Texts and Readings in Physical Sciences 19

Somendra Mohan Bhattacharjee
Mahan Mj
Abhijit Bandyopadhyay *Editors*

Topology and Condensed Matter Physics

HINDUSTAN BOOK AGENCY
Springer

Texts and Readings in Physical Sciences

Volume 19

The *Texts and Readings in Physical Sciences* series publishes high-quality textbooks, research-level monographs, lecture notes and contributed volumes. Undergraduate and graduate students of physical sciences and applied mathematics, research scholars and teachers would find this book series useful. The volumes are carefully written as teaching aids and highlight characteristic features of the theory. The books in this series are co-published with Hindustan Book Agency, New Delhi, India.

More information about this series at http://www.springer.com/series/15139

Somendra Mohan Bhattacharjee
Mahan Mj · Abhijit Bandyopadhyay
Editors

Topology and Condensed Matter Physics

HINDUSTAN BOOK AGENCY
Springer

Editors
Somendra Mohan Bhattacharjee
Institute of Physics
Bhubaneswar, Odisha
India

Mahan Mj
School of Mathematics
Tata Institute of Fundamental Research
Mumbai, Maharashtra
India

Abhijit Bandyopadhyay
Department of Physics
Ramakrishna Mission Vivekananda University
Howrah, West Bengal
India

This work is a co-publication with Hindustan Book Agency, New Delhi, licensed for sale in all countries in electronic form, in print form only outside of India. Sold and distributed in print within India by Hindustan Book Agency, P-19 Green Park Extension, New Delhi 110016, India. ISBN: 978-93-86279-66-8 © Hindustan Book Agency 2017.

ISSN 2366-8849 ISSN 2366-8857 (electronic)
Texts and Readings in Physical Sciences
ISBN 978-981-10-6840-9 ISBN 978-981-10-6841-6 (eBook)
https://doi.org/10.1007/978-981-10-6841-6

Library of Congress Control Number: 2017955364

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

About the Editors

Somendra Mohan Bhattacharjee is a senior professor at the Institute of Physics, Bhubaneswar, India. He received his Ph.D. degree from Carnegie Mellon University, U.S.A., in 1984. He is fellow of the Indian Academy of Sciences and the Indian National Science Academy, India. His research area is theoretical condensed matter physics, especially statistical physics.

Mahan Mj is a professor of mathematics at the Tata Institute of Fundamental Research, Mumbai, India, and a monk of the Ramakrishna Mission. Earlier, he was a professor of mathematics and dean of research at the Ramakrishna Mission Vivekananda University until 2015. He received his Ph.D. from University of California, Berkely, U.S.A. He is a recipient of the 2011 Shanti Swarup Bhatnagar Award in Mathematical Sciences and the Infosys Science Foundation Award 2015 for Mathematical Sciences.He specialises in geometric topology andhyperbolic geometry.

Abhijit Bandyopadhyay is an associate professor of Physics at Ramakrishna Mission Vivekananda University, Belur, Howrah, India. He specialises in high energy physics and cosmology.

Preface

Topology deals with the concept of continuity and properties unchanged under continuous deformations. In a real system, a continuous deformation, requiring small changes, would involve modes or excitations that cost low energies ($\to 0$). Since studies of low energy excitations are at the heart of traditional condensed matter physics (CMP), it is natural to expect that topology would play a role in the development of CMP. Curiously, it did not.

An early application of topology in CMP goes back to the use of Morse theory for van Hove singularities. Despite all its beauty and actions, it was not a hit – too much effort for too little gain. However, it did produce something of a ripple, which got amplified further by several disturbing results, like the Aharonov-Bohm effect in quantum mechanics, flux quantization in superconducting rings, the Abrikosov lattice in type-II superconductors, and so on. These hinted at a need to think afresh beyond calculus-based physics. While the new era saw topological ideas fruitfully exploited in the high energy physics context, the condensed matter physics community, on the other hand, remained hesitant. One had to wait for the Berezinski-Kosterlitz-Thouless transition involving vortices in a two dimensional xy magnet to appear on the scene, and then the use of homotopy theory for classification of defects in ordered media, to realize that topology could not be kept out of the reckoning. With the discovery of the quantum Hall effect and its various avatars, and their study via topological invariants, the language of topology became indispensable to CMP. The theory of topological insulators was the shot in the arm CMP was waiting for.

There was a parallel stream involving statistical mechanics. Although the idea of knots in ether as matter (or atoms), e.g., sodium as two linked rings because of the two D lines, got a quick burial in the nineteenth century, but still knot theory survived as a distinct field of topology. The discovery of knot and link structures in polymers and DNA by electron microscopy made studies of knots important in the chemical and biological domains. The study of integrable statistical mechanical models based on hydrogen bonded crystals, like ice, (vertex models, Yang-Baxter equation) in synergy with more modern developments in knot theory produced knot invariants and polynomials sought by the topologists. This was an example of physics helping mathematics. With braids as intertwined paths of two dimensional quantum particles, also called

directed polymers, braid groups led to exotic particle statistics, which are found in condensed matter systems. Such exotic nonabelian particles can be exploited in quantum computation, raising hopes of a secret rendezvous of Alice and Bob.

With so much to gain, topology needs to be an essential toolkit for physicists in general and condensed matter physicists in particular. But, to the best of our knowledge, the subject is not covered in standard physics courses. No doubt, there are authoritative books on topology and geometry in physics, but such are mostly for high energy physicists. It was in this context that a school was organized to introduce topology and its applications in condensed matter problems, at a pedagogic level, developed almost from scratch. This entailed introduction of topology with mathematical rigor while remaining accessible to physicists, side by side with the use of topology in quantum mechanics, statistical physics, and solid state physics.

At a time when mathematics is getting sidelined in physics curricula, the SERC school held in Kolkata in Nov-Dec 2015 was an attempt to halt, if not change, the trend. Moreover, this school tried to focus on a holistic approach to several frontier areas, bridging the language barrier among the disciplines of physics, and between physics and mathematics. Our ultimate hope is that a cross-fertilization will motivate newer research areas in this genre.

This book is a compilation of the lectures delivered in the SERC school. We thank all the contributors for coming and giving the lectures in 2015 in Kolkata, and for their extra effort to write these pedagogic lecture notes. We thank the referees for their meticulous work that improved the readability of the chapters. A summary of the book and guidelines on the usage of the book for courses are given separately in the following pages.

Somendra M Bhattacharjee, Institute of Physics, Bhubaneswar

Mahan Mj, TIFR, Mumbai

Abhijit Bandyopadhyay, Ramakrishna Mission Vivekananda University, Belur, Howrah

Acknowledgments

This book would not have seen the light of the day without the contributors writing up their lectures and making them accessible to a wider readership than the participants of the school. We thank A. Abanov, B. Ezhuthachan, J. K. Pachos, K. Sengupta, P. Ramadevi, R. Shankar, Subhro Bhattacharjee, S. Nechaev, S. Rao and S. Sen for their lecture notes. We also thank Ashoke Sen, P. S. Majumdar and James de Lisle for their lectures in the school. Special mention must be made of the members of the Department of Mathematics, Ramakrishna Mission Vivekananda University (RKMVU), A. Bhattacharya, D. Kulkarni, K. Biswas, Samik Basu, Somnath Basu, S. Maity, and U. Choudhury, for their diligent attempt to make the topics understandable by the physics students. Many tutorials were held to clarify points and answer questions, some of which can be found in the book as special sections.

The school could not have been run without the active help and advice of Dr Bobby Ezhuthachan, Sreetama Goswami, Aritra Saha, Dipak Patra, Sumon Sahu, Anirban Dinda and Suchetan Das. The school was enjoyable mainly because of the active participations of the participants during class- and off-hours. We also thank the students of physics and mathematics of RKMVU for participating in the physics-maths joint seminars for the year preceding the school.

Special thanks to Somnath Basu for many help in the editorial process and for drawing the two cartoons. Thanks are also due to Sreetama Goswami, Jayashree Behera, and Debraj Das for their generous help in editing the book.

We acknowledge the support of DST-SERB, Government of India, for funding this school which was held in S N Bose National Center for Basic Sciences under the aegis of RKMVU, Belur, Howrah. We thank the authorities and the supporting staff of RKMVU for their full support to make it a success. The authorities and the supporting staff of S N Bose Center are also thanked for their wholehearted cooperation. We acknowledge the keen interest taken by Shri D. K. Jain of Hindustan Book Agency and the TRiPS series Editor R. Ramaswamy.

Prolegomena

Layout

The book consists of two parts. The first part, consisting of rather short chapters (Chapters 2-7), is a quick, but more-or-less complete, review of topology. The focus has been on the fundamental concepts rather than on detailed proofs, all the while retaining the basic flavour of mathematics. There is an overview chapter (Chapter 1) at the beginning and a recapitulation chapter on group theory (Chapter 8). The physics section (Chapters 9-18) starts with introductory chapters and goes on to topics in quantum mechanics, statistical mechanics (polymers, knots and vertex models), solid state physics, exotic excitations (Dirac quasiparticles, Majorana modes, Abelian and nonabelian anyons), quantum spin liquids and quantum information processing.

Prerequisites

The readers are expected to be familiar with the usual notions of quantum mechanics (including path integrals, Dirac hamiltonian), statistical mechanics, and introductory solid state physics. In particular no field theoretic techniques are discussed, and so field theory is not a prerequisite. No extra mathematical background is assumed than what is expected of a graduate student of physics. It is to be noted that these are not review articles of current research but are more of a pedagogic nature; selected references are given to help the readers delve into current papers.

Summary

The overview chapter (Chap 1) can profitably be reread as one sails through the book.
Below, we elaborate more on the contents of the mathematics chapters to justify their inclusion rather than the physics chapters.

I Mathematics part — Topological tools

Various aspects of topology are covered in Chapters 2-7.

- A basic introduction and a tutorial comprise Chapter 2. This chapter with common examples shows why we need to define topology and topological spaces.

- The notion of fundamental group is introduced in Chapter 3. Here one explores or probes a topological space with the help of circles and identifies a topological invariant, the fundamental group, that remains invariant under a continuous transformation (like changing a solid torus into a cup). The general idea of homotopy groups is elaborated in part II; here a space is probed by means of higher dimensional spheres, thereby generalizing the notion of the fundamental group. Some computations of examples can be found in part III of this chapter.

- Often it may be sufficient to distinguish spaces by the "holes" in it and that can be achieved by Homology which is discussed in Chapter 4.

- Manifolds and differential forms are introduced in Chapter 5.
  Is the earth round or flat? The importance of the question (and the answer) is that one may draw a local map on a flat plane but not globally for the whole earth. The local versus global aspects of a space, is the topic of this chapter on manifolds by introducing tools from calculus. The approach unifies the three basic theorems for three dimensional vectors – Green's, Stokes' and Gauss' theorems - into a single one.

  At the next stage we need to generalize the notion of a Euclidean metric on flat Euclidean space by the notion of a metric on an arbitrary manifold. This calls for the basic tools of Riemannian geometry to be introduced. We do this in part II of this chapter.

