the Chern number is essentially connected to a winding number that increases if the respective phases get stronger.

We note that most of what we learned could have been anticipated by analogies to simplified models. However, our analysis is based on the most general assumption that, in principle, all symmetries are allowed to be broken. Still, we were able to identify the characteristic properties of all the topological phases in this sector of the phase space. The fact that not all criteria are rigorous (or equivalently that the contrast in the marginal distributions is not perfect) is a consequence of the insensitivity of topological phases w.r.t. geometric details. The location of the exact phase boundary could be extracted, e.g., by optimization of a traditional supervised learning model. We do not expect the result to be particularly useful, though, as it likely has a rather complicated dependence on the multitude of parameters of the most general model. This calculation is therefore incompatible with the type of understanding we sought out to obtain.

We note that the equivalence of Chern numbers 1,3 and $-1, -3$ can be explained in terms of the Hall conductivity. We have seen in Sec. 2.2.1 that the TKNN result relates the Hall conductivity to the Chern number. Experimental evidence then showed via comparison that the Chern number is essentially equivalent to the number of flux quanta per unit volume, i.e., $C = \Phi/\Phi_0$, where Φ is the magnetic flux per unit area. The area of a unit (equilateral) triangle with sides of length a is $\frac{3}{4}a^2$. Considering a uniform magnetic field B, we obtain $\Phi_{\mathrm{nn}} = \frac{3}{4}a^2B$ for the nearest neighbor lattice. On the other hand, the next-nearest neighbor links alone form three independent kagome lattices of "nearest neighbors" that are separated by lengths $a' = \sqrt{3}a$. The ratio of fluxes is therefore $\Phi_{\mathrm{nnn}}/\Phi_{\mathrm{nn}} = 3$, i.e., the number of flux quanta through this stretched lattice is three times as large, which accounts for the factor 3 between Chern indices. Although one is initially inclined to say that the ratio is 9 due to the contributions of the three independent lattices, the analogy is not 100% foolproof, since within our model we can still only have three bands in total, while three fully independent kagomes would have three triple degenerate bands. Therefore, effectively, we only get one larger kagome.

7.8.1 Broken Translational Symmetry

We now briefly summarize how also enlarged unit cells can be treated by our method, which is an ongoing effort in collaboration with Shinibali Bhattacharyya, Francesco Ferrari and Paul Wunderlich. In many realistic systems the translational symmetry is broken by some mechanism that can be related to electronic interactions or the crystalline environment itself. For the kagome systems that attract current interest, a 2×2 cell has been proposed [277, 278], which contains 12 sites. Taking into account nearest neighbor links only, we arrive at 24 hopping parameters, and therefore, an overall (24+11)-dimensional model, where we already subtracted one real parameter to fix the zero of the energy scale. In the following, we will investigate the topological properties at the van Hove singularity, i.e., at filling $5/12$, where we could not observe a finite band gap in the fully translationally symmetric model.

The model is again constructed automatically by our algorithm by generating all links and sorting them w.r.t. their coefficients $n_{1,2}$ in terms of $\mathbf{a}_{1,2}$ such that only positive n_1 is allowed and positive n_2 is preferred over negative in case two equivalent hoppings exist. Generally, any ambiguity is resolved by the condition that links which point from larger to smaller site index are preferred. This extremely general and uninspired choice of hoppings guarantees that we do not imprint any previous knowledge onto the data. We show our choice of the unit cell and all independent links that are used as features in our model in Fig. 7.52. We simply take the unit vectors that we used previously and multiply them by 2. The sites are then labeled arbitrarily as shown in the figure. Together with the number of sites, the number of nearest neighbor links has

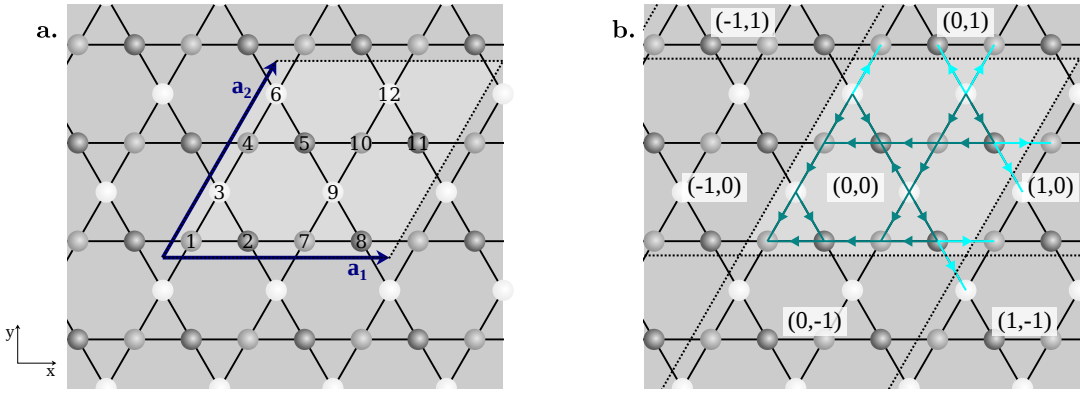

Figure 7.52: Definition of the unit cell for the 2×2 unit cell. **a.** The unit cell that and corresponding lattice vectors are drawn together with the site labels. We here simply chose $\mathbf{a}_{1,2} = 2\mathbf{a}_{1,2}^{1\times1}$. **b.** Independent nearest neighbor hoppings taken into account, where links colored teal are within the unit cell and those that connect neighboring unit cells are colored turquoise. The direction of the links is indicated by arrows.

increased by a factor of four. While the generic method of selecting independent hoppings from conjugate pairs is convenient, apparently, there is also no specific order within the directions of hopping links that are scattered seemingly randomly throughout the unit cell. The reference point is chosen as

$$\mathbf{x}_{\text{ref}} = (\underbrace{0,\ldots,0}_{\times 12},\underbrace{-1,\ldots,-1}_{\times 24}), \tag{7.150}$$

where we intentionally set all local potentials ε_i to zero. Allowing for finite values would not generate useful information, since after removing one parameter to fix the energy scale there would still be 11 parameters left that can appear in a variety of different configurations, which we will not be able to fully capture. In order to remedy this, another approach that we will introduce later is required.

We generate a data set with $n_{\text{samples}} = 10^7$ samples using a spread of $\alpha = 1.5$ for our uniform distribution, cf. Eq. 7.24. An overview over the data is shown in Fig. 7.53**a**, where we identify roughly 67% trivial samples and 16% non-trivial samples with $y = \pm 1$. Notably, we obtain a considerable number of insulating samples, however, almost all samples have separated bands according to a threshold of $\Delta E = 0.05$.

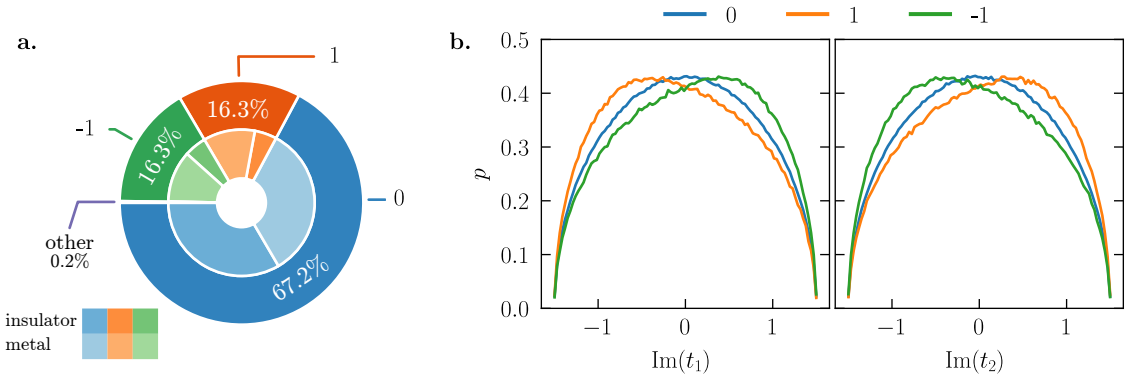

Figure 7.53: **a.** Composition of the data set for the completely generic 2×2 unit cell from Fig. 7.52. The data set contains 10^7 samples in total, of which most have Chern number 0. Roughly 32% of the samples are labeled $y = 1, -1$. **b.** Marginal distribution of the imaginary parts of the first two features. The two distributions are mirror images of each other indicating a particular relationship between hopping parameters among topological samples.

We compute the marginal distributions of all features next. However, since all hoppings are, in principle, equivalent, the importance scores will all be the same. We instead take a peek at the distributions themselves and identify two distinct shapes that are shown in Fig. 7.53**b** for the imaginary parts of t_1, t_2, which we pick as representatives. As for the small unit cell, the marginal distributions appear to identify a particular ordering of phases that we can extract by mapping all distributions to that of t_1 and inverting the hopping link for any t_i whose distribution is mirrored w.r.t. $p(\mathrm{Im}\,[t_1])$. The resulting pattern is again that found for the small unit cell, where all arrows wind with well-defined chirality around unit triangles. By choosing t_1 as the representative, we obtain clockwise winding.

So far, everything works just like in the small unit cell, despite the much larger number of degrees of freedom. We could now try to reduce the complexity by introducing constraints on the parameters such as, e.g., the local inversion symmetry that favors topological phases, in order to construct a simplified model. However, by doing this we are running a risk of discovering the same information that we have already obtained in the small unit cell, albeit at a different filling. A sensible analysis would, however, take notice of this previously acquired knowledge. In order to implement this, we need to make sure that our model breaks translational symmetry explicitly such that we cannot observe the same phases as before.

An analysis on all 24+11 parameters does not seem particularly promising, since statistics obtained from a completely unrestricted sample set will be blurred out due to non-trivial correlations between parameters. It therefore makes more sense to construct symmetric models that break translational symmetry explicitly by taking into account all non-trivial subgroups of the lattice's point group.

Due to the underlying triangular Bravais lattice, the perfect kagome lattice generically has D_6 symmetry, where D_6 is the dihedral group that contains 6-fold rotations and reflections, i.e., 12 group elements [283]. In order to work with these symmetric models more conveniently, we decided to use also a symmetric unit cell that is depicted in Fig. 7.54. Sites are labeled first

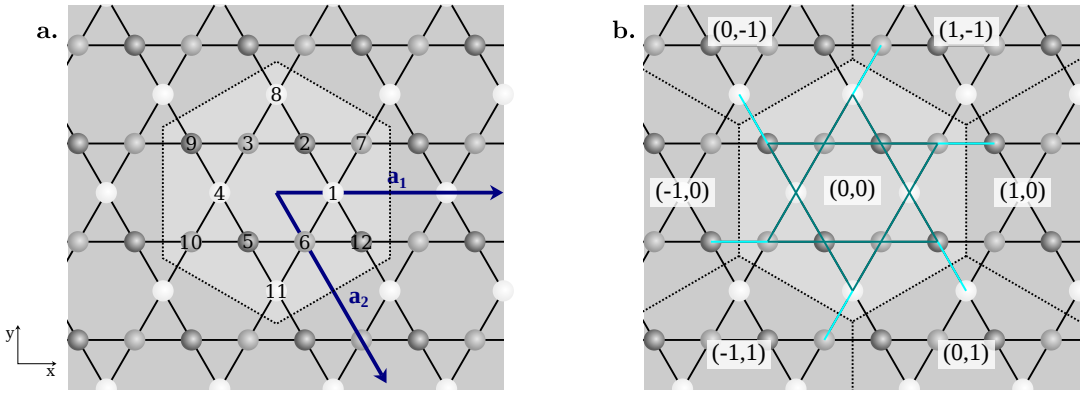

Figure 7.54: Definition of the symmetric unit cell for the 2×2 unit cell. **a.** The unit cell that and corresponding lattice vectors are drawn together with the site labels. Sites are labeled with increasing index from the center of the cell outwards and in counter-clockwise direction around the cell starting at $\mathbf{a_1}$. **b.** Independent nearest neighbor hoppings taken into account, where links colored teal are within the unit cell and those that connect neighboring unit cells are colored turquoise. We do not denote specific directions here, since these generally depend on the point group symmetry that is chosen.

around the inner hexagon. The same order is chosen for the hoppings such that t_1 through t_6 connect sites in the inner hexagon, t_7-t_{18} correspond to the links around the spikes of the Star of David and the remaining t_{19}-t_{24} connect to neighboring cells. Point group symmetries will place constraints on the hoppings, which is why we have not shown a particular direction of hoppings. For example, rotational C_6 symmetry would imply that $\varepsilon_{i<7} = \varepsilon$, $\varepsilon_{i\geq 7} = \varepsilon'$, $t_{i\leq 6} = t$,

$t_{7,9,11,13,15,17} = t'$, $t_{8,10,12,14,16,18} = t''$ and $t_{i\geq 19} = t'''$ with all directions of links corresponding to a well-defined chirality. This leaves a total of 4+2 degrees of freedom. The full D_6 symmetry on the other hand requires that $t = t^*$, $t' = t'^*$, $t''' = t'''^*$, i.e., there are only 1+4 independent features.