- Vector bundles and K-theory are discussed in Chapter 6.
  Most of physics involves vectors defined in space, like the electric field $\mathbf{E}(\mathbf{x})$ at a point $\mathbf{x}$, or the magnetization field $\mathbf{M}(\mathbf{x})$. This needs to be generalized if $\mathbf{x}$ is a point on a manifold, not necessarily the Euclidean space, and the vector is from an arbitrary vector space. The generalization of vector functions on Euclidean spaces to manifolds gives rise to the notion of vector bundles. K-theory is used to distinguish between vector bundles. Bott periodicity, a fundamental theorem of K-theory, is also introduced in this chapter.

  How to find global invariants of general vector bundles using methods of Riemannian geometry is the topic of Chern-Weil theory. An introduction is given in part II of this chapter.

- A short introduction to a few mathematical aspects of knot theory is in Chapter 7. It is elaborated further in various physics chapters.

Since group theory is a part of the basic language in this discourse, there is a chapter recapitulating group theory (Chapter 8).

II Physics problems:

The physics part starts with two introductory chapters (Chapters 9, 10) on the use of topology in different branches of physics and the idea of dimension. Next is a chapter (Chapter 11) on topology and geometry as used in quantum mechanics, followed by a chapter on applications in solid state physics (Chapter 12). An introduction to Dirac quasiparticles and Majorana modes are given in Chapter 13.

The connection between knot invariants and vertex models can be found in Chapter 14 while many newer ideas in the field of polymers can be found in Chapter 15. Nonabelian anyons are introduced in Chapter 16. Quantum spin liquids are discussed in Chapter 17. A discussion on the topological aspects of quantum computation can be found in Chapter 18.

Suggestions for courses

The book is designed for a two semester long course on "Topology and condensed matter physics" to give a holistic approach to various branches of physics that go under the name of "condensed matter physics". The interlink is the ideas from topology. With some pick-and-choose the topics of this book can be covered comfortably in one semester.

Selected chapters of the book may also be used as a short segment of other advanced courses like advanced quantum mechanics, advanced statistical mechanics, or solid state physics, or even a mathematical methods course. Possible selections are given below.

Suggestions for short courses are as follows:

- For Quantum Mechanics: Chapters 1-4, 7-11, 16
- For Statistical Mechanics: Chapters 1-4, 7-10, 14, 15
- For Solid state Physics focusing on exotic excitations: Chapters 1-4, 7-13, 14-18
- For a Mathematical methods course: Chapters 1-8.

Notation

We list below the notations for a few common spaces and groups used in this book. Other symbols are defined as and when introduced.

Symbol	Name	in Cartesian coordinates
B^n, D^n	n-Ball, a solid sphere in n-dimensions	$\sum_{j=1}^{j=n} x_j^2 \leq 1$
$\mathbb{C}$	Complex plane	$z = x + iy$
D^2	Disk	$x_1^2 + x_2^2 \leq 1$
$\mathbb{R}, \mathbb{R}^1$	Real line	$-\infty < x_1 < +\infty$
$\mathbb{R}^2$	Real plane	$-\infty < x_1, x_2 < +\infty$
$\mathbb{R}^n$	Real n-dimensional space	$-\infty < x_1, ..., x_n < +\infty$
S^1	Circle	$x_1^2 + x_2^2 = 1$
S^2	Sphere	$x_1^2 + x_2^2 + x_3^2 = 1$
S^n	n-sphere	$\sum_{j=1}^{j=n+1} x_j^2 = 1$
$\mathbb{Z}$	Integers	$n = 0, \pm 1, \pm 2, ...$
$\mathbb{Z}_2$	Integers modulo 2	$n = 0, 1$ (0 for even, 1 for odd integers)
$\mathbb{Z}_q$	Integers modulo q	$n = 0, 1, ..., q-1$ with $n + q \equiv n$

For strict inequalities, one often talks of "open ball", "open disk" etc.

List of contributors

- *Alexander Abanov* (aabanov@gmail.com) is at the Department of Physics and Astronomy, and Simons Center for Geometry and Physics, Stony Brook University, Stony Brook, NY 11794, USA
- *Atreyee Bhattacharya* (atreyee00@gmail.com) is at the Department of Mathematics, IISER Bhopal, Bhopal - 462006, Madhya Pradesh, India
- *Bobby Ezhuthachan* (bobby.ezhuthachan@gmail.com) is at the Ramakrishna Mission Vivekananda University, Belur Math, Howrah 711202, India
- *Dheeraj Kulkarni* (dheeraj.kulkarni@gmail.com) is at the Department of Mathematics, IISER Bhopal, Bhopal, Madhya Pradesh, 462066, India.
- *Jiannis K. Pachos* is at the School of Physics and Astronomy, University of Leeds, Leeds, LS2 9JT, United Kingdom.
- *Kingshook Biswas* (biswaskingshook@gmail.com) is at the Indian Statistical Institute, 203 B. T. Road, Kolkata 700 108, India
- *K. Sengupta* is at the Department of Theoretical Physics, Indian Association for the Cultivation of Science, Jadavpur, Kolkata-700032, India.
- *Mahan Mj* (mahan@math.tifr.res.in, mahan.mj@gmail.com) is at the Tata Institute of Fundamental Research. 5, Homi Bhabha Road, Mumbai-400005, India.
- *P. Ramadevi* (ramadevi@phy.iitb.ac.in) is at the Department of Physics, Indian Institute of Technology Bombay, Mumbai 400 076, India.
- *Samik Basu* (samik.basu2@gmail.com) is at Department of Mathematics and Computational Science, Indian Association for the Cultivation of Science, Kolkata - 700032, India.
- *Sergei Nechaev* is at Université Paris-Sud/CNRS, LPTMS, 91405 Orsay, France, CNRS/Independent University of Moscow, Poncelet Lab., Moscow, Russia, and P.N. Lebedev Physical Institute RAS, 119991, Moscow, Russia
- *Siddhartha Sen* (sen1941@gmail.com) is at the Trinity College, Dublin.
- *Soma Maity* (soma123maity@gmail.com) is at the Department of Mathematics, IISER Mohali, Knowledge City, Sector 81, PIN: 140306, India.

- *Somendra M. Bhattacharjee* (somen@iopb.res.in) is at the Institute of Physics, Bhubaneswar 751 005 India.
- *R. Shankar* (shankar.chennai@gmail.com) is at the The Insitute of Mathematical Sciences, Chennai, India.
- *Somnath Basu* (basu.somnath@gmail.com) is at the Department of Mathematics & Statistics, IISER Kolkata, Mohanpur, West Bengal 741246, India.
- *Subhro Bhattacharjee* (subhro@icts.res.in) is at the International Centre for Theoretical Sciences, TIFR, Bangalore 560089, India.
- *Sumathi Rao* is at the Harish-Chandra Research Institute, Chhatnag Road, Jhusi, Allahabad 211 019, India.
- *Utsav Choudhury* (utsav.choudhury@iiserkol.ac.in) is at the Department of Mathematics & Statistics, IISER Kolkata, Mohanpur, West Bengal 741246, India.
- *Ville Lahtinen* is at Freie Universität Berlin, Arnimallee 14, 14195 Berlin, Germany.

Contents

I Topological Tools

II Physics problems

1

Overview of Topological Ideas in Condensed Matter Physics

Siddhartha Sen

Topological methods have been found to be useful for understanding features of condensed matter physics. The quantum Hall effect, the problem of classifying defects and the topological insulator are examples where topological ideas successfully explain unusual physical effects.

In this overview chapter, I will introduce some topological ideas that have been used in physics, stressing intuitive features of the subject. These ideas are discussed in greater detail in other chapters of this book.

1.1 Introduction

Topology is the abstract study of continuity. It introduces mathematical structures which remain unchanged under continuous deformation. For this reason it is often called "rubber sheet geometry" as it focuses on features that do not change when continuous changes are made to the parameters of the system. It is thus a very useful tool for discovering "invariant under continuous deformation" features that may be present in a physics problem. Very often these features appear because of boundary conditions imposed on wave functions.

The reason topological ideas emerge from boundary conditions is because boundary conditions can turn a flat space into a space with twists or holes. To see how this can happen let us examine what happens to a space when periodic boundary conditions are imposed. Periodicity in one dimension turns a line into a circle. Periodicity in two directions, say the x and y directions,

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_1

requires the opposite sides in the x and y directions of a rectangle defined by the periodicity features of the problem be identified. This identification turns an originally flat rectangle into a mathematical torus which is a space with holes. This can be easily visualized. If we join the opposite sides of the rectangle in the x direction we end up with a cylinder. If we then join the two circular ends of the cylinder in order to impose periodicity in the y direction we get a torus. Thus the original flat sheet has now become a torus which has holes because of boundary conditions. If we had joined the edges of a rectangle in a different way we could get the Möbius strip. To get a Möbius strip we simply join one set of opposite sides after a twist. In this case no periodicity in the y direction is present. Gauge field interactions introduced on a space with holes, such as a torus or a Möbius strip, introduce consistency conditions in order to be well defined. As established in the mathematical discipline of fibre bundles, these consistency conditions can lead to quantization of physical variables such as the conductance, this happens for the Hall effect, or to the existence of zero energy states, this happens for the topological insulator.

From a mathematical point of view a gauge field is a way to relate wave functions located at different points. In mathematics, a gauge field is called a connection on a fibre bundle. We will informally introduce the idea of a fibre bundle and explain why it is a basic structure which is present in the description of all physical systems. The consistency conditions of topology are robust and are not affected by changing the parameter values of the system. They lead to observable effects.

We now give a few examples where topological reasoning is used.

1. Is our planetary system stable? To answer this question we do not need to determine the position of all the planets but only to know that the orbit of all the planets remain in a bounded region. In fact Poincaré invented topological methods to tackle this problem.

2. Are there gapless states? Here the interest is not on the precise details of the energy spectrum of a system but whether the Hamiltonian describing the system has zero energy modes. The existence of zero modes can often be found using symmetry arguments. But even if no symmetry is present topological ideas, for a space with holes, can require a Hamiltonian operator to have a zero modes. Here the idea of a fibre bundles and K-theory can be useful.

3. Are localized excitations possible in a system? Again there are topological ways of finding out if localized, stable excitations (solitons) exist for a given system. Here Homotopy groups of topology play a role.

4. Can one understand the nature of allowable possible defects? Defects can be classified using Homotopy groups from topology.

5. If a crystal symmetry group changes from a higher symmetry to a subgroup symmetry are there selection rules for the subgroup symmetry allowed? Here the topological method of Morse theory can help.

6. Can large quantum coherent systems be created? This is a problem of great interest. An interesting approach is to construct topologically protected quantum entangled ground states i.e., systems whose coherence can only be destroyed by noncontinuous deformations.

This list of examples is not comprehensive but gives a flavour of the wide variety of possible uses of topological reasoning.

In this collection of lectures some of the basic concepts of topology are explained and and physical examples are discussed where these topological tools have been found to be useful.

Let us now proceed to give an informal account of topology viewed as "rubber sheet geometry". We start with Euclidean geometry. Euclidean geometry describes space in terms of a set of points and a distance function between them which is defined by the theorem of Pythagoras. This Euclidean distance function $d(x, y)$ between two points x and $y \in \mathbb{R}^n$ does not change under rotations or translations of the coordinate system chosen. Rotations and translations are thus symmetries of Euclidean geometry. There are other geometries possible which have distance functions with different symmetry. With the help of the Euclidean distance function, or any other distance function, the concept of continuity can be defined which leads to the operation of taking a limit which leads to calculus. Thus the idea of continuity is fundamental for properly stating the laws of physics.