We show in Fig. 7.55 two possible subgroups of D_6. Fig. 7.55**a** represents the trivial subgroup, i.e., the group itself, which allows for three independent parameters (neglecting onsite terms). The links around the inner hexagon must be real due to reflection axes that go through the links' midpoints and the same applies to the outer links under a combined rotation and reflection. The remaining arrow directions were chosen arbitrarily, however, we note that the relation between teal arrows, i.e., links along the spikes of the Star of David, is dictated by symmetry. Note that this breaks the local inversion symmetry that we found to be a signature of topological phases in all lattice points unless everything is real and, more importantly, a well-defined chirality cannot be achieved. Therefore, we do not expect systems with D_6 symmetry to host topological phases, and indeed, after generating a data set subject to this symmetry we find virtually no topological samples with those reported topological having questionable band separation, i.e., likely being wrongly labeled.

In Fig. 7.55**b**, we illustrate the reduced C_6 symmetry, which allows for 4 (complex) degrees of freedom, again neglecting onsite terms. We notice that the specific choice of arrows drawn already realizes the characteristic configuration with globally well-defined winding of phases and is therefore expected to host topological phases. In fact, the trivial case, where all parameters are chosen equal, i.e., the truly inversion symmetric configuration, is simply a duplicate of the $C = \pm 1$ phase that we found in the minimal unit cell, i.e., we can also expect this phase here, although it has to be confirmed specifically at the changed filling. By choosing the hopping parameters $t' = t''$, i.e., equal along the spikes of the Star of David, we reduce the number of independent complex parameters to three. For this specific symmetry, which can also be obtained from D_6 by combining reflections with time-reversal, we performed another calculation and confirmed that we do find a considerable number of non-trivial samples with labels $y = \pm 1$. We note that, since it is statistically very unlikely that all parameters have the same values, all our samples break the full translational symmetry of the lattice.

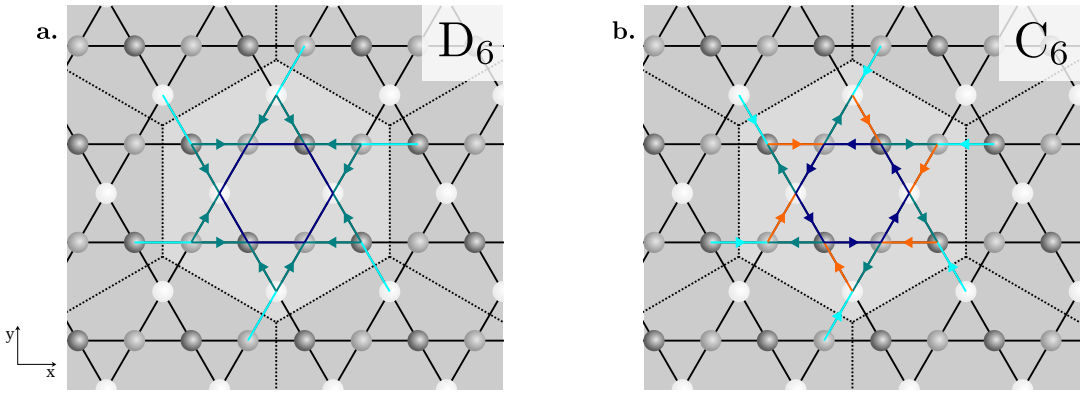

Figure 7.55: Illustration of the constraints imposed on the hopping parameters through point group symmetries. Equivalent hoppings are drawn in the same colors. Arrows indicate the direction corresponding to equal phase. **a.** The full D_6 symmetry is shown. There are three independent hoppings, the parameters corresponding to hopping around the inner hexagon and the outer links are required to be real. **b.** Reduced C_6 symmetry, i.e., six-fold rotations about the z-axis through the center of the unit cell only. There are four independent hoppings. The configuration shown realizes local inversion symmetry (considering the direction of the arrows only) at every lattice point.

Out of the 12 elements of D_6, 8 subgroups can be constructed. Additionally, by including

also time-reversal symmetry, the number of symmetry operations can be increased to 24. The strategy is now to generate data sets for specific symmetry groups that impose patterns that break translational symmetry by construction (except for $\mathbb{1}$ and time-reversal Θ), and then analyze these data sets to obtain a similar type of phase diagram as shown in Fig. 7.51 for the translationally symmetric system. Since this discussion is focused entirely on specific symmetries, which is in conflict with the completely unbiased approach that we usually strive for and seems unrealistic considering that most real materials break symmetries in one way or another, we would also study the effect of breaking these symmetries by allowing for small deviations from the perfectly symmetric configuration and investigating the stability of topological phases. Finally, in order to speed up the investigation we intend to investigate all possible symmetries at once by choosing a particular symmetry group at random during the sampling procedure. This will create statistics over symmetry labels that reveal their compatibility with topology. Finally, we intend to also include higher-order neighbors in the description.

At the time of writing, a number of promising results have already been obtained, however, the discussion of these and more will be part of a future publication. Therefore, we close this subject here and discuss in the remainder of this chapter possible generalizations of the method that we presented.

7.9 Perspective Towards Material Application

The motivation for this project was the search for a scheme that would allow us to engage in the actual process of engineering new topological materials. The results presented in the previous sections and the ongoing work on even more complex systems seem to indicate that this can, in fact, be achieved. Here, we have focused entirely on the development of the toolkit in terms of model systems. In the following, we describe the extension to real materials and motivate how the approach that we developed in this chapter can be used in a production setting.

The transformation to a realistic system is actually rather simple, since we already chose a formalism that is easily extended to realistic systems by virtue of a mapping to a tight-binding model. The latter is in practice done through wannierization of those DFT bands that are most important for the low energy physics by constructing maximally localized Wannier functions from the Kohn-Sham Bloch states. These can then be used to compute tight-binding parameters according to Eq. 3.58. All of this is conveniently already available in the Wannier90 package [284], which can operate on results from all common DFT codes through interfaces, thus enabling a route for the ab initio calculation of a tight-binding model for a particular material. This model is then simply used as the reference point in the statistical analysis, which subsequently probes the surrounding parameter space for possible topological phases. Clearly, we would in this case be interested mainly in small perturbations, since the validity of the tight-binding model deteriorates the larger our perturbations become. This can be controlled effectively by sampling from a Gaussian distribution instead of the uniform distribution that we used in our investigation of model systems. If topological candidate samples are detected, all data analysis methods developed for model systems can be employed to reduce the typically large number of parameters to the most essential ingredients that provide the proverbial knob which allows one to tune the system into a topological phase. As for the realization of this topological phase, one then has to construct a candidate material that corresponds to the perturbed tight-binding model. This step is non-trivial, however, we believe that intuition and experience is key. Our choice of the framework of tight-binding parameters that offers a rather direct real-space interpretation of the model and is therefore universally comprehensible has been chosen such that both theoretical and experimental colleagues can contribute their respective experience.

We note that the search for possible candidates could be narrowed down further by restricting

to experimental protocols that can be implemented realistically. These could include, e.g., distortions of the lattice through application of pressure or strain, doping, application of external fields.

Carrying out this scheme is left for future work, however, we remark that the our implementation of this protocol already accepts a reference point in WANNIER90 format. We focused here entirely on the Chern number, i.e., the quantum Hall phase in two-dimensional systems, which is a rather unique case, given that most materials give rise to three-dimensional electronic models. Nonetheless, this does not impede the general applicability of our scheme, since the topological label can simply be exchanged to distinguish, e.g., 3D topological insulators. Moreover, the methodology is not necessarily restricted to topological class labels and one can, in principle, also study other phase diagrams—under the condition that a mapping to discrete class labels exists.

7.10 Information Theoretical View

We will briefly illustrate an information theoretical view onto the same problem of topological classification. As a quick recap: we are faced with data (X, Y) that defines a relationship $f : X \mapsto Y$, where $X \in \mathbb{C}^n$ and $Y \in \mathbb{Z}$. Immediately, the different definition and target sets of f make it clear that there cannot be a unique mapping since $|\mathbb{Z}| < |\mathbb{C}^n|$, where $|\cdot|$ denotes the cardinality of the corresponding set.

This bears resemblance to hash functions, which, too, lack the property of invertibility as they map from a virtually unbounded set $\{n^m \mid m \in \mathbb{N}\}$ to, e.g., n^{128} for a hash of length 128, where n is the size of the alphabet. Unlike hash functions found in cryptography, however, inputs that are close together in terms of, e.g., the 2-norm are expected to also yield the same classifier in the majority of cases, i.e., all samples that do not lie close to a phase transition. This property practically enables machine learning of the classification. On the contrary, crytographic hash functions are constructed such that they cannot be learned by practical machine learning algorithms.

Instead of cryptographic hash functions we can interpret the classifier as an extremely lossy code—an analogy that we will motivate in the following. The classifier maps each data point, which could be considered a word in the language of data compression, to a class label that is much smaller in size. Words that get mapped onto the same class label are considered equivalent after data compression. While this is usually not intended for the sake of reversibility that allows one to read the original data, it turns out that a classification in terms of an order parameter is just that. Let us assume that we have class labels A, B. Then, all words that map onto A are considered equivalent. This interpretation is astoundingly true for a topological classification, where Hamiltonians can be smoothly connected, and therefore belong to the same equivalence class iff their topological class labels are identical. Hence, there is a correspondence between the mathematical concept of equivalence classes and data compression.

The entropy of the original "message", i.e., our input data, is extremely large due to both the high number of degrees of freedom and the high variability of values along each dimension. Assuming a Monte Carlo method like the formerly introduced data sampling scheme, we obtain

$$H(X) = H(x_1, x_2, \ldots) = \sum_{j=1}^{n_{\text{features}}} H(x_j) = -\sum_{j=1}^{n_{\text{features}}} \sum_{i=1}^{N_j} p_{j,i} \log p_{j,i}, \tag{7.151}$$

where $p_{j,i}$ is the probability of letter i to occur for feature j and N_j the size of the alphabet. For our data set we assume uniformly distributed data and since here we have a continuum of letters, we approximate the probability density function through a discrete distribution $p_{j,i} = \frac{1}{N_j}$, where

N_j is the number of bins. Then,

$$H(X) = \sum_{j=1}^{n_{\text{features}}} \log N_j. \tag{7.152}$$

Apparently, this is much larger than the size of the set of class labels, which contains just a handful of numbers. Therefore, we are far below the optimal compression limit $H(X)$ given by the Shannon source coding theorem, which means that recovering the original data from the class labels is impossible. What is possible, however, is to recover a representative of the corresponding equivalence class, which is formalized via the following code:

For each input and class check if a smooth connection between input and the representative of the class exists.

In order to invent such a code—which is a requirement for finding topological order in data with an unsupervised learning technique—one has to find a proper similarity measure that expresses the "smooth" connection between words. This is a highly non-trivial task and it seems rather unlikely that this can be achieved with generic machine learning algorithms. It has been demonstrated in Refs. [196, 198], however, that, given a cleverly chosen distance measure, one can indeed detect topological order from unlabeled data. In their case, the authors have chosen an elaborate similarity measure that is computed via an optimization scheme. This high level of complexity is apparently not always required, cf. Ref. [217], where instead the Chebychev distance defined on the single particle Hilbert space is used. In a more general setting, one could make use of the recently introduced concept of homotopic distance [285].

7.11 Interacting Systems

During the entire discussion in this chapter we completely neglected electronic interactions. The motivation for this is fueled mainly by the possibility to describe the topology of interacting systems through an auxiliary non-interacting model—the topological Hamiltonian, cf. Sec. 2.5.2. In addition, our earlier investigation of correlation effects, cf. Chapter 6, revealed that momentum-dependent corrections do not lead to qualitatively different physics, and therefore, our method can easily be adapted to also include a local self-energy term. Available results for interacting systems indicate that usually topological phases are simply shifted [137, 139, 141, 175] and in all but strongly correlated systems no new physics compared to the non-interacting case are expected to appear.

In weakly interacting systems we expect that our general methodology is very relevant and to some degree interaction information can be added in terms of a simple mean-field picture. In the following, we want to elaborate on possible applications of similar ideas to intrinsically interacting problems that go beyond the scope of this work and are expected to be tackled in future research projects.

Generally, a treatment within the topological Hamiltonian framework appears manageable, since it is essentially a single particle model. However, if non-local corrections to the self-energy shall be taken into account, we believe that this makes sense only in the way we have implemented our statistical analysis in Chapter 6, where the tight-binding model remained fixed and the only free parameters were given by a parameterization of the self-energy. Combining the two approaches, i.e., sampling over both single particle and self-energy parameters, would most likely lead to redundancies in the expression of the final model, and therefore, somewhat limit the amount of information that is accessible via such an approach. It is expected that worthwhile information that differentiates between interaction and single particle effects can only be obtained from non-overlapping parameters, i.e., parameters that appear only in one of

the two terms. Such a requirement could be derived in terms of a set of linearly independent functions $f_\alpha(k)$ through an expansion

$$H_t(k) = H_0(k) + \Sigma(k) = \sum_\alpha (h_\alpha + \Sigma_\alpha) f_\alpha(k), \tag{7.153}$$

where the non-overlapping condition can be expressed as

$$\min\{|h_\alpha|, |\Sigma_\alpha|\} = 0 \quad \forall \alpha. \tag{7.154}$$

Unfortunately, due to the required generality of the self-energy, Eq. 7.154 can not really be satisfied without restricting the possible solutions of the many-body problem.

We expect that a more promising approach can be formulated by leaving the realm of tight-binding parameters and instead take as data the possible solutions to a manybody problem. Since many different formalisms exist, we have to evaluate them regarding their suitability. In our opinion, a description in terms of correlation functions is not necessarily the most straightforward option, since these functions are rather complicated, and there exists no obvious choice for a basis in which these functions are best described. This is a challenge also, in particular, since the physical and analytic properties of Green's functions impose many constraints on expansion parameters.

A much simpler approach could instead be based on many-body wave functions in analogy to the variational Monte Carlo (VMC) technique, where a reasonably general ansatz is chosen and parameterized through a set of variables that compose the data set. If the ansatz is chosen carefully, the only constraint necessary is normalization, which is equivalent in nature to the scale degree of freedom in the tight binding model. Given such a parameterization, one can then investigate the properties of wave functions that realize topological phases using the exact same methods that were applied in the non-interacting case.

One of the prerequisites to such a supervised approach is the existence of a way to label individual samples. For many-body wave functions, e.g., the Hall conductivity formula by Niu et al., cf. Sec. 2.5.1, comes to mind. However, a careful analysis reveals that the topological index depends not on the ground state wave function alone, but on a set of wave functions obtained for different boundary conditions. The sensitivity to boundary conditions is obviously not included in a single randomly sampled wave function, and so it appears as if an analysis like this cannot be made. In addition to the formula by Niu et al., the topological phase is generally related to a change in polarization when a parameter is varied along a closed path. According to Resta, we have [80]

$$P = \frac{1}{2\pi} \mathrm{Im}\left[\log\langle\psi_0| e^{i\frac{2\pi}{L}X} |\psi_0\rangle\right], \tag{7.155}$$

where L is the size of the system and $|\psi_0\rangle$ the ground state. Of course, the variation along a path is again a manifestation of the same problem that the topology is not uniquely determined by the ground state wave function alone. Therefore, we propose to sample evolutions of wave functions along a path $R(t)$ that represents the entire set of parameters at each step t. Due to the requirement of periodicity in t that arises from the assumption of a closed path, i.e., $R(t+T) = R(t)$, the functional behavior of R on t can be expanded in terms of a set of periodic functions. Here, a similar approach as in our self-energy analysis in Chapter 6 is thinkable and a careful data analysis could reveal interesting relationships between the behavior of the wave function and the topological phase.

The best case scenario results in a prediction of particular "topological wave functions". It is then an open problem to find actual realizations of these states, however, we believe that in using a more general ansatz this method can provide a wave function with a reduced parameter set that could either lead to a better convergence of the VMC algorithm or provide new variational wave functions that could reveal interesting physical properties of topological phases.

7.12 Summary

We have here proposed a scheme aimed at understanding and ultimately predicting novel realizations of topological states. Since predictions are a difficult matter—especially if the exact type of information required is not entirely well-defined—we started out with a rather general discussion of the type of abstract information that is typically encoded in terms of phase diagrams. Systems for which a phase diagram is known are considered "understood", therefore, being able to extract this information would also improve our understanding of topological phases. In real materials, the number of parameters (here, we chose a representation in terms of tight-binding parameters) is usually rather large, and therefore, an all-encompassing understanding cannot be expected. Instead we devised a scheme that allows us to reduce the number of parameters to the ones that are most essential, i.e., that contain most of the information about the topological phase, and thereby arrive at reduced models that are much easier to comprehend. We explored the possibility to straight-forwardly apply common machine learning algorithms to this problem, however, neither one of those resulted in a completely satisfactory description. Therefore, we developed a statistical approach that is based on the information theoretical concept of statistical distance of marginal distributions as an information measure together with descriptors of two- and three-parameter correlations. For the development of the method, the Haldane model served as our testbed and we were able to show that by starting from a completely generic and totally uninspired tight-binding model on the honeycomb lattice we can systematically construct the Haldane model as a characteristic model for the topological phases in this lattice. In addition, we found other realizations with broken symmetries that cannot be described in terms of the Haldane model. We note that we took special care of not making use of any prior knowledge about the Haldane phase diagram that we had studied in detail in Chapter 6.

As a demonstration of the method we performed an analysis of the general topological phase diagram for systems on the kagome lattice. Again, we started out with a data set that contained zero prior knowledge by constructing a generic model with up to third-nearest neighbors. Based on this data we were able to identify characteristic configurations for topological phases through the inspection of marginal distributions and correlations between parameters. For the description of correlations we performed feature engineering and were able to identify necessary conditions for topological phases solely from an analysis of the data. Finally, we arrived at a phase diagram that contains describes prototypes for each of the observed topological phases.

The procedure was then applied also to kagome systems with broken translational symmetry, where we discussed the use of point group symmetries in order to deal with the statistical noise that is necessarily present in completely unconstrained models.

Finally, we discussed possibilities for further adaptation of the method towards real materials and possibly also to interacting systems.

Chapter 8

Conclusion

In this thesis we have closely investigated several aspects in the context of topological phases in condensed matter systems. In particular, we have focused on the description of Chern insulators that appear in systems without time-reversal symmetry. Motivated by the important work on the topological Hamiltonian by Wang and Zhang that relates the topological invariant of an interacting system to the Chern number of an auxiliary non-interacting Hamiltonian, we studied the influence of common approximate numerical methods on the resultant phase diagram. The topological Hamiltonian is constructed from two parts—the non-interacting Hamiltonian and the full momentum-dependent self-energy at frequency zero. Taking into account the fact that the Chern number is a measure of the Berry curvature in momentum-space, we expected that it is highly sensitive also to the momentum-dependence of the self-energy. This is especially interesting given the popularity of the dynamical mean-field theory (DMFT) in the field, since DMFT neglects this momentum-dependence entirely. For our study we chose the most general model imaginable—the ionic Hubbard model on the square lattice, which combines in itself a strong antiferromagnetic instability that leads to a strong momentum-dependence of the self-energy even at small interaction strengths, and the mass term (a.k.a., the ionic potential) that is found in most topological models throughout the literature. In order to quantify the importance of the momentum-dependence, we introduced a measure that we call the "self-energy dispersion amplitude", which simply measures the variation of the self-energy throughout the Brillouin zone with the simple motivation that, provided this measure is very small, the DMFT result should be trustworthy. We computed the self-energy dispersion amplitude explicitly by means of the two-particle self-consistent method (TPSC) that produces a fully momentum-dependent solution and is applicable in the weak to intermediate coupling regime. This revealed a type of phase diagram that distinguishes a local regime, where non-local interaction effects are not important and a non-local regime where the opposite is the case. Despite the strong momentum-dependence on the square lattice, we found the regime where the local approximation is applicable to be rather large indicating that previous DMFT studies should, in fact, be trustworthy. We noticed that the transition between the two regimes can be understood in terms of the competition between the potential energy contributions from the ionic potential and the Hubbard interaction that balance each other out at the transition. In a more quantitative description we managed to relate this competition to an order parameter that we computed with exact diagonalization that—like TPSC—takes into account non-local effects and DMFT itself. Since this parameter only depends on local quantities the DMFT solution agrees very well with that obtained by other methods, which reveals an internal error indicator for DMFT. We carried out the same approach also for the triangular lattice and found consistent results.

Having established that DMFT should be applicable in a large region of the phase diagram simply due to the absence of a strong momentum-dependence, we then turned our atten-

tion to the more general case, i.e., how a finite momentum-dependence affects the topological classification—here, applied to the Haldane-Hubbard model. Since all available methods are merely approximate and for this reason bear the risk of simply overlooking the important information, we decided to tackle this case from an entirely different perspective via the use of a stochastic algorithm, a.k.a, statistical method. We constructed this method to be as unbiased as possible by investigating all possible self-energies and making no assumptions other than a certain degree of smoothness that is expected for the low to intermediate coupling regime and the conservation of the symmetries of the non-interacting Hamiltonian. The systematic approach that we followed involved decomposing the total self-energy into a local and a non-local part, which for the Haldane model can be represented in terms of Pauli matrices. By computing the Chern number for a finite but unbiased subset of all possible self-energies, our statistical method provided a probabilistic result for the location of the true phase transition that turned out to lie approximately on top of the local transition on average, subject to a finite variance. We investigated this variance also quantitatively and found that, in fact, the probability for the location of the transition line decays exponentially with increasing distance from the local transition, and the corresponding length scale is given approximately by the self-energy dispersion amplitude. Given an expected strength of the momentum-dependence, we can therefore immediately predict a window of possibility for the actual topological transition that is valid irrespective of the numerical method used to solve the many-body problem. We finally used our findings also to discuss the phase diagram of the Haldane-Hubbard model, where different numerical methods predicted contrasting transition lines, and found that the predominant source of the deviation was most likely the different treatment of magnetic order that mainly affects the local self-energy.