Continuity of a real valued function $f(x)$ at a point x_0 is the statement that for any $\epsilon > 0$, we can find a $\delta > 0$ such that if $d(x, x_0) < \delta$ (an open interval) then $d(f(x), f(x_0)) < \epsilon$ (a different open interval). In words, continuity means that close by points are mapped to close by points, so that for a choice of the distance between x and x_0 less than a δ, the two points $f(x)$ and $f(x_0)$ mapped by the function will be separated by a distance less than ϵ, where δ depends on the choice made for ϵ.

For topology, regarded as geometry, we have to find some way of capturing its characteristic feature, which is "invariance under continuous deformations". The notion of distance is no longer relevant as any distance function introduced between points will change under deformation. We need to find a set of deformation invariant objects to replace the set of points of Euclidean geometry as the building block of topology. Such a set of objects can be found. Let us explain how.

In the definition of continuity on the real line, two key ideas were used. The first was the idea of distance. The second was the idea of an open interval. The open interval was defined to be the set of all points x that satisfy the condition. defined by $d(x, x_0) < \delta$. We will use this idea to define the generalized geometry required in topology.

Euclidean geometry, we pointed out, is defined in terms of a set of points and a distance function between them. The set of points is the undefined fundamental object of the subject. In topology the analogue of the set of points is a collection of open sets and the analogue of the distance function is a set of rules which tells us properties that the collection of open sets must have.

Intuitively it seems reasonable that an open set, for example an open interval on the real line, remains an open set under continuous deformations. The precise conditions required for open sets to give a "topological space" will be described shortly. The important point to grasp is that the open set in topology is not defined by using a distance function but is the fundamental object of the subject and the aim of the subject is to study objects that are invariant under continuous deformations. These invariants, discovered in topology, are found to be important for physics.

Our next step is to give an informal account of a few useful mathematical ideas from topology that are relevant in the context of physics. These ideas are:

1. The idea of a manifold.
2. The idea of differential forms.
3. The idea of homotopy, homology and cohomology.
4. The idea of Fibre Bundles and Vector Bundles.

We introduce the mathematical ideas and, where possible, also give practical tools for using them. Let us explain what these ideas are and why they are useful.

We start with the arena in which physical events happen: space. Let us recapitulate a bit. In our introduction we introduced Euclid's geometry as a mathematical structure where space was represented by points and a distance function given by the theorem of Pythagoras. The distance function could be used to tell us how close two points were to one another. Using this idea of closeness, the idea of limits could be formulated which then leads to calculus: the key mathematical idea for modeling physical systems. Thus the notion of distance and closeness are fundamental ideas on which physics is built. Let us expand on what we said before.

The fundamental notion of distance due to Pythagoras defines Euclidean space. There are however many situations where Euclidean distance is not appropriate. For instance the distance between two points on the surface of a sphere. However two points on the surface of a sphere that are close together do have a distance which can be approximated by the Euclidean distance. Spaces that have the property that the distance between points close together can be approximated by the Euclidean distance are called manifolds. For such spaces a scheme of representing them by patches of Euclidean space is possible. By representing a patch of a manifold by Euclidean space we mean that an invertible map taking points on a patch of the manifold to Euclidean space exists. If the map is smooth then this procedure can be used to introduce calculus on a manifold. The idea is to carry out usual calculus operations (differentiation and integration) in the Euclidean space associated with a patch of the manifold and then use the map to transfer the result to the manifold. This is a conceptual step. In many situations the precise nature of the map is not important. It is clear that many different maps are possible. Each map provides Euclidean coordinates for the manifold. If two patches overlap then points of the manifold

in the overlap region belonging to two different patches will have two different coordinate representations and formula for relating the two sets of coordinates need to be given.

Thus calculations on the manifold are coordinate dependent. For this reason mathematicians invented objects that contain geometrical/physical information about a manifold but are coordinate independent. Differential forms are examples of this type.

Finally a manifold can have features that remain unchanged even when the manifold is deformed preserving continuity. For example, a torus cannot be changed into a sphere without tearing. The features of a manifold that are preserved under continuous deformations are called topological invariants. Homotopy, homology and cohomology groups are examples of such topological invariants.

Thus the reason the topics listed are worth studying is that they make it possible to carry out calculus calculations on manifolds, to present calculations in a coordinate independent way and highlight global features of a manifold that have important physical consequences.

Finally, given a manifold, in order to study physical phenomena on it, the introduction of another space is always required. Suppose, for example, we want to study wind flows on the earth. We then represent the surface of the earth described using patches but we also need to introduce a second space to describe the wind speed and direction. This space is three dimensional Euclidean space. These two spaces together can model the physics problem. The two spaces, the two dimensional patch, and the three dimensional Euclidean space, have to be combined together in a smooth way.

The construction described has been studied in great detail by mathematicians under the name *fibre bundle*.

The way we will explain the ideas listed is by looking at examples and by explaining key concepts. We start with manifolds.

1.2 Manifolds

A manifold M is described by overlapping patches, called charts, that completely cover it and each chart has an invertible map from a portion of the manifold to an Euclidean space. For instance, the Euclidean space $\mathbb{R}^n$ is a manifold and a single chart (the identity map from $\mathbb{R}^n$ to itself) covers it. However, it can also be covered by a countably infinite number of charts as follows. Local charts can be defined in terms of spherical balls (of rational radius) centred at rational points, i.e., points with each coordinate a rational number, in Euclidean space. This collection of local charts is countably infinite. The key point is that the rationals are dense in the reals (meaning that any real number

is the limit of a sequence of rational numbers) which, in turn, implies that the rational points are dense in Euclidean space. Thus the collection covers $\mathbb{R}^n$.[1]

To clarify ideas let us consider a simple manifold, namely a circle S^1. There are many different ways of describing S^1. Algebraically we can think of it as points $(x, y) \in \mathbb{R}^2$ such that $x^2+y^2 = 1$, or as points on the real line where points x and $x + L$ are identified with a fixed L, or as the space defined by the set of equations: $x^2 + y^2 + z^2 = 1$ and $z = 0$. Each one of these descriptions has an extension that can be useful for studying more general spaces. The method of defining S^1 where points x and $x + L$ are identified (i.e., glued together) is called a quotient construction.

We can introduce more structures to the manifold. For example, we can represent S^1 as a differentiable manifold. This means that the gluing maps introduced to describe S^1 are chosen to be smooth i.e., differentiable and not just continuous. Given two overlapping charts for a patch on a manifold M, we may go from the Euclidean space to M via (the inverse of) one chart and then back to the Euclidean space via the other chart; such maps are called transition maps. We say M is smooth if all transition maps, as maps from the Euclidean space to itself, are smooth or differentiable up to all orders. Once we have a smooth manifold, we may do usual calculus (differentiation and integration) in the Euclidean space associated with a patch of the manifold and then use the map to transfer the result to the manifold. However, two charts will give rise to two different coordinate representations. The transition maps being smooth is the fundamental point which shows that a formula written in terms one set of coordinates is smooth implies the same formula in the other set of coordinates is also smooth. Thus, the notion of smoothness is well defined on the whole manifold and does not depend on the local chart chosen to represent a given smooth quantity defined on M.

The first step then that is needed to represent S^1 as a manifold is to introduce charts and smooth invertible maps to cover it and provide a coordinate system for describing S^1. If functions are introduced in each chart which are differentiable then by using the smooth invertible maps the notion of differentiable functions can be transferred to S^1, i.e., we now have calculus on S^1. The technical details of this scheme will not be given. Our aim is to explain the ideas involved. In applications the precise form of coordinates is rarely required.

To describe S^1 a minimum of two coordinate charts are required. There are many ways of introducing charts. Let us describe one way. We take S^1 to be the set of points on $\mathbb{R}^2$ defined by $S^1 = \{(x, y)|x^2+y^2 = 1\}$. A chart maps the open set U_1, defined to be the circle minus the point $(1, 0)$, i.e., $U_1 = S^1 - \{(1, 0)\}$, to the Euclidean line (the y axis via a stereographic projection). We draw a straight line from point $(1, 0)$ to the given point in U_1 and see where it intersects the line $x = 0$. This gives a point on the y axis, namely, the point E_p. Explicitly

[1] There is also the notion of compact manifolds. A compact manifold is a manifold which is also a compact topological space. For example, Euclidean space is a manifold but it is not compact, while S^1 is an example of a compact manifold. More details may be found in later chapters.

we have

$$\varphi : U_1 \to \mathbb{R},\ (x, y) \mapsto \frac{y}{1-x}.$$

The resulting coordinates cover the entire circle except the point $(1, 0)$. Similarly, one can think of another chart $U_2 = S^1 - \{(-1, 0)\}$ which covers the circle except the point $(-1, 0)$, given formally as

$$\psi : U_2 \to \mathbb{R},\ (x, y) \mapsto \frac{y}{x+1}.$$

These two charts thus cover all points of the circle. It follows that the overlap is $S^1 - \{(\pm 1, 0)\}$, and the transition map

$$\psi \circ \varphi^{-1} : \mathbb{R} - \{0\} \to \mathbb{R} - \{0\},\ t \mapsto \frac{1}{t},$$

is smooth. It is now possible to do calculus on S^1 by doing it on each chart. For instance, a function on S^1 will be called smooth if it is continuous and smooth on each chart. With this simple example of S^1, we see how charts, once introduced, give a local description of a manifold in terms of Euclidean space and calculus can be introduced on the manifold.

1.3 Differential Forms

Differential forms and their dual, the vector fields, are natural global objects that appear when the operations of calculus are introduced on a manifold.[2] Functions are the simplest examples of differential forms. They are differential zero forms. In general differential forms are antisymmetric tensor fields that appear in physics and geometry. For example, the electromagnetic field $F_{\mu\nu}$ and the curvature tensor $R_{\mu\nu}$ are differential two forms, while vector fields appear as generators of change and symmetry. Both objects can be defined in a coordinate independent way on a manifold. However they can also be introduced using a local Euclidean coordinates description. The local description is useful for calculations while the coordinate free description is useful for spotting general features of a problem. Thus a local description for a p-form, $\omega^{(p)}$ is:

$$\omega^{(p)} = \sum \omega_{i_1, i_2, \cdots, i_p}(x) dx^{i_1} \wedge dx^{i_2} \wedge \cdots \wedge dx^{i_p},$$

where $\omega^{(p)}$ is an object on the manifold, while the expression written down gives a representation of this object using the coordinate building block differentials $dx^{i_1}, dx^{i_2}, \ldots$. These differentials satisfy the basic antisymmetric rule of $\wedge$ multiplication, $dx \wedge dy = -dy \wedge dx$. Due to the antisymmetry, terms which look like $dx^{i_1} \wedge \cdots \wedge dx^{i_p}$ with repeated indices are zero. In such a case we

[2] More technical details on forms can be found in Chapter 5

may take $\omega_{i_1,\cdots,i_p}$ to be zero, i.e., $\omega_{i_1,i_2,\cdots,i_p} = 0$, if there are repeated indices. Moreover, if indices are nonrepeating then

$$dx^{i_1} \wedge \cdots \wedge dx^{i_p} = \text{sign}(\sigma) dx^{\sigma(i_1)} \wedge \cdots \wedge dx^{\sigma(i_p)}$$

where σ is a permutation of the set $\{i_1, \cdots, i_p\}$ and $\text{sign}(\sigma)$ is the sign of the permutation σ. This negative sign can be compensated by choosing $\omega_{i_1,i_2,\cdots,i_p}(x)$ to be odd under odd permutation. Note that in the summation each i_k is from $1, \cdots, n$. This leads to many identical terms in the expression above, since, by our choice, any permutation of the labels $i_1, \cdots, i_p$ gives the same contribution. Thus, there are $p!$ equal contributions to $\omega^{(p)}$. Very often this feature is recognized by introducing a $\frac{1}{p!}$ term. Thus we can write

$$\begin{aligned} \omega^{(p)} &= \sum_{i_1<\cdots<i_p} \omega_{i_1,\cdots,i_p}(x) dx^{i_1} \wedge \cdots \wedge dx^{i_p}, \\ &= \frac{1}{p!} \sum_{i_1,\cdots,i_p=1}^{n} \omega_{i_1,\cdots,i_p}(x) dx^{i_1} \wedge \cdots \wedge dx^{i_p} \end{aligned}$$

For example, if $\omega = f_{12} dx^1 \wedge dx^2 + f_{21} dx^2 \wedge dx^1$ then

$$\omega = \omega_{1,2}\, dx^1 \wedge dx^2 = \frac{1}{2!}(\omega_{1,2} dx^1 \wedge dx^2 + \omega_{2,1} dx^2 \wedge dx^1),$$

where $\omega_{1,2} = -\omega_{2,1} = f_{12} - f_{21}$.