Fueled by the success of the statistical viewpoint, we found another interesting application in a similar context—the understanding of topological phase diagrams in high dimensions as is the case for realistic systems that are usually characterized through a multitude of independent degrees of freedom. While understanding a phase diagram is predominantly of academic interest, this machinery is also expected to be valuable for a more general audience as it promises to pave the road towards a systematic way for engineering new topological materials. Since a growing amount of research has already been done in the field of prediction—predominantly making use of artificial intelligence or machine learning methods—we decided to evaluate possible uses of such existing algorithms to our specific problem, that is, finding a connection between the topological invariant and the tight-binding parameters of a given system. We showed first that this is rather difficult in terms of clustering algorithms, since the data is inherently structureless, which requires us to add the topological invariant as a label to the data. The structure that is introduced through the label can then be analyzed by a plethora of supervised learning techniques. Here, we are not interested in simply reproducing the label, but in understanding how the label is connected to the input data. Therefore, we chose with decision trees an architecture that is known to be interpretable by design. After training these models on our randomly generated data set we showed that—despite its reputation—the resulting classifier is rather difficult to understand in this particular case and the only information that can reasonably be extracted is the importance of the respective input variables for the classification.

We then introduced our statistical approach to the same problem and showed that using statistical and information theoretical tools, the same information can be obtained from the initial data set directly, without performing any training whatsoever. This saves time for training and avoids a possible bias of the model by completely taking this "middle man" out of the equation. We used the Haldane model as an example and showed that by investigating a cloud of data points centered around the Haldane configuration, a lot of properties of the model can, indeed, be extracted through data analytic means. We then removed the initial bias and

demonstrated that even without knowledge of the Haldane model, its properties can be deduced from a data set that is constructed generically on the honeycomb lattice. In addition, we found configurations that are dissimilar to the Haldane phases, since they break the symmetry between nearest neighbor hopping parameters.

Having arrived at a workflow and established its capability to describe the Haldane model, we moved on to investigate topological phases in the kagome lattice. By applying the same methodology, we were able to uncover non-trivial relationships between the tight-binding parameters and the Chern number, and arrived finally at a type of phase diagram that reveals qualitatively how the distinct topological phases that were contained in our data set are linked to configurations of hopping parameters—once again delivering proof of the power of the method.

Outlook

The investigations into the effects of non-local self-energies on the topological classification lead to a rather final conclusion such that obvious additional research would focus primarily on either improving numerical methods to get closer to the exact result instead of our probabilistic estimate, or applying our statistical method to other systems or other topological invariants.

The, in our opinion, most interesting avenues that are opened up by this work, however, follow along the lines of the investigation of topological phase diagrams. Here, we focused on two-dimensional systems without time-reversal symmetry that are classified by the Chern number only. Of course, the method can easily be adapted by swapping the Chern label for another topological invariant, so that, in principle, also three-dimensional topological insulators could be investigated along these lines. We have motivated in detail how this method can be applied directly to material research by using, e.g., density functional theory results to shift focus to a particular interesting region of the phase space that could be accessible from existing compounds through experimental means. On the other hand, we believe that there is also potential for understanding interacting systems, in particular, in combination with wave function techniques such as variational Monte Carlo. It would also be interesting to explore the application of data analytic methods such as those presented in this work to the optimized models of traditional machine learning approaches. This could be used as a way to overcome the issues with interpretability that most of these general purpose methods have.

Acknowledgments

I would like to thank everyone who has been a part of my journey towards completing this final milestone in my scientific career. First of all, I thank my PhD advisor Prof. Dr. Roser Valentí for giving me absolute freedom to follow whatever thought I found interesting and generally providing a very friendly and comfortable working environment that I'll be leaving behind me with many positive memories. I thank also Prof. Dr. Falko Pientka for his willingness to referee this thesis.

Many thanks go out to Karim Zantout and Julian Stobbe for proof-reading most of the thesis and offering helpful advice. I especially enjoyed the collaboration with Karim in some of the projects discussed here.

I thank my parents for supporting me all the way through many levels of higher education which eventually led to my authorship of this document.

To everyone that has been a part of our group during the last four and a half years, that is, (in no particular order) Michaela, Kira, Karim, Sananda, Dominik, David, Aleksandar, Lucas, Benjamin, Niclas, Ying, Adrian, Simon, Marius, Max, Francesco, Shinibali, Young-Joon, Paul, Vladislav, Aaram, Steve, Julian, Lisa and whomever I inevitably forgot, I extend my deepest thanks for always keeping the spirits up and making lunch a highlight of every day. I will miss the daily trips to REWE and our Tischkicker tournaments that were continually proposed especially by Aleksandar.

I thank Anne Metz for her assistance with many organizational tasks and for understanding our frequent frustration with bureaucratic nonsense.

Thanks go out also to the students from our sister group, who shared offices and time with us. In particular, I thank Hendrik for lots of fun, cake and coffee, although I never drank any.

I thank Steffen for setting up a Minecraft server, where we could find distraction from day-to-day business. Also, how awesome was Outer Wilds? Keep exploring!

Lastly, I thank all my friends who contributed to other parts of my life during this very demanding time of writing.

Bibliography

[1] B. Andrei Bernevig and Taylor L. Hughes. *Topological Insulators and Topological Superconductors*. Princeton University Press, 2013.

[2] David Vanderbilt. *Berry Phases in Electronic Structure Theory: Electric Polarization, Orbital Magnetization and Topological Insulators*. Cambridge University Press, 2018.

[3] M. Z. Hasan and C. L. Kane. Colloquium: Topological insulators. *Reviews of Modern Physics*, 82(4):3045–3067, October 2010.

[4] Thomas Mertz. Topological Quantum Phase Transitions and the Hofstadter-Hubbard Model. Master's thesis, Goethe Universität Frankfurt am Main, 7 2016.

[5] Edwin H. Hall. *American Journal of Mathematics*, 2:287–292, 1879.

[6] D. J. Thouless, M. Kohmoto, M. P. Nightingale, and M. den Nijs. Quantized Hall Conductance in a Two-Dimensional Periodic Potential. *Phys. Rev. Lett.*, 49:405–408, Aug 1982.

[7] J M Kosterlitz and D J Thouless. Long range order and metastability in two dimensional solids and superfluids. (Application of dislocation theory). *Journal of Physics C: Solid State Physics*, 5(11):L124–L126, jun 1972.

[8] J M Kosterlitz and D J Thouless. Ordering, metastability and phase transitions in two-dimensional systems. *Journal of Physics C: Solid State Physics*, 6(7):1181–1203, apr 1973.

[9] F. D. M. Haldane. Nonlinear Field Theory of Large-Spin Heisenberg Antiferromagnets: Semiclassically Quantized Solitons of the One-Dimensional Easy-Axis Néel State. *Phys. Rev. Lett.*, 50:1153–1156, Apr 1983.

[10] Michael Schirber. Nobel Prize—Topological Phases of Matter. *Physics*, 9(116), 2016.

[11] H. Aoki and T. Ando. Effect of localization on the Hall conductivity in the two-dimensional system in strong magnetic fields. *Solid State Communications*, 38(11):1079–1082, 1981.

[12] R. E. Prange. Quantized hall resistance and the measurement of the fine-structure constant. *Phys. Rev. B*, 23:4802–4805, May 1981.

[13] D J Thouless. Localisation and the two-dimensional Hall effect. *Journal of Physics C: Solid State Physics*, 14(23):3475–3480, aug 1981.

[14] R. B. Laughlin. Quantized Hall conductivity in two dimensions. *Phys. Rev. B*, 23:5632–5633, May 1981.

[15] B. I. Halperin. Quantized Hall conductance, current-carrying edge states, and the existence of extended states in a two-dimensional disordered potential. *Phys. Rev. B*, 25:2185–2190, Feb 1982.

[16] Felix Bloch. Über die Quantenmechanik der Elektronen in Kristallgittern. *Zeitschrift für Physik*, 52:555–600, 1929.

[17] Charles Kittel. *Introduction to Solid State Physics*. Wiley, 8 edition, 2004.

[18] Neil Ashcroft and N. David Mermin. *Solid State Physics*. Saunders College Publishing, 1 edition, 1976.

[19] Ryogo Kubo, Satoru J. Miyake, and Natsuki Hashitsume. Quantum Theory of Galvanomagnetic Effect at Extremely Strong Magnetic Fields. volume 17 of *Solid State Physics*, pages 269–364. Academic Press, 1965.

[20] P G Harper. The General Motion of Conduction Electrons in a Uniform Magnetic Field, with Application to the Diamagnetism of Metals. *Proceedings of the Physical Society. Section A*, 68(10):879–892, oct 1955.

[21] Douglas R. Hofstadter. Energy levels and wave functions of Bloch electrons in rational and irrational magnetic fields. *Phys. Rev. B*, 14:2239–2249, Sep 1976.

[22] J. E. Avron, R. Seiler, and B. Simon. Homotopy and Quantization in Condensed Matter Physics. *Phys. Rev. Lett.*, 51:51–53, Jul 1983.

[23] Michael V. Berry. Quantal phase factors accompanying adiabatic changes. *Proc. R. Soc. Lond. A*, 392:45–57, 1984.

[24] Mikio Nakahara. *Geometry, Topology and Physics*. CRC Press, second edition, 2003.

[25] Mahito Kohmoto. Topological invariant and the quantization of the Hall conductance. *Annals of Physics*, 160(2):343–354, 1985.

[26] Yasuhiro Hatsugai. Edge states in the integer quantum Hall effect and the Riemann surface of the Bloch function. *Physical Review B*, 48(16):11851–11862, October 1993.

[27] Yasuhiro Hatsugai. Chern number and edge states in the integer quantum Hall effect. *PRL*, 71(22):3697–3700, November 1993.

[28] Frank Wilczek and A. Zee. Appearance of Gauge Structure in Simple Dynamical Systems. *Phys. Rev. Lett.*, 52:2111–2114, Jun 1984.

[29] K. v. Klitzing, G. Dorda, and M. Pepper. New Method for High-Accuracy Determination of the Fine-Structure Constant Based on Quantized Hall Resistance. *Phys. Rev. Lett.*, 45:494–497, Aug 1980.

[30] Klaus von Klitzing. The quantized Hall effect. *Rev. Mod. Phys.*, 58:519–531, Jul 1986.

[31] von Klitzing, K. The Quantum Hall Effect. *Europhys. News*, 13(4):2–4, 1982.

[32] Alexander Altland and Martin R. Zirnbauer. Nonstandard symmetry classes in mesoscopic normal-superconducting hybrid structures. *Phys. Rev. B*, 55:1142–1161, Jan 1997.

[33] Élie Cartan. Sur une classe remarquable d'espaces de Riemann. *Bulletin de la Société Mathématique de France*, 54:214–264, 1926.

[34] Élie Cartan. Sur une classe remarquable d'espaces de Riemann. II. *Bulletin de la Société Mathématique de France*, 55:114–134, 1927.

[35] Andreas P. Schnyder, Shinsei Ryu, Akira Furusaki, and Andreas W. W. Ludwig. Classification of topological insulators and superconductors in three spatial dimensions. *Phys. Rev. B*, 78:195125, Nov 2008.

[36] Andreas P. Schnyder, Shinsei Ryu, Akira Furusaki, and Andreas W. W. Ludwig. Classification of Topological Insulators and Superconductors. *AIP Conference Proceedings*, 1134(1):10–21, 2009.

[37] 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.

[38] Martin R. Zirnbauer. Riemannian symmetric superspaces and their origin in random-matrix theory. *Journal of Mathematical Physics*, 37(10):4986–5018, 1996.

[39] Raoul Bott. The periodicity theorem for the classical groups and some of its applications. *Advances in Mathematics*, 4(3):353–411, 1970.

[40] Yuan-Ming Lu and Dung-Hai Lee. Inversion symmetry protected topological insulators and superconductors, 2014.

[41] Liang Fu and C. L. Kane. Topological insulators with inversion symmetry. *Phys. Rev. B*, 76:045302, Jul 2007.

[42] Barry Bradlyn, L. Elcoro, Jennifer Cano, M. G. Vergniory, Zhijun Wang, C. Felser, M. I. Aroyo, and B. Andrei Bernevig. Topological quantum chemistry. *Nature*, 547(7663):298–305, July 2017.

[43] Qian Niu, D. J. Thouless, and Yong-Shi Wu. Quantized hall conductance as a topological invariant. *Phys. Rev. B*, 31:3372–3377, Mar 1985.

[44] R. Resta. The insulating state of matter: a geometrical theory. *Eur. Phys. J. B*, 79:121–137, 2011.

[45] D. N. Sheng, L. Sheng, Z. Y. Weng, and F. D. M. Haldane. Spin Hall effect and spin transfer in a disordered Rashba model. *Phys. Rev. B*, 72:153307, Oct 2005.

[46] D. N. Sheng, Z. Y. Weng, L. Sheng, and F. D. M. Haldane. Quantum Spin-Hall Effect and Topologically Invariant Chern Numbers. *Phys. Rev. Lett.*, 97:036808, Jul 2006.

[47] Takahiro Fukui and Yasuhiro Hatsugai. Topological aspects of the quantum spin-Hall effect in graphene: Z_2 topological order and spin Chern number. *Phys. Rev. B*, 75:121403, Mar 2007.

[48] Emil Prodan. Robustness of the spin-Chern number. *Phys. Rev. B*, 80:125327, Sep 2009.

[49] Zhong Wang and Shou-Cheng Zhang. Simplified Topological Invariants for Interacting Insulators. *Phys. Rev. X*, 2:031008, Aug 2012.

[50] Zhong Wang and Shou-Cheng Zhang. Strongly correlated topological superconductors and topological phase transitions via Green's function. *Phys. Rev. B*, 86:165116, Oct 2012.

[51] Zhong Wang and Binghai Yan. Topological Hamiltonian as an exact tool for topological invariants. *Journal of Physics: Condensed Matter*, 25(15):155601, mar 2013.

[52] K. Ishikawa and T. Matsuyama. Magnetic field induced multi-component QED_3 and quantum Hall effect. *Zeitschrift für Physik C - Particles and Fields*, 33:41–45, 1986.

[53] Hiroto So. Induced Topological Invariants by Lattice Fermions in Odd Dimensions. *Progress of Theoretical Physics*, 74(3):585–593, 09 1985.

[54] E. Y. Loh, J. E. Gubernatis, R. T. Scalettar, S. R. White, D. J. Scalapino, and R. L. Sugar. Sign problem in the numerical simulation of many-electron systems. *Phys. Rev. B*, 41:9301–9307, May 1990.

[55] James Gubernatis, Naoki Kawashima, and Philipp Werner. *Quantum Monte Carlo Methods: Algorithms for Lattice Models*. Cambridge University Press, 2016.

[56] V. Gurarie. Single-particle Green's functions and interacting topological insulators. *Phys. Rev. B*, 83:085426, Feb 2011.

[57] Grigory E. Volovik. *The Universe in a Helium Droplet*. Oxford University Press, 2009.

[58] D. C. Tsui, H. L. Stormer, and A. C. Gossard. Two-Dimensional Magnetotransport in the Extreme Quantum Limit. *Phys. Rev. Lett.*, 48:1559–1562, May 1982.

[59] R. B. Laughlin. Anomalous Quantum Hall Effect: An Incompressible Quantum Fluid with Fractionally Charged Excitations. *Phys. Rev. Lett.*, 50:1395–1398, May 1983.

[60] A.Yu. Kitaev. Fault-tolerant quantum computation by anyons. *Annals of Physics*, 303(1):2–30, 2003.

[61] Saptarshi Mandal and Arun M Jayannavar. An introduction to Kitaev model-I, 2020.

[62] Alexei Kitaev and Chris Laumann. Topological phases and quantum computation, 2009.

[63] C. L. Kane and E. J. Mele. Z_2 Topological Order and the Quantum Spin Hall Effect. *Phys. Rev. Lett.*, 95:146802, Sep 2005.

[64] Zhong Wang, Xiao-Liang Qi, and Shou-Cheng Zhang. Topological invariants for interacting topological insulators with inversion symmetry. *Phys. Rev. B*, 85:165126, Apr 2012.

[65] Zhong Wang and Shou-Cheng Zhang. Topological Invariants and Ground-State Wave functions of Topological Insulators on a Torus. *Phys. Rev. X*, 4:011006, Jan 2014.

[66] Stephan Rachel. Interacting topological insulators: a review. *Reports on Progress in Physics*, 81(11):116501, oct 2018.

[67] Jan Carl Budich and Björn Trauzettel. From the adiabatic theorem of quantum mechanics to topological states of matter. *physica status solidi (RRL) – Rapid Research Letters*, 7(1-2):109–129, 2013.

[68] Meng Guo, Pavel Putrov, and Juven Wang. Time reversal, SU(N) Yang–Mills and cobordisms: Interacting topological superconductors/insulators and quantum spin liquids in 3+1D. *Annals of Physics*, 394:244–293, 2018.

[69] Dmytro Pesin and Leon Balents. Mott physics and band topology in materials with strong spin-orbit interaction. *Nature Physics*, 6(5):376–381, May 2010.

[70] Koji Kudo, Tsuneya Yoshida, and Yasuhiro Hatsugai. Higher-Order Topological Mott Insulators. *Phys. Rev. Lett.*, 123:196402, Nov 2019.

[71] Takahiro Fukui, Yasuhiro Hatsugai, and Hiroshi Suzuki. Chern Numbers in Discretized Brillouin Zone: Efficient Method of Computing (Spin) Hall Conductances. *Journal of the Physical Society of Japan*, 74(6):1674–1677, 2005.

[72] Yasuhiro Hatsugai. Explicit Gauge Fixing for Degenerate Multiplets: A Generic Setup for Topological Orders. *Journal of the Physical Society of Japan*, 73(10):2604–2607, 2004.

[73] Yasuhiro Hatsugai. Characterization of Topological Insulators: Chern Numbers for Ground State Multiplet. *Journal of the Physical Society of Japan*, 74(5):1374–1377, 2005.

[74] Gregory H. Wannier. The Structure of Electronic Excitation Levels in Insulating Crystals. *Phys. Rev.*, 52:191–197, Aug 1937.

[75] Gregory H. Wannier. Dynamics of Band Electrons in Electric and Magnetic Fields. *Rev. Mod. Phys.*, 34:645–655, Oct 1962.

[76] Nicola Marzari and David Vanderbilt. Maximally localized generalized Wannier functions for composite energy bands. *Phys. Rev. B*, 56:12847–12865, Nov 1997.

[77] 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.

[78] W. Kohn. Analytic Properties of Bloch Waves and Wannier Functions. *Phys. Rev.*, 115:809–821, Aug 1959.

[79] 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.

[80] Raffaele Resta. Quantum-Mechanical Position Operator in Extended Systems. *Phys. Rev. Lett.*, 80:1800–1803, Mar 1998.

[81] Raffaele Resta and David Vanderbilt. *Theory of Polarization: A Modern Approach*, pages 31–68. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[82] S. Kivelson. Wannier functions in one-dimensional disordered systems: Application to fractionally charged solitons. *Phys. Rev. B*, 26:4269–4277, Oct 1982.

[83] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

[84] John W. Negele and Henri Orland. *Quantum Many-Particle Systems*. CRC Press, 2018.

[85] Alexander Altland and Ben D. Simons. *Condensed Matter Field Theory*. Cambridge University Press, 2010.

[86] Alexander L. Fetter and John Dirk Walecka. *Quantum Theory of Many-Particle Systems*. Dover, 2003.

[87] J. Hubbard. Electron Correlations in Narrow Energy Bands. *Proceedings of the Royal Society of London Series A*, 276(1365):238–257, November 1963.

[88] Intel® 64 and IA-32 Architectures Software Developer's Manual. https://software.intel.com/content/www/us/en/develop/articles/intel-sdm.html, (retrieved: 02.07.2021).

[89] Jürgen Schnack, Jörg Schulenburg, and Johannes Richter. Magnetism of the $N = 42$ kagome lattice antiferromagnet. *Phys. Rev. B*, 98:094423, Sep 2018.

[90] Tjark Heitmann and Jürgen Schnack. Combined use of translational and spin-rotational invariance for spin systems. *Phys. Rev. B*, 99:134405, Apr 2019.

[91] Sean Eron Anderson. Bit Twiddling Hacks. https://graphics.stanford.edu/~seander/bithacks.html, (retrieved: 15.06.2021).

[92] GCC online docs: Other Built-in Functions Provided by GCC. https://gcc.gnu.org/onlinedocs/gcc/Other-Builtins.html, (retrieved: 02.07.2021).

[93] Intel® Intrinsics Guide. https://software.intel.com/sites/landingpage/IntrinsicsGuide/#text=_popcnt64, (retrieved: 02.07.2021).

[94] J. G. F. Francis. The QR Transformation A Unitary Analogue to the LR Transformation—Part 1. *The Computer Journal*, 4(3):265–271, 01 1961.

[95] J. G. F. Francis. The QR Transformation—Part 2. *The Computer Journal*, 4(4):332–345, 01 1962.

[96] V.N. Kublanovskaya. On some algorithms for the solution of the complete eigenvalue problem. *USSR Computational Mathematics and Mathematical Physics*, 1(3):637 – 657, 1962.

[97] LAPACK - Linear Algebra PACKage. https://www.netlib.org/lapack. (retrieved: 19.10.2020).

[98] Alston S. Householder. Unitary Triangularization of a Nonsymmetric Matrix. *Journal of the Association for Computing Machinery*, 5:4, 1958.

[99] J. M. Ortega and H. F. Kaiser. The LLT and QR methods for symmetric tridiagonal matrices. *The Computer Journal*, 6(1):99–101, 01 1963.

[100] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[101] R. B. Lehoucq, D. C. Sorensen, and C. Yang. ARPACK User's Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods, 1997. https://www.caam.rice.edu/software/ARPACK/UG/ug.html, (retrieved: 10.06.2021).

[102] Cornelius Lanczos. An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators. *Journal of Research of the National Bureau of Standards*, 45(4), 1950.

[103] C. C. PAIGE. Computational Variants of the Lanczos Method for the Eigenproblem. *IMA Journal of Applied Mathematics*, 10(3):373–381, 12 1972.

[104] J. Jaklič and P. Prelovšek. Lanczos method for the calculation of finite-temperature quantities in correlated systems. *Phys. Rev. B*, 49:5065–5068, Feb 1994.