Now, p-forms can be added and multiplied by scalars to generate new p-forms. They thus form a vector space. For example, on the Euclidean space of dimension n, any 1-form is given by

$$\omega = f_1 dx^1 + \cdots + f_n dx^n,$$

where f_i's are smooth functions. We already mentioned that a 0-form is a smooth function. Therefore any 1-form is a linear combination of n generating 1-forms with coefficients coming from the space (actually a ring) of smooth functions. If the dimension of the manifold is n then differential one forms are elements of a n dimensional vector space with basis vectors $dx^i, i = 1, ..., n$ while differential p-forms belong to a vector space of dimension ${}^nC_p$ with basis vectors $dx^{i_1} \wedge dx^{i_2} \wedge \cdots \wedge dx^{i_p}$.[3]

On a local chart, the generating 1-forms dx^i can be multiplied. This leads to the notion of multiplying a p-form and a q-form together to generate a $(p+q)$-form. This multiplication is associative as it is so on charts, but noncommutative

[3]Formally, these structures are called modules . We say that the space of 1-forms (on $\mathbb{R}^n$) is a finitely generated module over the ring of smooth functions. In fact, it is of rank n with a basis given by $dx^1, \cdots, dx^n$. More generally, the space of p-forms on $\mathbb{R}^n$ is a module over the smooth functions of rank $\binom{n}{p}$ with a basis given by $dx^{i_1} \wedge \cdots \wedge dx^{i_p}, i_1 < \cdots < i_p$. If the manifold M is compact then a theorem due to Swan implies that the space of p-forms is a finitely generated module over the ring of functions.

(by the antisymmetry rule). Explicitly,

$$\begin{aligned} \omega^{(p)} \wedge \mu^{(q)} &= (-1)^{pq} \mu^{(q)} \wedge \omega^{(p)} \ : \ \text{gradedcommutative} \\ \mu_1^{(q_1)} \wedge (\mu_2^{(q_2)} \wedge \mu_3^{(q_3)}) &= (\mu_1^{(q_1)} \wedge \mu_2^{(q_2)}) \wedge \mu_3^{(q_3)} \ : \ \text{associative.} \end{aligned}$$

Differential forms are also elements of an exterior algebra.[4] The coefficients $\omega_{i_1, i_2, \cdots, i_p}(x^i(p))$ in the local description of a p-form can be given a coordinate independent interpretation. We do not go into this in detail.

1.3.1 Examples

In dimension two we have the following forms:

- Zero form: $f_0(x)$, a function.
- One form: $f_1(x)dx^1 + f_2(x)dx^2$, and
- Two form: $f_3(x)dx^1 \wedge dx^2$.

In three dimensions we can have

- Zero form: f_0,
- One form: $f_1 dx^1 + f_2 dx^2 + f_3 dx^3$,
- Two form: $g_1 dx^1 \wedge dx^2 + g_2 dx^3 \wedge dx^1 + g_3 dx^2 \wedge dx^3$,
- Three form: $h\ dx^1 \wedge dx^2 \wedge dx^3$.

1.4 Vector fields

Vector fields and differential forms are dual objects. Differential forms describe antisymmetric tensor fields in a coordinate independent way while vector fields are operators that change coordinates and hence change functions and forms.

We explain vector fields by looking at an example. Suppose there is a moving point p on a manifold M of dimension n. We want to describe this motion. This can be done by using a local coordinate description of $p(t)$ as $x^i(p(t)), i = 1, 2, ..., n$. As the point $p(t)$ moves, it generates a curve on the manifold M. Its velocity at time t is the tangent to this curve at the coordinate point $x^i(p(t))$, which is

$$\frac{dx^i(p(t))}{dt}, \ i = 1, 2, ..., n.$$

Using this idea, we consider the way a function $f(p)$ of $p(t)$ changes. We have

$$\frac{df}{dt} = \sum_i \frac{dx^i(p(t))}{dt} \frac{\partial f}{\partial x^i}.$$

[4] The particular noncommutative product rule is also referred to as graded commutativity. The space of all differential forms is a (graded) vector space with a multiplication that is associative and graded commutative.

The next step is to write this result as

$$\begin{aligned} \frac{df}{dt} &= \mathbf{X}f \\ \mathbf{X} &= \sum_i v^i \frac{\partial}{\partial x^i} \\ v^i &= \frac{dx^i}{dt}. \end{aligned}$$

Thus we have introduced an operator $\mathbf{X}$ which changes a function. This operator is the vector field. Recall that two vector spaces are dual to one another if an element of one acting on an element of the other gives a scalar quantity. This is equivalent to the definition of the dual space of a vector space V, as the set of all linear maps from V to $\mathbb{R}$. It was pointed out earlier that since differential forms of the same order can be added and multiplied by a scalar to give a differential form they are elements of a vector space. Similarly vector fields of the same dimension form a vector space. These two vector spaces, when they are of the same dimension are a dual to one another which can be displayed as follows.

When we consider differential 1-forms at a point p in M, we get a vector space of dimension n generated by the basis $dx_i(p)$. Working on a local chart, a differential 1-form defines a map from the vector space of (local) vector fields to $\mathbb{R}$ as follows. We write

$$df = \sum_i \frac{\partial f}{\partial x^i} dx^i,$$

and define

$$\langle df, \mathbf{X} \rangle = \mathbf{X}f.$$

In particular for $f = x^j$ and $\mathbf{X} = \frac{\partial}{\partial x^i}$, we have

$$\langle df, \mathbf{X} \rangle = \langle dx^j, \frac{\partial}{\partial x^i} \rangle = \frac{\partial x^j}{\partial x^i} = \delta_i^j,$$

where $\delta_i^j = 1$, when $i = j$ and is zero otherwise, so that, in general,

$$\langle \omega^{(1)}, X \rangle = \left\langle \sum_i f_i dx^i, \sum_j v^j \frac{\partial}{\partial x^j} \right\rangle = \sum_i f_i v^i.$$

This result makes the dual link between forms and vector fields clear by showing how they can be paired to give a number.[5] Very often we will refer to the vector fields $\frac{\partial}{\partial x^i}$ as elements of the tangent space of a manifold at point p and refer to dx^i as elements of the dual cotangent space at p, with $i = 1, 2, ..., n$, where the dimension of the manifold is n. We next consider the operations of calculus on forms. We start with the exterior derivative.

[5] We may also view this as a map from the dual of vector fields at p to the vector space of 1-forms at p. In fact, this map is a linear isomorphism.

1.5 The exterior derivative (d)

The exterior derivative maps a p-form to a $(p+1)$-form. Locally a p-form is represented as a finite sum of the wedge product of p-terms. The action of the d-operator increases the number of such terms by one in the following way

$$d\omega^{(p)} = \sum_{i_1,\cdots,i_p,j} \frac{\partial \omega^{(p)}_{i_1,\cdots,i_p}(x)}{\partial x^j} dx^j \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_p}.$$

Let us discuss a few examples of d operator here. In two dimensions we can get

$$\begin{aligned} d(f_1 dx^1 + f_2 dx^2) &= \frac{\partial f_1}{\partial x^2} dx^2 \wedge dx^1 + \frac{\partial f_2}{\partial x^1} dx^1 \wedge dx^2, \\ &= -\left(\frac{\partial f_1}{\partial x^2} - \frac{\partial f_2}{\partial x^1}\right) dx^1 \wedge dx^2, \\ &= \left(\frac{\partial f_2}{\partial x^1} - \frac{\partial f_1}{\partial x^2}\right) dx^1 \wedge dx^2, \end{aligned}$$

which is basically the "curl" (area).

Next consider the two form in dimension 3:

$$\begin{aligned} & d(g_1 dx^1 \wedge dx^2 + g_2 dx^3 \wedge dx^1 + g_3 dx^2 \wedge dx^3) \\ = \; & \frac{\partial g_1}{\partial x^3} dx^3 \wedge dx^1 \wedge dx^2 + \frac{\partial g_2}{\partial x^2} dx^2 \wedge dx^3 \wedge dx^1 + \frac{\partial g_3}{\partial x^1} dx^1 \wedge dx^2 \wedge dx^3, \\ = \; & \left(\frac{\partial g_3}{\partial x^1} + \frac{\partial g_2}{\partial x^2} + \frac{\partial g_1}{\partial x^3}\right) dx^1 \wedge dx^2 \wedge dx^3. \end{aligned}$$

This is the familiar "divergence".

An immediate important consequence of the above representation of d is the result

$$d^2 = 0.$$

To see this, we simply use the definition of the d-operator. This simple identity is one of the fundamental identities and is quite ubiquitous in mathematics. For instance, it can be interpreted as the Jacobi identity in Lie algebras. It is a generalization of the two well known vector calculus results that (i) the divergence of the curl of any vector is zero, $\nabla \cdot (\nabla \times \mathbf{A}) = 0$, for any vector $\mathbf{A}$ and (ii) the curl of the gradient of a scalar function is zero, $\nabla \times \nabla\phi = 0$, for any scalar function ϕ. We may recall that these two important vector relations allow us to define the vector and the scalar potentials in electromagnetism. We now discuss some of the hidden features of this simple result.

The result suggests the question: if $d\omega^{(p)} = 0$, does it mean that we must have $\omega^{(p)} = d\eta^{(p-1)}$ for some $(p-1)$-form η? Forms for which $d\omega^{(p)} = 0$ are called closed. Forms which can be written as $\omega^{(p)} = d\eta^{(p-1)}$ are called exact. Note that exact forms are closed due to $d^2 = 0$. We will see that

locally all closed forms are exact, but this need not be the case globally. The local result is known as the Poincaré lemma while the possibility of the result not holding due to global considerations is the subject matter of de Rham cohomology, where closed forms modulo exact forms are shown to lead to groups that are topological invariants of the manifold. They are cohomology groups of the manifold.

1.6 A brief discussion on de Rham cohomology

We now show that Poincaré's lemma need not hold globally. To do this we introduce two spaces, namely Z^p and B^p which consist of all closed and all exact p-forms respectively on M defined as stated as follows

$$\begin{aligned} Z^p &: \{\omega^{(p)} \mid d\omega^{(p)} = 0\} \Rightarrow \text{ this is called space of all closed } p-\text{forms}, \\ B^p &: \{\omega^{(p)} \mid \omega^{(p)} = d\eta^{(p-1)}\} \Rightarrow \text{ this is called space of all exact } p-\text{forms}. \end{aligned}$$

Since an exact form is always closed, we can write

$$B^p \subset Z^p.$$

One can define the quotient space as

$$\begin{aligned} H^p(M,\mathbb{R}) &= Z^p(M,\mathbb{R})/B^p(M,\mathbb{R}) \\ &= p\text{th de Rham cohomology group.} \end{aligned}$$

The group property can be seen easily if we suppose that the number of closed p-forms on a manifold is finite. One can then choose a basis of these forms. The coefficients of these basis p-forms can be chosen as arbitrary real numbers, i.e., they are elements of $\mathbb{R}$. Elements of $\mathbb{R}$ form an Abelian group under addition, with zero as its identity element. Thus one can associate with Z^p and B^p, Abelian groups which reflect the number of elements that are closed or exact. The quotient of these groups is the de Rham cohomology group.[6] If Poincaré's lemma was globally valid, the spaces Z^p and B^p would be equal and the cohomology group would be trivial, i.e., it consists of a single element corresponding to the identity element of $\mathbb{R}$, which is zero.

We now give an example where this is not the case. The simplest counterexample is the one dimensional manifold S^1. For this 1–manifold, only forms in degree 0 and 1 exist. An arbitrary zero form will be described by a function parameterized by an angle θ which is periodic, while a one form can be written as $\omega^{(1)} = f_1(\theta)d\theta$, again with $f_1(\theta)$ being periodic in θ. We now examine the space Z^1 and B^1.

$$\begin{aligned} Z^1 &= \{\omega^{(1)} \mid d\omega^{(1)} = 0\}, \\ B^1 &= \{\omega^{(1)} \mid \omega^{(1)} = d\omega^{(0)}\}. \end{aligned}$$

[6]The finiteness assumption here is for simplicity— it is not necessary. In general, the set of closed p-forms is a group as it is a vector space. The same holds for exact p-forms. Both these are Abelian groups and the quotient of these groups Z^p/B^p is also a group.

Let $\omega^{(1)} = f_1(\theta)d\theta$, where $0 \leq \theta \leq 2\pi$. We write $f_1(\theta)$ as

$$f_1(\theta) = \sum_{n=-\infty}^{\infty} c_1^n e^{in\theta}.$$

Then $d\omega^{(1)} = 0$ trivially. We note from the above equations that

$$\begin{aligned} \omega^{(0)} &= \sum_{n=-\infty}^{\infty} c_0^n e^{in\theta}, \\ \text{or, } d\omega^{(0)} &= \sum_{n=-\infty}^{\infty} (in) c_0^n e^{in\theta} d\theta. \end{aligned}$$

Since we want to find out all the exact forms, we need $\omega^{(1)} = d\omega^{(0)}$. Hence comparing coefficients, we get

$$c_1^n = c_0^n(in), \text{ for } n \neq 0,$$

which means that c_0^n, which is a real number can not be eliminated by a term coming from $d\omega^{(0)}$. Thus we can write

$$\begin{aligned} H^1(S^1, \mathbb{R}) &= Z^1(S^1, \mathbb{R})/B^1(S^1, \mathbb{R}) \\ &= \mathbb{R}c_0^n, \end{aligned}$$

where $\mathbb{R}c_0^n$ is the real line generated by c_0^n, which is also the vector space generated by c_0^n. Thus we write

$$H^1(S^1, \mathbb{R}) = \mathbb{R}.$$

Here $Z^1 \neq B^1$, i.e., Poincaré lemma does not extend globally in this case.

1.7 Betti numbers

Let us start with one important comment about the example in the previous section. The cohomology group in this example was the group, under addition of real numbers. It was generated by just one independent real number, c_n^0, and for this reason is said to have rank one. The rank, an integer number, is called the Betti number associated with a cohomology group. In this case the Betti number is one for the cohomology group $H^1(S^1)$. We write the Betti number as $b_1(S^1) = 1$. It counts, in this case, the number of one dimensional holes present in the manifold S^1.

For a two torus T^2, the surface of a doughnut, the manifold is $T^2 = S^1 \times S^1$. The de Rham cohomology groups of T^2 can be calculated by constructing the number of independent zero, one and two dimensional forms that are closed but not exact following the procedure described. The cohomology groups are

$$\begin{aligned} H^0(T^2) &= \mathbb{R}, \\ H^1(T^2) &= \mathbb{R} + \mathbb{R}, \\ H^2(T^2) &= \mathbb{R}. \end{aligned}$$

The associated Betti numbers of T^2 are: $b_0 = 1, b_1 = 2, b_2 = 1$. There are tables listing cohomology groups and Betti numbers for standard manifolds, such as spheres and tori of arbitrary dimension. We list the Betti numbers for S^n, the n-sphere: $b_n = 1, b_0 = 1$, and all other Betti numbers $b_i = 0$.

1.8 Homotopy and Cohomology Groups

The de Rham cohomology groups were introduced as an answer to the question: Is a closed form (i.e., a form $\omega^{(p)}$ with the property $d\omega^{(p)} = 0$) always exact (i.e., $\omega^{(p)} = d\eta^{(p-1)}$)? The fact that the answer is no was used to distinguish the circle from the disc. Thus a method for detecting the presence of the hole in the circle (thought of as the unit circle in the plane) was found using differential forms. There are other ways of spotting holes in manifolds. Homology groups are another set of topological invariants which detect holes in a given space. The fundamental group, which was invented by Poincaré, detects holes in a space with a chosen base point. This group is the first in a series of groups one can associate to a based space. These groups are called homotopy groups. Although the idea of homotopy groups is very intuitive, these are rather subtle invariants and generally quite intractable. The objects used to construct these groups are loops and higher dimensional closed surfaces like spheres, which can encircle a hole and thus detect their presence.

Let us sketch the approach for a two dimensional surface, M, with a hole in it, while a formal approach is given later. We can pick a point on the surface, a base point $x_0 \in M$, and draw closed loops on this surface that starts at the base point; such loops are called based loops. These loops will be of two types. There will be loops that circle the hole and those that do not. Again for loops that encircle the hole there will be ones which circle the hole an arbitrary (but finite) number of times in either the clockwise or anticlockwise direction. The idea of homotopy is to introduce a notion of equivalence between two loops that can be smoothly deformed into each other. This generates an equivalence relation between based loops. Two loops that can be deformed into each other will be said to be homotopic. Thus given a loop α which starts and finishes at base point x_0 of the surface, there will be an equivalent class $[\alpha]$ of loops homotopic to α. Now a loop is nothing but a smooth map of the circle S^1 to the surface. Poincaré defined a group operation on the equivalence class of loops by introducing a way of combining two loops to form a product loop . The combination rule for loops was the group operation. Following such a procedure Poincaré constructed the first homotopy group, also called the Fundamental group of Poincaré, $\pi_1(M, x_0)$.

The rule for joining loops was simple. Each loop was taken to be a map $\alpha_i(t)$ on the surface with the parameter t in the closed interval $[0 \leq t \leq 1]$ chosen so that $\alpha_i(0) = \alpha_i(1) = x_0$, a fixed point on the surface. The important point was that a loop was continuous, and was to be parameterized by a variable t which was an element of a closed unit interval in such a way that the beginning and the end of the loop were the fixed point x_0. The way Poincaré introduced

a combination rule for loops was to consider two loops that started and ended at the same point x_0, and made the first loop return to its starting point in half the time interval followed by an equally fast moving second loop so that at the end of one unit the two loops are traversed. Thus the combined loops form one loop. Poincaré went on to show that by using the idea of homotopic loops the combination rule for homotopic loops form a group.

There is also a notion of two spaces X, Y being homotopic if X can be deformed smoothly to Y by means of a smooth parameter. A startling example of this is that n dimensional Euclidean space and the origin of the space, which is a point, are spaces of the same homotopy type. A continuous deformation which establishes this equivalence is the map $x \to tx$, where x is a point of the n dimensional space and t as a real number. For $t = 1$ we get the point x while for $t = 0$ we get the origin. This example makes it clear that homotopy related spaces need not have the same dimension. Let us write down mathematical expressions summarizing what we have said.

A loop α in a space M is thus a continuous map from the closed unit interval $[0, 1]$ to M with end points fixed at a point $x_0 \in M$. We write

$$\alpha : [0, 1] \to M, \quad \alpha(0) = \alpha(1) = x_0.$$

Thus a loop has three properties. It is continuous, it starts and ends at a fixed point of M and the parameter t used to describe its position in M varies in the closed interval $[0, 1]$.

Now if α, β are two loops in M, both based at x_0, we define the combination rule of Poincaré to form a new closed loop $\gamma = \alpha \cdot \beta$, where for all the loops the beginning and end points are fixed at $x_0 \in M$, i.e.,

$$\alpha(0) = \alpha(1) = \beta(0) = \beta(1) = \gamma(0) = \gamma(1) = x_0.$$

The rule for "multiplying" or combining closed loops is,

$$\gamma(t) \equiv (\alpha \cdot \beta)(t) = \begin{cases} \alpha(2t) & \text{for } t \in [0, \frac{1}{2}] \\ \beta(2t-1) & \text{for } t \in [\frac{1}{2}, 1] \end{cases}$$

The parameter factors $2t$ for α and $2t-1$ for β are introduced so that both α, β have parameters ranging over the interval $[0, 1]$ even though t ranges over $[0, \frac{1}{2}]$, for loop α and over $[\frac{1}{2}, 1]$ for loop β. The two closed loops meet at x_0 when $t = \frac{1}{2}$. Finally we write down what is meant by two loops being homotopic. Two loops α, β in M are homotopic if they can be continuously deformed into each other, i.e., there is a "homotopy" parameter s belonging to a closed interval $[0, 1]$ such that for $s = 0$ we get the loop α and for $s = 1$ we get the loop β. In other words, α_s, always based at x_0 for all s, interpolates between α, β as s varies. Here is how this can be written. We have a homotopy map $\alpha_s(t)$ which is a continuous map from $[0, 1] \times [0, 1]$ to M defined to have the property,

$$\alpha_s : [0, 1] \to M, \text{ such that } \alpha_0(t) = \alpha(t), \alpha_1(t) = \beta(t),$$

with $\alpha_s(0) = \alpha_s(1) = x_0$. If two loops α, β can be deformed into each other in this way they are said to be homotopic and we write $\alpha \approx \beta$. The equivalence class of loops homotopic to α is written as $[\alpha]$. The multiplication rule between two equivalent class of loops is defined to be $[\alpha] * [\beta] = [\alpha \cdot \beta]$, i.e., representative elements from each class are multiplied and then their equivalence class is taken. It can be proved that such a procedure is well defined, i.e., it does not depend on the particular representative elements chosen to define multiplication. Using this idea of multiplying equivalent class of loops Poincaré showed that they form a group. This was called the *fundamental group* of the space M in which the loops were introduced. The group captured a topological property of M. It was the birth of algebraic topology.