[105] J. Schnack and O. Wendland. Properties of highly frustrated magnetic molecules studied by the finite-temperature Lanczos method. *European Physical Journal B*, 78(4):535–541, December 2010.

[106] Claudius Gros and Roser Valentí. Cluster expansion for the self-energy: A simple many-body method for interpreting the photoemission spectra of correlated Fermi systems. *Phys. Rev. B*, 48:418–425, Jul 1993.

[107] D. Sénéchal, D. Perez, and M. Pioro-Ladrière. Spectral Weight of the Hubbard Model through Cluster Perturbation Theory. *Phys. Rev. Lett.*, 84:522–525, Jan 2000.

[108] David Sénéchal, Danny Perez, and Dany Plouffe. Cluster perturbation theory for Hubbard models. *Phys. Rev. B*, 66:075129, Aug 2002.

[109] David Sénéchal. Cluster Perturbation Theory. In A. Avella and F. Mancini, editors, *Strongly Correlated Systems*, volume 171 of *Springer Series in Solid-State Sciences*. Springer, 2012.

[110] Thomas Maier, Mark Jarrell, Thomas Pruschke, and Matthias H. Hettler. Quantum cluster theories. *Rev. Mod. Phys.*, 77:1027–1080, Oct 2005.

[111] M. Potthoff. Self-energy-functional approach to systems of correlated electrons. *European Physical Journal B*, 32(4):429–436, April 2003.

[112] Michael Potthoff and Matthias Balzer. Self-energy-functional theory for systems of interacting electrons with disorder. *Phys. Rev. B*, 75:125112, Mar 2007.

[113] F. Manghi. Multi-orbital Cluster Perturbation Theory for transition metal oxides, 2013.

[114] F Grandi, F Manghi, O Corradini, C M Bertoni, and A Bonini. Topological invariants in interacting quantum spin Hall: a cluster perturbation theory approach. *New Journal of Physics*, 17(2):023004, jan 2015.

[115] F. Manghi. Correlated electrons in a crystalline topological insulator. *Phys. Rev. B*, 103:115114, Mar 2021.

[116] Stéphane Pairault, David Sénéchal, and A.-M. S. Tremblay. Strong-Coupling Expansion for the Hubbard Model. *Phys. Rev. Lett.*, 80:5389–5392, Jun 1998.

[117] Pairault, S., Sénéchal, D., and A.-M.S. Tremblay. Strong-coupling perturbation theory of the Hubbard model. *Eur. Phys. J. B*, 16(1):85–105, 2000.

[118] Walter Metzner and Dieter Vollhardt. Correlated Lattice Fermions in $d = \infty$ Dimensions. *Phys. Rev. Lett.*, 62:324–327, Jan 1989.

[119] Antoine Georges and Gabriel Kotliar. Hubbard model in infinite dimensions. *Phys. Rev. B*, 45:6479–6483, Mar 1992.

[120] Antoine Georges, Gabriel Kotliar, and Werner Krauth. Superconductivity in the two-band Hubbard model in infinite dimensions. *Zeitschrift fur Physik B Condensed Matter*, 92(3):313–321, September 1993.

[121] U. Brandt and C. Mielsch. Thermodynamics and correlation functions of the Falicov-Kimball model in large dimensions. *Zeitschrift fur Physik B Condensed Matter*, 75(3):365–370, September 1989.

[122] P. W. Anderson. Localized Magnetic States in Metals. *Phys. Rev.*, 124:41–53, Oct 1961.

[123] Emanuel Gull, Andrew J. Millis, Alexander I. Lichtenstein, Alexey N. Rubtsov, Matthias Troyer, and Philipp Werner. Continuous-time Monte Carlo methods for quantum impurity models. *Rev. Mod. Phys.*, 83:349–404, May 2011.

[124] Antoine Georges, Gabriel Kotliar, Werner Krauth, and Marcelo J. Rozenberg. Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions. *Rev. Mod. Phys.*, 68:13–125, Jan 1996.

[125] Gabriel Kotliar and Dieter Vollhardt. Strongly Correlated Materials: Insights From Dynamical Mean-Field Theory. *Physics Today*, 57(3):53–59, March 2004.

[126] Antoine Georges. Strongly Correlated Electron Materials: Dynamical Mean-Field Theory and Electronic Structure. In Adolfo Avella and Ferdinando Mancini, editors, *Lectures on the Physics of Highly Correlated Electron Systems VIII: Eighth Training Course in the Physics of Correlated Electron Systems and High-Tc Superconductors*, volume 715 of *American Institute of Physics Conference Series*, pages 3–74, August 2004.

[127] Eva Pavarini, Erik Koch, Dieter Vollhardt, and Alexander Lichtenstein, editors. *DMFT at 25: Infinite Dimensions*. Verlag des Forschungszentrum Jülich, 2014.

[128] Dieter Vollhardt. *Dynamical Mean-Field Theory of Strongly Correlated Electron Systems*.

[129] A. N. Kolmogorov. *Foundations of the Theory of Probability*. Dover, 1998.

[130] George Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series. Brooks/Cole Publishing Company, 1990.

[131] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[132] David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[133] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[134] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.

[135] Thomas Mertz, Karim Zantout, and Roser Valentí. Self-energy dispersion in the Hubbard model. *Phys. Rev. B*, 98:235105, Dec 2018.

[136] Masatoshi Imada, Atsushi Fujimori, and Yoshinori Tokura. Metal-insulator transitions. *Rev. Mod. Phys.*, 70:1039–1263, Oct 1998.

[137] Daniel Cocks, Peter P. Orth, Stephan Rachel, Michael Buchhold, Karyn Le Hur, and Walter Hofstetter. Time-Reversal-Invariant Hofstadter-Hubbard Model with Ultracold Fermions. *Phys. Rev. Lett.*, 109:205303, Nov 2012.

[138] Jan Carl Budich, Ronny Thomale, Gang Li, Manuel Laubach, and Shou-Cheng Zhang. Fluctuation-induced topological quantum phase transitions in quantum spin-Hall and anomalous-Hall insulators. *Phys. Rev. B*, 86:201407, Nov 2012.

[139] Jan Carl Budich, Björn Trauzettel, and Giorgio Sangiovanni. Fluctuation-driven topological Hund insulators. *Phys. Rev. B*, 87:235104, Jun 2013.

[140] Tuomas I. Vanhala, Topi Siro, Long Liang, Matthias Troyer, Ari Harju, and Päivi Törmä. Topological Phase Transitions in the Repulsively Interacting Haldane-Hubbard Model. *Phys. Rev. Lett.*, 116:225305, Jun 2016.

[141] Pramod Kumar, Thomas Mertz, and Walter Hofstetter. Interaction-induced topological and magnetic phases in the Hofstadter-Hubbard model. *Phys. Rev. B*, 94:115161, Sep 2016.

[142] Bernhard Irsigler, Tobias Grass, Jun-Hui Zheng, Mathieu Barbier, and Walter Hofstetter. Topological Mott transition in a Weyl-Hubbard model: Dynamical mean-field theory study. *Phys. Rev. B*, 103:125132, Mar 2021.

[143] F. D. M. Haldane. Model for a Quantum Hall Effect without Landau Levels: Condensed-Matter Realization of the "Parity Anomaly". *Phys. Rev. Lett.*, 61:2015–2018, Oct 1988.

[144] B. Andrei Bernevig, Taylor L. Hughes, and Shou-Cheng Zhang. Quantum Spin Hall Effect and Topological Phase Transition in HgTe Quantum Wells. *Science*, 314(5806):1757–1761, 2006.

[145] J. P. F. LeBlanc, Andrey E. Antipov, Federico Becca, Ireneusz W. Bulik, Garnet Kin-Lic Chan, Chia-Min Chung, Youjin Deng, Michel Ferrero, Thomas M. Henderson, Carlos A. Jiménez-Hoyos, E. Kozik, Xuan-Wen Liu, Andrew J. Millis, N. V. Prokof'ev, Mingpu Qin, Gustavo E. Scuseria, Hao Shi, B. V. Svistunov, Luca F. Tocchio, I. S. Tupitsyn, Steven R. White, Shiwei Zhang, Bo-Xiao Zheng, Zhenyue Zhu, and Emanuel Gull. Solutions of the Two-Dimensional Hubbard Model: Benchmarks and Results from a Wide Range of Numerical Algorithms. *Phys. Rev. X*, 5:041041, Dec 2015.

[146] N. Paris, K. Bouadim, F. Hebert, G. G. Batrouni, and R. T. Scalettar. Quantum Monte Carlo Study of an Interaction-Driven Band-Insulator–to–Metal Transition. *Phys. Rev. Lett.*, 98:046403, Jan 2007.

[147] J. Hubbard and J. B. Torrance. Model of the Neutral-Ionic Phase Transformation. *Phys. Rev. Lett.*, 47:1750–1754, Dec 1981.

[148] R. Resta and S. Sorella. Many-Body Effects on Polarization and Dynamical Charges in a Partly Covalent Polar Insulator. *Phys. Rev. Lett.*, 74:4738–4741, Jun 1995.

[149] Raffaele Resta and Sandro Sorella. Electron Localization in the Insulating State. *Phys. Rev. Lett.*, 82:370–373, Jan 1999.

[150] Michele Fabrizio, Alexander O. Gogolin, and Alexander A. Nersesyan. From Band Insulator to Mott Insulator in One Dimension. *Phys. Rev. Lett.*, 83:2014–2017, Sep 1999.

[151] C. D. Batista and A. A. Aligia. Exact Bond Ordered Ground State for the Transition between the Band and the Mott Insulator. *Phys. Rev. Lett.*, 92:246405, Jun 2004.

[152] K. Bouadim, N. Paris, F. Hébert, G. G. Batrouni, and R. T. Scalettar. Metallic phase in the two-dimensional ionic Hubbard model. *Phys. Rev. B*, 76:085112, Aug 2007.

[153] Aaram J. Kim, M. Y. Choi, and Gun Sang Jeon. Finite-temperature phase transitions in the ionic Hubbard model. *Phys. Rev. B*, 89:165117, Apr 2014.

[154] Soumen Bag, Arti Garg, and H. R. Krishnamurthy. Phase diagram of the half-filled ionic Hubbard model. *Phys. Rev. B*, 91:235108, Jun 2015.

[155] Mohsen Hafez-Torbati and Götz S. Uhrig. Orientational bond and Néel order in the two-dimensional ionic Hubbard model. *Phys. Rev. B*, 93:195128, May 2016.

[156] Abhisek Samanta and Rajdeep Sensarma. Superconductivity from doublon condensation in the ionic Hubbard model. *Phys. Rev. B*, 94:224517, Dec 2016.

[157] T. Miyao. Existence of charge-density waves in two-dimensional ionic Hubbard model. *ArXiv e-prints*. arXiv:1606.04197 (2016).

[158] Karla Loida, Jean-Sébastien Bernier, Roberta Citro, Edmond Orignac, and Corinna Kollath. Probing the Bond Order Wave Phase Transitions of the Ionic Hubbard Model by Superlattice Modulation Spectroscopy. *Phys. Rev. Lett.*, 119:230403, Dec 2017.

[159] G. Rohringer, H. Hafermann, A. Toschi, A. A. Katanin, A. E. Antipov, M. I. Katsnelson, A. I. Lichtenstein, A. N. Rubtsov, and K. Held. Diagrammatic routes to nonlocal correlations beyond dynamical mean field theory. *Rev. Mod. Phys.*, 90:025003, May 2018.

[160] Y.M. Vilk, Liang Chen, and A.-M.S. Tremblay. Two-particle self-consistent theory for spin and charge fluctuations in the Hubbard model. *Physica C: Superconductivity*, 235-240:2235–2236, 1994.

[161] Y.M. Vilk and A.-M.S. Tremblay. Non-Perturbative Many-Body Approach to the Hubbard Model and Single-Particle Pseudogap. *J. Phys. I France*, 7(11):1309–1368, 1997.

[162] Karim Zantout, Steffen Backes, and Roser Valentí. Two-Particle Self-Consistent Method for the Multi-Orbital Hubbard Model. *Annalen der Physik*, n/a(n/a):2000399, 2021.

[163] Philipp Werner, Armin Comanac, Luca de' Medici, Matthias Troyer, and Andrew J. Millis. Continuous-Time Solver for Quantum Impurity Models. *Phys. Rev. Lett.*, 97:076405, Aug 2006.

[164] Hartmut Hafermann, Philipp Werner, and Emanuel Gull. Efficient implementation of the continuous-time hybridization expansion quantum impurity solver. *Computer Physics Communications*, 184(4):1280 – 1286, 2013.

[165] A. N. Rubtsov, M. I. Katsnelson, and A. I. Lichtenstein. Dual fermion approach to nonlocal correlations in the Hubbard model. *Phys. Rev. B*, 77:033101, Jan 2008.

[166] A. N. Rubtsov, M. I. Katsnelson, A. I. Lichtenstein, and A. Georges. Dual fermion approach to the two-dimensional Hubbard model: Antiferromagnetic fluctuations and Fermi arcs. *Phys. Rev. B*, 79:045133, Jan 2009.

[167] B Bauer, L D Carr, H G Evertz, A Feiguin, J Freire, S Fuchs, L Gamper, J Gukelberger, E Gull, S Guertler, A Hehn, R Igarashi, S V Isakov, D Koop, P N Ma, P Mates, H Matsuo, O Parcollet, G Pawłowski, J D Picon, L Pollet, E Santos, V W Scarola, U Schollwöck, C Silva, B Surer, S Todo, S Trebst, M Troyer, M L Wall, P Werner, and S Wessel. The ALPS project release 2.0: open source software for strongly correlated systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(05):P05001, may 2011.

[168] Emanuel Gull, Philipp Werner, Sebastian Fuchs, Brigitte Surer, Thomas Pruschke, and Matthias Troyer. Continuous-time quantum Monte Carlo impurity solvers. *Computer Physics Communications*, 182(4):1078–1082, 2011.

[169] Thomas Mertz, Karim Zantout, and Roser Valentí. Statistical analysis of the Chern number in the interacting Haldane-Hubbard model. *Phys. Rev. B*, 100:125111, Sep 2019.

[170] Gregor Jotzu, Michael Messer, Rémi Desbuquois, Martin Lebrat, Thomas Uehlinger, Daniel Greif, and Tilman Esslinger. Experimental realization of the topological Haldane model with ultracold fermions. *Nature*, 515(7526):237–240, November 2014.

[171] Wilfrido A. Gómez-Arias and Gerardo G. Naumis. Analytical calculation of electron group velocity surfaces in uniform strained graphene. *International Journal of Modern Physics B*, 30(3):1550263, December 2016.

[172] Gordon W. Semenoff. Condensed-Matter Simulation of a Three-Dimensional Anomaly. *Phys. Rev. Lett.*, 53:2449–2452, Dec 1984.

[173] Thomas Mertz and Roser Valentí. Engineering topological phases guided by statistical and machine learning methods. *Phys. Rev. Research*, 3:013132, Feb 2021.

[174] Jakub Imriška, Lei Wang, and Matthias Troyer. First-order topological phase transition of the Haldane-Hubbard model. *Phys. Rev. B*, 94:035109, Jul 2016.

[175] Igor S. Tupitsyn and Nikolay V. Prokof'ev. Phase diagram topology of the Haldane-Hubbard-Coulomb model. *Phys. Rev. B*, 99:121113, Mar 2019.

[176] Alexander Lau, Carmine Ortix, and Jeroen van den Brink. Topological Edge States with Zero Hall Conductivity in a Dimerized Hofstadter Model. *Phys. Rev. Lett.*, 115:216805, Nov 2015.

[177] N. E. Bickers, D. J. Scalapino, and S. R. White. Conserving Approximations for Strongly Correlated Electron Systems: Bethe-Salpeter Equation and Dynamics for the Two-Dimensional Hubbard Model. *Phys. Rev. Lett.*, 62:961–964, Feb 1989.

[178] Markus König, Steffen Wiedmann, Christoph Brüne, Andreas Roth, Hartmut Buhmann, Laurens W. Molenkamp, Xiao-Liang Qi, and Shou-Cheng Zhang. Quantum Spin Hall Insulator State in HgTe Quantum Wells. *Science*, 318(5851):766–770, 2007.

[179] Frank Schindler, Zhijun Wang, Maia G. Vergniory, Ashley M. Cook, Anil Murani, Shamashis Sengupta, Alik Yu. Kasumov, Richard Deblock, Sangjun Jeon, Ilya Drozdov, Hélène Bouchiat, Sophie Guéron, Ali Yazdani, B. Andrei Bernevig, and Titus Neupert. Higher-order topology in bismuth. *Nature Physics*, 14(9):918–924, July 2018.

[180] Juan Carrasquilla. Machine learning for quantum matter. *Advances in Physics: X*, 5(1):1797528, 2020.

[181] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[182] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. Software available from tensorflow.org.

[183] Li Huang and Lei Wang. Accelerated Monte Carlo simulations with restricted Boltzmann machines. *Phys. Rev. B*, 95:035105, Jan 2017.

[184] Xiao Yan Xu, Yang Qi, Junwei Liu, Liang Fu, and Zi Yang Meng. Self-learning quantum Monte Carlo method in interacting fermion systems. *Phys. Rev. B*, 96:041119, Jul 2017.

[185] Peter Broecker, Juan Carrasquilla, Roger G. Melko, and Simon Trebst. Machine learning quantum phases of matter beyond the fermion sign problem. *Scientific Reports*, 7:8823, August 2017.

[186] Sirui Lu, Shilin Huang, Keren Li, Jun Li, Jianxin Chen, Dawei Lu, Zhengfeng Ji, Yi Shen, Duanlu Zhou, and Bei Zeng. Separability-entanglement classifier via machine learning. *Phys. Rev. A*, 98:012315, Jul 2018.

[187] Ryosuke Jinnouchi, Jonathan Lahnsteiner, Ferenc Karsai, Georg Kresse, and Menno Bokdam. Phase Transitions of Hybrid Perovskites Simulated by Machine-Learning Force Fields Trained on the Fly with Bayesian Inference. *Phys. Rev. Lett.*, 122:225701, Jun 2019.

[188] Ryosuke Jinnouchi, Ferenc Karsai, and Georg Kresse. On-the-fly machine learning force field generation: Application to melting points. *Phys. Rev. B*, 100:014105, Jul 2019.

[189] S. Pilati, E. M. Inack, and P. Pieri. Self-learning projective quantum Monte Carlo simulations guided by restricted Boltzmann machines. *Phys. Rev. E*, 100:043301, Oct 2019.

[190] Ryo Nagai, Ryosuke Akashi, and Osamu Sugino. Completing density functional theory by machine learning hidden messages from molecules. *npj Computational Mathematics*, 6:43, January 2020.

[191] Taegeun Song, Roser Valenti, and Hunpyo Lee. Analytic continuation of the self-energy via Machine Learning techniques, 2020.

[192] M. Michael Denner, Mark H. Fischer, and Titus Neupert. Efficient learning of a one-dimensional density functional theory. *Phys. Rev. Research*, 2:033388, Sep 2020.

[193] Kenny Choo, Giuseppe Carleo, Nicolas Regnault, and Titus Neupert. Symmetries and Many-Body Excitations with Neural-Network Quantum States. *Phys. Rev. Lett.*, 121:167204, Oct 2018.

[194] Lei Wang. Discovering phase transitions with unsupervised learning. *Phys. Rev. B*, 94:195105, Nov 2016.

[195] Matthew J. S. Beach, Anna Golubeva, and Roger G. Melko. Machine learning vortices at the Kosterlitz-Thouless transition. *Phys. Rev. B*, 97:045207, Jan 2018.

[196] Joaquin F. Rodriguez-Nieva and Mathias S. Scheurer. Identifying topological order through unsupervised machine learning. *Nature Physics*, 15(8):790–795, May 2019.

[197] Kenta Shiina, Hiroyuki Mori, Yutaka Okabe, and Hwee Kuan Lee. Machine-Learning Studies on Spin Models. *Scientific Reports*, 10:2177, February 2020.

[198] Mathias S. Scheurer and Robert-Jan Slager. Unsupervised Machine Learning and Band Topology. *Phys. Rev. Lett.*, 124:226401, Jun 2020.

[199] Pengfei Zhang, Huitao Shen, and Hui Zhai. Machine Learning Topological Invariants with Neural Networks. *Phys. Rev. Lett.*, 120:066401, Feb 2018.

[200] Ning Sun, Jinmin Yi, Pengfei Zhang, Huitao Shen, and Hui Zhai. Deep learning topological invariants of band insulators. *Phys. Rev. B*, 98:085402, Aug 2018.

[201] Zixian Su, Yanzhuo Kang, Bofeng Zhang, Zhiqiang Zhang, and Hua Jiang. Disorder induced phase transition in magnetic higher-order topological insulator: A machine learning study. *Chinese Physics B*, 28(11):117301, oct 2019.

[202] Yurui Ming, Chin-Teng Lin, Stephen D. Bartlett, and Wei-Wei Zhang. Quantum topology identification with deep neural networks and quantum walks. *npj Computational Mathematics*, 5:88, August 2019.

[203] Benno S. Rem, Niklas Käming, Matthias Tarnowski, Luca Asteria, Nick Fläschner, Christoph Becker, Klaus Sengstock, and Christof Weitenberg. Identifying quantum phase transitions using artificial neural networks on experimental data. *Nature Physics*, 15(9):917–920, July 2019.

[204] Marcello D. Caio, Marco Caccin, Paul Baireuther, Timo Hyart, and Michel Fruchart. Machine learning assisted measurement of local topological invariants. *arXiv e-prints*, page arXiv:1901.03346, January 2019.

[205] Hoi Chun Po, Ashvin Vishwanath, and Haruki Watanabe. Complete theory of symmetry-based indicators of band topology. *Nature Communications*, 8:50, June 2017.

[206] Christie S. Chiu, Da-Shuai Ma, Zhi-Da Song, B. Andrei Bernevig, and Andrew A. Houck. Fragile topology in line-graph lattices with two, three, or four gapped flat bands. *Phys. Rev. Research*, 2:043414, Dec 2020.

[207] Armin Sahinovic and Benjamin Geisler. Active learning and element embedding approach in neural networks for infinite-layer versus perovskite oxides, 2021.

[208] Tiantian Zhang, Yi Jiang, Zhida Song, He Huang, Yuqing He, Zhong Fang, Hongming Weng, and Chen Fang. Catalogue of topological electronic materials. *Nature*, 566(7745):475–479, February 2019.

[209] Nikolas Claussen, B. Andrei Bernevig, and Nicolas Regnault. Detection of topological materials with machine learning. *Phys. Rev. B*, 101:245117, Jun 2020.

[210] Yi Jiang, Dong Chen, Xin Chen, Tangyi Li, Guo-Wei Wei, and Feng Pan. Topological representations of crystalline compounds for the machine-learning prediction of materials properties. *npj Computational Mathematics*, 7:28, January 2021.