Homotopy groups are important in physics because by using them it is possible to distinguish spaces which cannot be deformed into each other. There are, for instance, an increasing number of examples in condensed matter where there are defects or textures which cannot be removed by continuous deformations. Such defects can be characterized by homotopy groups. Thus line defects are classified by π_1 while more intricate defects can be classified by higher homotopy groups. We will later show that the presence of Dirac points on a topological insulator can also be understood using homotopy groups.

For a two dimensional surface with a hole it is intuitively reasonable that the surface can be deformed to a circle S^1 and our discussion suggests that $\pi_1(S^1, x_0) = \mathbb{Z}$, where the integer $\mathbb{Z}$ describes the winding number, clockwise and anticlockwise of the loop around S^1. If we have $n \in \pi_1(S^1)$, it means we have a loop of winding number n.

The higher dimensional homotopy groups $\pi_n(M, x_0)$ come from continuous (loosely called smooth) maps of S^n to M. The cohomology groups $H^n(S^m)$ for S^m are known for all values of n, m but this is not true for the homotopy groups $\pi_n(S^m, x_0)$. They are not known for arbitrary integers n, m. But many cases are known and two important cases that are known are $\pi_n(S^n, x_0) = \mathbb{Z}$ and $\pi_3(S^2, x_0) = \mathbb{Z}$. The second example is very interesting as it shows the power of the homotopy group to spot global features, namely that the two spaces S^3 and $S^1 \times S^2$ are related by a subtle procedure which we now briefly describe. A simple way to do this is by representing S^3 by two complex variables (z_1, z_2) where $z_1 = x_1 + ix_2, z_2 = x_3 + ix_4$ which satisfy the condition corresponding to

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 = |z_1|^2 + |z_2|^2 = 1,$$

where $|z|$ is the modulus of the complex variable z, since this condition defines points on S^3 . It is clear that both z_1 and z_2 cannot be zero at the same time if they represent points on S^3. There are thus two cases: case one when z_1 is not zero, and case two when z_2 is not zero. When z_1 is not zero, we can write

$$(z_1, z_2) = \frac{z_1}{|z_1|} (|z_1|, w), \quad \text{where} \quad w = |z_1| \frac{z_2}{z_1},$$

is a complex number. What this shows is that the pair (z_1, z_2) of points on S^3 can be written as the product of a circle S^1 and a two sphere S^2.[7] To see this we note that $\frac{z_1}{|z_1|}$ is of fixed modulus, i.e., it represents a point on S^1 while the pair $(|z_1|, w)$, where w is a complex number and $|z_1|$, in any case, is a real number, represents a point on S^2. We can check that $|z_1|^2 + |w|^2 = 1$. Thus for z_1 non zero we have shown that S^3 can be written locally as a product of $S^1 \times D^2$. A similar argument can be given for the case when z_2 is not zero. Thus we can always represent S^3 locally as $S^1 \times D^2$. This structure of S^3 allows us to map S^3 to S^2 by mapping the S^2 hidden in S^3 to S^2. This leads to a non zero homotopy group $\pi_3(S^2, x_0)$. The intuitive picture given shows the subtle nature of homotopy maps. The corresponding cohomology group $H^3(S^2)$ is the trivial group, 0. It does not capture the subtle twisting relationship between S^3 and $S^1 \times S^2$.

In our discussion we gave two descriptions of S^3 as product spaces depending on whether z_1 or z_2 is non zero. These two descriptions when joined together give S^3 which is not a product space. Thus we have an example where gluing two product spaces gives a different space. The general study of gluing two spaces together, subject to certain conditions, to form new spaces is the subject matter of fibre bundle theory which we will consider in the next section. For the moment we draw attention to a rather elegant feature present when our two $S^1 \times D^2$ spaces are glued together to give S^3. It is clear that the joining together is not trivial. We will say a twist is involved, namely that the two S^1 circles present in the two descriptions get linked. This linking feature can be intuitively understood as in one case we have a circle in the z_1 plane while in the other the circle is in the z_2 plane.

The map $S^3 \to S^2$ described is known as the Hopf map. It shows up in many physics applications; for example, the description of all possible states that can be obtained from the superposition of two wave functions involves a Hopf map as we now show. A superposition of two normalized state to form a normalized state can be described in terms of two complex superposition numbers a, b subject to the normalization condition $|a|^2 + |b|^2 = 1$. Thus the allowed points a, b are on S^3, but in describing the superposed state there is an overall phase choice allowed. Taking this into account means (a, b) is equivalent to $(\lambda a, \lambda b)$ where $\lambda = e^{i\theta}$ is a phase. We are free to choose either λa or λb to be real. Let's choose $\lambda a = c$ where c is real. Thus $(a, b) \to \lambda^{-1}(c, w)$, with $w = \lambda b$ is a representation of S^3 in terms of $S^1 \times S^2$. This is exactly the situation we have just considered. Thus a Hopf map describes the physically inequivalent states contained in (a, b).

There are tables listing homotopy groups. Thus, in a physics problem, one need not know how to calculate homotopy groups but only to understand how and where they can appear in a physics problem. This requires spotting that in a given situation there is a map from S^n to a given space X of interest. If this happens then the group $\pi_n(X)$ will be relevant. Such maps show up very

[7] See Chapter 6 for more details.

Table 1.1: Homotopy groups of Spheres

Homotopy Group	S^1	S^2	S^3	S^4	Comment
π_1	$\mathbb{Z}$	0	0	0	Winding Number
π_2	0	$\mathbb{Z}$	0	0	Point Defect
π_3	0	$\mathbb{Z}$	$\mathbb{Z}$	0	Hopf fibration
π_4	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}$	Witten Anomaly

often as boundary conditions. If the group $\pi_n(X)$ is nontrivial, it means the system has topological features which will imply physical consequences. Thus homotopy groups are important for physics. We list a few homotopy groups, which have appeared in physics problems, in the form of a table in Table 1.1.

1.9 Fibre Bundles and Vector Bundles

We are now ready to take one more step in the mathematics of spaces by introducing the important idea of fibre bundles and vector bundles. We have explained at the beginning of the chapter how fibre bundles are natural objects present in the description of most physical systems. However their special features become important only when the physical space of interest is twisted or has holes in it. When this happens novel physical features emerge. An example of a space with a twist is provided by the topological insulator, where a time reversal system with strong spin-orbit coupling leads to the possibility of a twisted structure in momentum space. This happens because time reversal invariance leads to the degeneracy of energy levels for spin up and spin down states at zero momentum. Periodicity in momentum space implies that this degeneracy holds for momentum points $0, \pi$ that are related by periodicity. Thus a loop in momentum space starting at $k = 0$ and ending at $k = \pi$ for a particle with spin can be a space with a twist. The twist represents the spin up state at $k = 0$ changing to a spin down state at $k = \pi$. Such a space is topologically stable: it is a Möbius strip.

The topological twist can be spotted by K-groups which are constructed by introducing a way to add and subtract vector spaces so that they form a group. For the spin system with time reversal symmetry the vector space is a representation of the group $SO(3) = \frac{SU(2)}{\mathbb{Z}_2}$. A nonzero K-group has physical implications. For instance in the case of the topological insulator a nonvanishing K-group implies the presence of a zero mode, i.e., of a gapless state.

K-Theory is a generalized cohomology theory. It is used to classify vector bundles. Wave functions on a Brillouin zone are vector bundles. Vector bundles are special fibre bundles. Fibre bundles are spaces built by gluing together two spaces: A base space B (in our case the Brillouin zone) and a fibre space F

which is a vector space (in our case the space of wave functions). For a vector bundle there is a group G that acts linearly and transitively on fibre vector space, i.e., it moves points round without having fixed points and any two points in the fibre space can be connected by a group operation. The group G is called the group of the fibre.

The gluing procedure is described by first representing the base space B by a collection of contractible overlapping open sets $U_\alpha, U_\beta, ...$ so that on each open set U_α the bundle E is simply the product of the spaces U_α and F. The map from E to this product space is a map ϕ_α which describes the bundle E locally. If two open sets U_α and U_β overlap then the bundle E will have two coordinate descriptions in the overlap region. A map linking these two descriptions is required for consistency. This map $g_{\alpha|\beta}$ is called a transition function. It is the crucial step in the construction of a fibre bundle as it glues together untwisted spaces and can do so as to generate the twists and holes.

Finally, a bundle E has a projection map $\pi : E \to B$. It describes the way in which overlapping descriptions of E must be glued together to form the global bundle space E. The local product structure of the bundle E is called a "local trivialization".

In any one of these local trivializations, a point in the bundle can be described by the pair $(x, f(x))$ where x is a base point and $f(x)$ the corresponding location of the point on the fibre. This is thus a map from the base space to the bundle space. As x ranges over the base we get $f(x)$ ranging over the different fibres. The collection $f(x)$ is called a section of the bundle[8]. It is in general not a function as different open sets are needed and after a global tour on the base when one returns to the same point x one need not return to the same $f(x)$ but to a rotated $gf(x)$ where $g \in G$ is the group of the fibre. Sections represent physical information. For instance they can be wave functions. If the bundle E can be described as the product of B and F then the bundle is said to be a trivial bundle. In this case the section becomes a function. An example would be when B is a circle S^1 and the fibre is a line L, and the total space is a cylinder (a trivial bundle) while if the gluing is done with a twist we get the Möbius strip (a nontrivial bundle). If the fibre space F is a vector space then the bundle is said to be a vector bundle. If the dimension of the fibre vector space is k then we have a vector bundle of rank k. This is the basic intuitive idea of a vector bundle. Different ways of gluing lead to bundles that are not equivalent. Thus finding out all possible bundles given the base B, the fibre F, and the group of the fibre G will depend on the way the base space is represented by open sets and then how these are glued together. This is the problem of classifying a bundle.

There is a standard procedure for classifying vector bundles which reduces the classification problem to that of determining the homotopy class of maps from the base space B of the bundle of interest to a space called the classifying

[8]To be precise, $f(x)$ is a section of the localized bundle as global sections may not exist (for example, in twisted principal bundles).

space BG. [9] Fortunately we do not need to know the details of the space BG. All we will need for our calculations is G, which is the group of the fibre. Thus this result from fibre bundle theory tells us that bundles, with S^n as base space, are classified by the $(n-1)$th homotopy group $\pi_{n-1}(G)$. The point to note is that this result requires us to calculate the homotopy group not for a space but for a group. We will briefly explain how this is done by replacing the group by a space associated with it. Thus we will now show that the groups $U(1)$, $SU(2)$, $SO(3)$ can be identified with spaces $S^1, S^3, \mathbb{R}P^3$ respectively.

1.10 Groups and their manifolds

In the last section, it was stated that the class of fibre bundles over S^n with G as the group of the fibre can be identified with $\pi_{n-1}(G)$, the $(n-1)$th homotopy group of G. It is evident that the cohomology groups of G (considered as a space by ignoring the group structure) will capture some properties of the Lie group. Indeed all the classical Lie groups have spaces associated with them that completely capture their cohomological properties. For instance, $SU(3)$ is a bundle over S^5 with S^3 as the fibre but it is not the product bundle. Using methods from algebraic topology we can still extract cohomological information about $SU(3)$ in terms of S^3 and S^5. In general, a lot of the cohomological information of $SU(N)$ can be described by those of $S^3 \times ... \times S^{2N-1}$. As suggested above, we now discuss the examples of the groups $U(1)$, $SU(2)$, $SO(3)$. We start with $U(1)$.