[211] Kamal Choudhary, Kevin F. Garrity, Nirmal J. Ghimire, Naween Anand, and Francesca Tavazza. High-throughput search for magnetic topological materials using spin-orbit spillage, machine learning, and experiments. *Phys. Rev. B*, 103:155131, Apr 2021.

[212] Evert P. L. van Nieuwenburg, Ye-Hua Liu, and Sebastian D. Huber. Learning phase transitions by confusion. *Nature Physics*, 13(5):435–439, February 2017.

[213] Dong-Ling Deng, Xiaopeng Li, and S. Das Sarma. Machine learning topological states. *Phys. Rev. B*, 96:195145, Nov 2017.

[214] Wenqian Lian, Sheng-Tao Wang, Sirui Lu, Yuanyuan Huang, Fei Wang, Xinxing Yuan, Wengang Zhang, Xiaolong Ouyang, Xin Wang, Xianzhi Huang, Li He, Xiuying Chang, Dong-Ling Deng, and Luming Duan. Machine Learning Topological Phases with a Solid-State Quantum Simulator. *Phys. Rev. Lett.*, 122:210503, May 2019.

[215] Oleksandr Balabanov and Mats Granath. Unsupervised learning using topological data augmentation. *Phys. Rev. Research*, 2:013354, Mar 2020.

[216] Eliska Greplova, Agnes Valenti, Gregor Boschung, Frank Schäfer, Niels Lörch, and Sebastian D Huber. Unsupervised identification of topological phase transitions using predictive models. *New Journal of Physics*, 22(4):045003, apr 2020.

[217] Yanming Che, Clemens Gneiting, Tao Liu, and Franco Nori. Topological quantum phase transitions retrieved through unsupervised machine learning. *Phys. Rev. B*, 102:134213, Oct 2020.

[218] Cyrill Bösch, Tena Dubček, Frank Schindler, Andreas Fichtner, and Marc Serra-Garcia. Discovery of topological metamaterials by symmetry relaxation and smooth topological indicators. *Phys. Rev. B*, 102:241404, Dec 2020.

[219] Vittorio Peano, Florian Sapper, and Florian Marquardt. Rapid Exploration of Topological Band Structures Using Deep Learning. *Phys. Rev. X*, 11:021052, Jun 2021.

[220] Douglas Adams. *The Hitchhiker's Guide to the Galaxy*. Pan Books, London, 1979.

[221] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! Criticism for Interpretability. *Advances in Neural Information Processing Systems 29*, 29:2280–2288, 2016.

[222] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. February 2017.

[223] Tim Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. June 2017.

[224] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of Neural Networks is Fragile. October 2017.

[225] Christoph Molnar. *Interpretable Machine Learning*. Online Publication, 2021. `https://christophm.github.io/interpretable-ml-book/`.

[226] Mario Krenn and Anton Zeilinger. Predicting research trends with semantic and neural networks with an application in quantum physics. *Proceedings of the National Academy of Sciences*, 117(4):1910–1916, 2020.

[227] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990.

[228] Michael W. Berry. *Survey of Text Mining*. Springer, New York, 2004.

[229] Michael W. Berry. *Survey of Text Mining II*. Springer, London, 2008.

[230] Harold Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321–377, 1936.

[231] Karl Pearson F.R.S. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[232] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991.

[233] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[234] N. L. Holanda and M. A. R. Griffith. Machine learning topological phases in real space. *Phys. Rev. B*, 102:054107, Aug 2020.

[235] J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 5.1:281–297, 1967.

[236] J. A. Hartigan and M. A. Wong. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[237] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[238] W. K. Hastings. Monte Carlo Sampling Methods using Markov Chains and their Applications. *Biometrika*, 57(1):97–109, April 1970.

[239] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[240] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Trans. Database Syst.*, 42(3), July 2017.

[241] Nathan Argaman and Guy Makov. Density functional theory: An introduction. *American Journal of Physics*, 68(1):69–79, 2000.

[242] E. Engel and R.M. Dreizler. *Density Functional Theory: An Advanced Course*. Theoretical and Mathematical Physics. Springer Berlin Heidelberg, 2011.

[243] X. Wu, V. Kumar, and J. Ross Quinlan et al. Top 10 algorithms in data mining. *Knowl Inf Syst*, 14:1–37, 2008.

[244] L. Breiman, J.H. Friedman, R.A. Olshen, and Stone C.J. Classification and Regression Trees. *Biometrics*, 40(3):874–874, 1984.

[245] Quinlan J.R. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[246] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[247] Gareth James, Daniela Witten, Trevor Hastie, and Rob Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

[248] Brian D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[249] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[250] K.G. Murty and S.N. Kapadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39:117–129, 1987.

[251] Maw-Sheng Chern. On the computational complexity of reliability redundancy allocation in a series system. *Operations Research Letters*, 11(5):309–315, 1992.

[252] Yi-Chih Hsieh, Yung-Cheng Lee, and Peng-Sheng You. Solving nonlinear constrained optimization problems: An immune evolutionary based two-phase approach. *Applied Mathematical Modelling*, 39(19):5759–5768, 2015.

[253] Igor V. Tetko, David J. Livingstone, and Alexander I. Luik. Neural network studies. 1. Comparison of overfitting and overtraining. *Journal of Chemical Information and Computer Sciences*, 35(5):826–833, 1995.

[254] K.P. Burnham and D.R. Anderson. *Model Selection and Multimodel Inference*. Springer, 2002.

[255] Corrado Gini. *Variabilità e Mutuabilità*. Bologna, 1912.

[256] P.M. Dixon, J. Weiner, T. Mitchell-Olds, and R. Woodley. Bootstrapping the Gini Coefficient of Inequality. *Ecology*, 68(5):1548–1551, 1987.

[257] Lidia Ceriani and Paolo Verme. The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *The Journal of Economic Inequality*, 10:421–443, 2012.

[258] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

[259] L. Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001.

[260] Kendall E. Atkinson. *An introduction to numerical analysis*. Wiley, 2nd edition, 1989.

[261] M.E.J. Newman and G.T. Barkema. *Monte Carlo Methods in Statistical Physics*. Clarendon Press, 1999.

[262] Solomon Kullback. *Information Theory and Statistics*. Dover, 1968.

[263] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1942.

[264] A. Bhattacharyya. On a Measure of Divergence between Two Multinomial Populations. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 7(4):401–406, 1946.

[265] T. Kailath. The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Transactions on Communication Technology*, 15(1):52–60, 1967.

[266] Encyclopedia of Mathematics. Hellinger distance. `http://encyclopediaofmath.org/index.php?title=Hellinger_distance&oldid=47206`.

[267] Leandro Pardo. *Statistical Inference Based on Divergence Measures*. Chapman and Hall/CRC, 2005.

[268] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians, 2020.

[269] Alfréd Rényi. On measures of entropy and information. *The 4th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561, 1960.

[270] I. Csiszár. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2:191–213, 1972.

[271] Y.A. Rozanov. *Probability Theory: A Concise Course*. Dover, 1977.

[272] S. Watanabe. Information Theoretical Analysis of Multivariate Correlation. *IBM Journal of Research and Development*, 4(1):66–82, 1960.

[273] Jianji Wang and Nanning Zheng. Measures of Correlation for Multiple Variables. *arXiv e-prints*, page arXiv:1401.4827, January 2014.

[274] Hoang Vu Nguyen, Emmanuel Müller, Jilles Vreeken, Pavel Efros, and Klemens Böhm. Multivariate Maximal Correlation Analysis. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 775–783, Bejing, China, 22–24 Jun 2014. PMLR.

[275] Yisen Wang, Simone Romano, Vinh Nguyen, James Bailey, Xingjun Ma, and Shu-Tao Xia. Unbiased Multivariate Correlation Analysis. 2017.

[276] Brenden R. Ortiz, Lídia C. Gomes, Jennifer R. Morey, Michal Winiarski, Mitchell Bordelon, John S. Mangum, Iain W. H. Oswald, Jose A. Rodriguez-Rivera, James R. Neilson, Stephen D. Wilson, Elif Ertekin, Tyrel M. McQueen, and Eric S. Toberer. New kagome prototype materials: discovery of KV_3Sb_5, RbV_3Sb_5, and CsV_3Sb_5. *Phys. Rev. Materials*, 3:094407, Sep 2019.

[277] Yu-Xiao Jiang, Jia-Xin Yin, M. Michael Denner, Nana Shumiya, Brenden R. Ortiz, Gang Xu, Zurab Guguchia, Junyi He, Md Shafayat Hossain, Xiaoxiong Liu, Jacob Ruff, Linus Kautzsch, Songtian S. Zhang, Guoqing Chang, Ilya Belopolski, Qi Zhang, Tyler A. Cochran, Daniel Multer, Maksim Litskevich, Zi-Jia Cheng, Xian P. Yang, Ziqiang Wang, Ronny Thomale, Titus Neupert, Stephen D. Wilson, and M. Zahid Hasan. Unconventional chiral charge order in kagome superconductor KV3Sb5. *Nature Materials*, Jun 2021.

[278] M. Michael Denner, Ronny Thomale, and Titus Neupert. Analysis of charge order in the kagome metal AV_3Sb_5 (A =K,Rb,Cs). *arXiv e-prints*, page arXiv:2103.14045, March 2021.

[279] Hengxin Tan, Yizhou Liu, Ziqiang Wang, and Binghai Yan. Charge density waves and electronic properties of superconducting kagome metals, 2021.

[280] E. Uykur, B. R. Ortiz, S. D. Wilson, M. Dressel, and A. A. Tsirlin. Optical detection of charge-density-wave instability in the non-magnetic kagome metal KV_3Sb_5, 2021.

[281] Feng Du, Shuaishuai Luo, Brenden R. Ortiz, Ye Chen, Weiyin Duan, Dongting Zhang, Xin Lu, Stephen D. Wilson, Yu Song, and Huiqiu Yuan. Pressure-induced double superconducting domes and charge instability in the kagome metal KV3Sb5. *Physical Review B*, 103(22), Jun 2021.

[282] F. H. Yu, T. Wu, Z. Y. Wang, B. Lei, W. Z. Zhuo, J. J. Ying, and X. H. Chen. Concurrence of anomalous Hall effect and charge density wave in a superconducting topological kagome metal, 2021.

[283] Declan Davis. Dihedral Group D_6. MathWorld–A Wolfram Web Resource, created by Eric W. Weisstein. `https://mathworld.wolfram.com/DihedralGroupD6.html` (retrieved: 27.07.2021).

[284] Giovanni Pizzi, Valerio Vitale, Ryotaro Arita, Stefan Blügel, Frank Freimuth, Guillaume Géranton, Marco Gibertini, Dominik Gresch, Charles Johnson, Takashi Koretsune, Julen Ibañez-Azpiroz, Hyungjun Lee, Jae-Mo Lihm, Daniel Marchand, Antimo Marrazzo, Yuriy Mokrousov, Jamal I Mustafa, Yoshiro Nohara, Yusuke Nomura, Lorenzo Paulatto, Samuel Poncé, Thomas Ponweiser, Junfeng Qiao, Florian Thöle, Stepan S Tsirkin, Małgorzata Wierzbowska, Nicola Marzari, David Vanderbilt, Ivo Souza, Arash A Mostofi, and Jonathan R Yates. Wannier90 as a community code: new features and applications. *Journal of Physics: Condensed Matter*, 32(16):165902, jan 2020.

[285] E. Macías-Virgós and D. Mosquera-Lois. Homotopic distance between maps. *Mathematical Proceedings of the Cambridge Philosophical Society*, page 1–21, 2021.