1.10.1 $U(1)$

The group $U(1)$ is a unitary group which means that $U^\dagger = U^{-1}$, i.e., the adjoint is the inverse. For $U(1)$ the adjoint is the same as U^*, the complex conjugate as it is a 1×1 matrix. Thus we can write a general $U(1)$ group element as $g(\theta) = e^{i\theta}$ where θ is real. But we know from Euler's theorem that

$$e^{i\theta} = \cos\theta + i \sin\theta$$

Thus a general element is fixed by choosing a value for θ restricted to lie on a circle S^1 since the function representing $U(1)$ is a periodic function of θ and we have $0 \leq \theta \leq 2\pi$. Thus the space associated with the group $U(1)$ is S^1.

[9] A technical aside: The study of topological and geometric properties of classifying spaces is the subject of characteristic classes. Although cohomology groups of classical classifying spaces are known, we are interested in the homotopy class of maps from B to BG. In general, this may be a hard problem but when the base space B is a sphere the question reduces to homotopy groups of BG. However, there is an isomorphism between the homotopy groups of BG and that of G, shifted by one. Thus the question reduces to understanding the higher homotopy groups of the Lie group G.

1.10.2 $SU(2)$

The group $SU(2)$ is a 2×2 unitary matrix which has determinant equal to one. The property $U^\dagger = U^{-1}$ is satisfied by any 2×2 matrix of the form,

$$\begin{pmatrix} a & b \\ -b^* & a^* \end{pmatrix}$$

with $|a|^2 + |b|^2 = 1$, the determinant condition. But writing $a = x_1 + ix_2$ and $b = x_3 + ix_4$ the condition becomes $(x_1)^2 + (x_2)^2 + (x_3)^2 + (x_4)^2 = 1$ which is a point on S^3. Thus each point on S^3 gives an allowed $SU(2)$ matrix and is its associated manifold.

1.10.3 $SO(3)$

The group $SO(3)$ is an orthogonal 3×3 matrix that represents rotations in three dimensions. We determine the parameter space needed to describe a general rotation. We will see it is the space $\mathbb{R}P^3$. To describe a rotation we need to fix the axis of rotation and prescribe the rotation angle. The axis of rotation can be fixed by the polar angles α, β. The angle θ can be taken to be the radius of the sphere. Thus α, β give an axis direction and θ gives the length along this axis. Since $-\pi \leq \theta \leq +\pi$ the radius of the parameter space sphere is π. But a rotation along an axis by an angle $\pm\pi$ give the same point. Thus the parameter space sphere of radius π has to have the additional property that the two ends of any axis through the centre of the sphere meeting the surface of the sphere must be identified as they represent the same rotation. This is the space $\mathbb{R}P^3$.

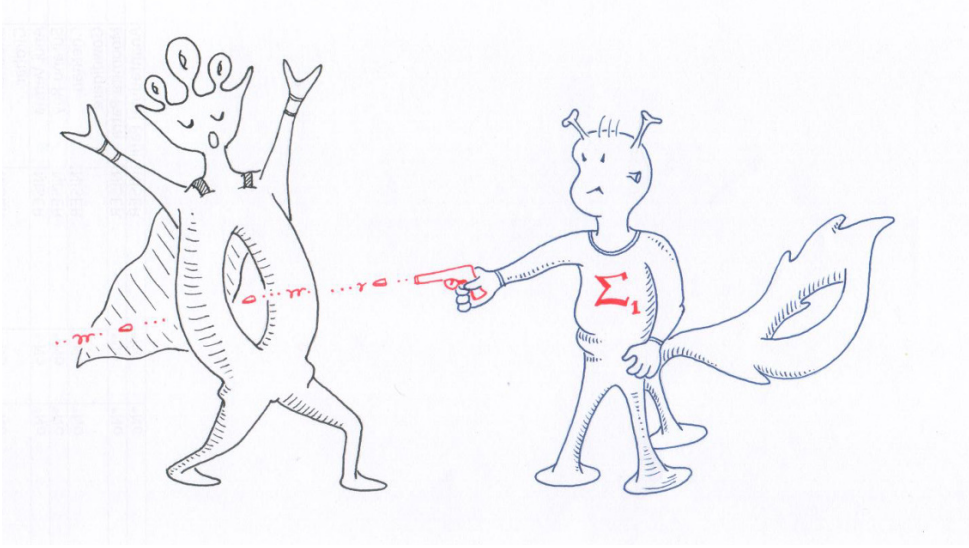
For understanding the topology of the topological insulator we need tools required to classify vector bundles when the base space M is not an n-sphere S^n since the classifying theorem requires finding the homotopy equivalence classes of maps from $M \to BG$. For example, for the topological insulator, $M = T^3$, the three torus, these maps do not give standard homotopy groups. New methods are required.

1.11 Conclusion

In this brief tour basic ideas of topology are sketched but techniques for calculating the topological invariants in the form of various groups introduced here are not described. To learn these techniques, see the following chapters of this book, while for more details books on mathematics and research papers or conference proceedings in physics need to be consulted.

Part I

Topological Tools



2

Set Topology

Somnath Basu and Atreyee Bhattacharya

In the first part we discuss metric spaces and continuous maps between them. Reformulating the notion of continuity leads to the concept of topological spaces. Some important properties of such spaces are discussed. We end with the concept of topological manifolds, as a particular class of examples of topological spaces.
The second part is a tutorial on equivalence relations and quotient sets. The main aim here is to recall a few relevant definitions in this context and more importantly make the reader comfortable with these definitions by providing a series of examples. Examples are discussed keeping in mind their significance in advanced topics of Mathematics such as topology and geometry.

I Topology: A Quick Review

2.1 Equivalence relation

Let us start with the notion of an equivalence relation. This shall be used throughout in several examples that we will encounter.

Definition 2.1.1. Let X be a set. An *equivalence relation* $\sim$ on X is an identification between elements of X satisfying:
(a) [reflexive] $x \sim x$ for any $x \in X$;
(b) [symmetric] if $x \sim y$ then $y \sim x$;
(c) [transitive] if $x \sim y$ and $y \sim z$ then $x \sim z$.
The equivalence class of x, denoted by $[x]$, is the set of all elements in X that are related to x.

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_2

The set of equivalence classes will be denoted by $X/\sim$. We shall see a host of examples and identify $X/\sim$ with standard objects from analysis, geometry and topology. However, the identifications, for now, are just bijections between sets. In later sections we shall revisit some of these examples and prove that the identifications can be upgraded to an equivalence of *topological spaces*.

Example 2.1.2. An equivalence relation on X can be specified by giving a subset of $X \times X$ satisfying the three properties in Definition 2.1.1.
(1) Let $X = \{1, 2, 3\}$ and $R = \{(1, 1), (1, 2), (2, 1), (2, 2), (3, 3)\}$. The equivalence classes are $[1] = [2]$ and $[3]$.
(2) Let $X = \mathbb{R}$ and $x \sim y$ if $x - y \in \mathbb{Z}$. Then it can be seen that $X/\sim$ is in bijective correspondence with the unit circle S^1 in $\mathbb{R}^2$. The map that sends $[t]$ to $e^{2\pi it}$ is one such bijection.
(3) Let $X = V$ be a finite dimensional vector space over $\mathbb{R}$. Given a subspace W, we define $v_1 \sim_W v_2$ if $v_1 - v_2 \in W$. This defines an equivalence relation and the resulting set $V/\sim_W$ (usually denoted by V/W) can be given the structure of a vector space. It is tempting to say that V/W is in bijection with $W^\perp$, the orthogonal complement of W in V. However, there is no such natural bijection. In case V has a positive definite inner product then the orthogonal projection of V to $W^\perp$ induces a linear isomorphism from V/W to $W^\perp$.
(4) Let $X = V - \{0\}$ with V as in (3) above. Consider the equivalence relation where $v \sim w$ if $v = \lambda w$ for some $\lambda \in \mathbb{R} - \{0\}$. The set of equivalence classes is denoted by $\mathbb{P}(V)$ and called the *space of lines* in V or *projectivization* of V.
(5) Let $X = G$ be a group. We say $g_1 \sim g_2$ if $g_1 = gg_2g^{-1}$. The equivalence classes are called the *conjugacy classes* of G. For an abelian group, the conjugacy relation is simply the reflexive relation on G.
(6) Let $\mathbb{D}^n$ denote the closed unit disk in $\mathbb{R}^n$. The equivalence relation generated by identifying all points on $\mathbb{D}^n$ of length 1 is denoted by $\sim$. Then it can be shown that $\mathbb{D}^n/\sim$ is identifiable with S^n, the unit sphere in $\mathbb{R}^{n+1}$.

These are not surprising examples but should be kept in mind. In a later section we shall expect the unfamiliar reader to go through these armed with the notion of *quotient topology*.

2.2 From metric spaces to topology

Let us recall the familiar notion of metric spaces.

Definition 2.2.1. A set X equipped with a non-negative real valued function $d : X \times X \to \mathbb{R}$ is called a *metric space* if the following holds:
(i) $d(x, y) = 0$ if and only if $x = y$;
(ii) d is symmetric, i.e., $d(x, y) = d(y, x)$;
(iii) d satisfies the *triangle inequality*, i.e., $d(x, y) + d(y, z) \geq d(x, z)$.

We shall denote a metric space by (X, d), or simply by X when d is clear in the context. The function d is called the *distance function*. To each point x

in a metric space (X, d) and every $\epsilon > 0$ we have the set

$$B_\epsilon(x) := \{y \in X \mid d(x, y) < \epsilon\}$$

called the *open ball* of radius ϵ, centered at x. Given $y \in B_\epsilon(x)$ with $r = d(x, y) < \epsilon$, we observe that $B_{\epsilon-r}(y) \subseteq B_\epsilon(x)$. When $\epsilon = 0$ then $B_\epsilon(x) = \varnothing$.

Example 2.2.2. We present several well-known examples.
(1) Let d be the Euclidean distance on $X = \mathbb{R}^n$. The open balls in this metric are actually the open balls in the usual sense.
(2) Let $X = \mathbb{R}^n$ with

$$d((x_1, \cdots, x_n), (y_1, \cdots, y_n)) := \max_i\{|x_i - y_i|\}.$$

It can be seen that the open balls (of radius ϵ) are open cubes with side length 2ϵ.
(3) Let X be any set and let $d(x, y) = 1$ if $x \neq y$. It follows that open balls are either the point or the whole space.
(4) Given a metric space (X, d) we may consider a new function

$$\mathrm{d}(x, y) := \frac{d(x, y)}{1 + d(x, y)}.$$

This defines a distance function and (X, d) becomes a metric space where the distance between any two points is uniformly bounded by 1. Note that in the new metric $B_1(x) = X$ for any $x \in X$.

Let τ_d consist of all subsets of X which are arbitrary unions of finite intersections of open balls. By construction τ_d contains $\varnothing$ and X. Moreover, it is closed under arbitrary unions. Given $U, V \in \tau_d$ one verifies that $U \cap V$ is also in τ_d. This implies that τ_d is closed under finite intersections. This collection τ_d is called the *topology induced by the metric* on X.

Remark 2.2.3. *It is well-known that the topology induced by the metrics in (1) and (2) are the same, i.e., they both have the same collection of open sets. In the case of (3) we see that $\tau_d = \mathcal{P}(X)$, the power set of X. For (4) one can check that $\tau_d = \tau_{\mathrm{d}}$.*

Before we explain why this passage from d to τ_d is important, we recall the notion of convergence of sequences and continuous maps (between metric spaces).

Definition 2.2.4. Let (X, d) be a metric space. A sequence $\{x_n\}_{n\geq 1} \subseteq X$ is said to *converge* to $x \in X$ if given $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that $d(x_n, x) < \epsilon$ for all $n \geq N$.
A map $f : (X, d) \to (Y, \mathrm{d})$ is called *continuous* if it maps convergent sequences to convergent sequences, i.e., if $\{x_n\}$ is a sequence converging to x then $\{f(x_n)\}$ converges to $f(x)$.

Note that a sequence (in a metric space) can converge to at most one point. An alternate but equivalent definition of continuity states that a map is *continuous at* x if for any $\epsilon > 0$ there exists $\delta > 0$, depending on ϵ and x, such that if $d(x, x') < \delta$ then $\mathrm{d}(f(x), f(x')) < \epsilon$. A function is called *continuous* if it is continuous at each point. If a common δ, independent of x, exists for a given ϵ then we say that f is *uniformly continuous*.

The significance of τ_d stems from the following key observation.

Proposition 2.2.5. *Let $f : (X, d) \to (Y, \mathrm{d})$ be a map between metric spaces. Then the following are equivalent:*
(1) the map f is continuous;
(2) if $U \in \tau_{\mathrm{d}}$ then $f^{-1}(U) \in \tau_d$.

Proof. Let $\{x_n\}$ be a sequence converging to x. Assuming (2) and setting $U = B_\epsilon(f(x))$ we gather that

$$f^{-1}(U) = \{w \in X \mid \mathrm{d}(f(w), f(x)) < \epsilon\} \in \tau_d.$$

Thus, $f^{-1}(U) = \cup_{i \in I} U_i$, where U_i is a finite intersections of open balls in X. Choose an U_i such that $x \in U_i$. It follows that there exists $\delta > 0$ such that $B_\delta(x) \subseteq U_i$. Moreover, there exists N such that $x_n \in B_\delta(x)$ for all $n \geq N$. In particular, $f(x_n) \in U$ for all such n.

To prove the converse it suffices to show that $f^{-1}(B_\epsilon(y)) \in \tau_d$ for any $y \in Y$. Let $x \in f^{-1}(B_\epsilon(y))$, i.e., $d(f(x), y) = r < \epsilon$. By continuity of f at x there exists δ (possibly depending on x) such that if $w \in B_\delta(x)$ then $\mathrm{d}(f(w), f(x)) < \epsilon - r$. By the triangle inequality, $\mathrm{d}(f(w), y) < \epsilon$. Consequently, $B_\delta(x) \subseteq f^{-1}(B_\epsilon(y))$ and the latter is a union of $B_\delta(x)$ as x varies. □

2.3 Topological spaces: Definition and examples

Having extracted the abstract essence of continuity, we now define topological spaces.

Definition 2.3.1. A *topological space* (X, τ) is a set X equipped with a collection τ of subsets of X satisfying the following:
(i) the empty set and X are in τ;
(ii) the collection τ is closed under finite intersections;
(iii) the collection τ is closed under arbitrary unions.
An element $U \in \tau$ is called an *open set* while its complement is called a *closed set*.

Definition 2.3.2. A map $f : (X, \tau_X) \to (Y, \tau_Y)$ between topological spaces is called *continuous* if $f^{-1}(\tau_Y) \subseteq \tau_X$. If f maps open (respectively closed) sets to open (respectively closed) sets then f is called an *open* (respectively *closed*) map.

Given two topologies τ, τ' on X, we say they are the same if $\tau = \tau'$.

Remark 2.3.3. *The condition for continuity of a function can be restated in terms of closed sets by saying that inverse images of closed sets should be closed sets.*

Observe that metric spaces are topological spaces and different metrics can induce the same topology (cf. Remark 2.2.3). A topological space (X, τ) is called *metrizable* if there is a metric which induces the topology τ. If a space X is metrizable then any two distinct points *can be separated by open sets*, i.e., given $x \neq y$ there exists $U, V \in \tau$ such that $U \cap V = \varnothing$ and $x \in U, y \in V$. This follows by setting $r = d(x, y)$, $U = B_{r/2}(x)$ and $V = B_{r/2}(y)$. Topological spaces where points can be separated by open sets are called *Hausdorff*.

Example 2.3.4. Let us consider several standard examples of topology on a given set X.
(1) The *trivial topology* on X is $\tau_{\mathrm{tr}} = \{\varnothing, X\}$. Any map $f : (Y, \tau) \to (X, \tau_{\mathrm{tr}})$ is continuous while f is open if $f(U) = X$ for every non-empty open set U.
(2) The *discrete topology* on X is $\tau_{\mathrm{dis}} = \mathcal{P}(X)$. This corresponds to Example 2.2.2 (3). Any map $f : (X, \tau_{\mathrm{dis}}) \to (Y, \tau)$ is continuous.
(3) The *cofinite topology* on X is the collection consisting of complements of finite sets along with the empty set. If X is finite then this is exactly the discrete topology. However, if X is infinite then any two non-empty open sets intersect. In this case the topology is not Hausdorff and cannot be induced by a metric. For instance, a map $f : (\mathbb{R}, \tau_{\mathrm{cofin}}) \to (\mathbb{R}, \tau_d)$ with d as in Example 2.2.2 (2) is continuous if and only if it is constant. The cofinite topology on $\mathbb{R}$ is often called the *Zariski topology*.

The usual topology on the real line has open sets which are finite or countable disjoint unions of sets of the form $(-\infty, b), (a, b), (a, \infty)$. Topological spaces can be weird in the sense that some of its properties may deviate sharply from what we are used to in Euclidean spaces and its open subsets. We present a few examples below.

Example 2.3.5. (1) The *profinite topology* on $\mathbb{Z}$ consists of open sets which are defined to be arbitrary unions of (non-constant) arithmetic progressions that extend in both directions. This topology is neither discrete nor trivial.
(2) Given a space (X, τ) and a subset $A \subseteq X$ we define the *induced topology* on A by declaring $\tau_A := \{U \cap A \,|\, U \in \tau\}$ to be a topology on A. For instance, the induced topology on $\mathbb{Q}$ from the standard real line is different than $(\mathbb{Q}, \tau_{\mathrm{dis}})$.
(3) The real line admits another, quite different, topology. Consider $X = \mathbb{R}$ and $\tau = \{(-r, r) \,|\, r > 0\} \cup \{\varnothing, \mathbb{R}\}$. This topology is not Hausdorff. Moreover, the identity map $i : (\mathbb{R}, \tau) \to \mathbb{R}$, where the latter is the standard topology on the real line, is not continuous. If we switch the topology then the identity map is continuous.
(4) Finite sets as topological spaces are rather interesting objects. There are exactly four topologies on $X = \{a, b\}$ given by $\tau_{\mathrm{tr}}, \tau_{\mathrm{cofin}}, \tau_1 = \{\varnothing, \{a\}, X\}$ and $\tau_2 = \{\varnothing, \{b\}, X\}$. A well-known example is the *pseudo-circle*, a non-Hausdorff

topological space $X = \{a, b, c, d\}$ with

$$\tau = \{\varnothing, \{a\}, \{b\}, \{a, b\}, \{a, b, c\}, \{a, b, d\}, X\}.$$

This space is weakly homotopy equivalent to the circle.

The notion of convergent sequences make sense for topological spaces as well.

Definition 2.3.6. A sequence $\{x_n\}$ in a topological space (X, τ) is said to be *convergent* if there exists $x \in X$ such that for any open set U containing x there exists $N \in \mathbb{N}$ such that $\{x_n\}_{n \geq N} \subseteq U$. We say that the sequence *converges* to x and that x is *a* limit of the sequence.

If we consider the sequence $\{1, -1, 1, -1, \cdots\}$ in Example 2.3.5 (3) then this converges to both -1 as well as 1. Although a convergent sequence may have several limits, in a Hausdorff space this limit is unique. This is one of the main reasons to work primarily with Hausdorff spaces in analytical geometry and topology.

For a subset A of a space X, the *interior* of A is defined to be the largest open set contained in A, i.e.,

$$\operatorname{Int} A = \bigcup_{U \in \tau, U \subseteq A} U.$$

The *closure* of A is defined to be the smallest closed set containing A, i.e.,

$$\overline{A} = \bigcap_{F^c \in \tau, A \subseteq F} F.$$

The *boundary* of A is the set $\overline{A} - A$. The set A is called *dense* if $\overline{A} = X$.

2.4 Topological spaces: Some key properties

In order to distinguish between topological spaces we need an appropriate notion of equivalence.

Definition 2.4.1. A map $f : (X, \tau_X) \to (Y, \tau_Y)$ is called a *homeomorphism* if f is a bijection, and both f and f^{-1} are continuous. We say two topological spaces are *equivalent* (or *homeomorphic*) if there exists a homeomorphism between the two spaces.

It can be shown that any open ball in $\mathbb{R}^n$ is homeomorphic to $\mathbb{R}^n$ (in the standard topology). The requirement that both f and f^{-1} be continuous is necessary as a continuous bijection need not have a continuous inverse. To see this, consider the unit circle S^1 with the topology inherited from $\mathbb{R}^2$. The map

$$f : [0, 2\pi) \to S^1, \ t \mapsto e^{it}$$

is continuous and bijective but f is not open. Using stereographic projection we may prove that S^1 minus a point is homeomorphic to $\mathbb{R}$. This generalizes to higher dimensions for the unit sphere S^n in $\mathbb{R}^{n+1}$. Let us now look at a slightly more interesting construction.

Example 2.4.2. Consider $\widehat{\mathbb{R}^n} := \mathbb{R}^n \sqcup \{\infty\}$ equipped with the topology

$$\tau = \tau_d \cup \{(\mathbb{R}^n - K) \cup \{\infty\} \mid K \text{ is a closed and bounded subset of } \mathbb{R}^n\}.$$

It can be shown that $\widehat{\mathbb{R}^n}$ is homeomorphic to S^n.

Several key properties which are preserved under continuity are often used to distinguish between topological spaces.

Definition 2.4.3. A topological space (X, τ) is called *connected* if X cannot be written as the disjoint union of two non-empty open sets. It is called *path-connected* if for any two points $x, y \in X$ there exists a *path* joining x to y, i.e., continuous map $\gamma : [0, 1] \to X$ such that $\gamma(0) = x, \gamma(1) = y$.

It is clear that these properties are preserved under homeomorphisms. In fact, the image of a connected (respectively path-connected) space under a continuous map is connected (respectively path-connected). Euclidean spaces with the usual metric/topology are connected and path-connected. For instance, $[0, 1]$ is connected for if $[0, 1] = U \sqcup V$ is a disconnection with $U = [0, a_0) \sqcup (a_1, a_2) \sqcup \cdots$ then $a_0 \in V$ which is open. Therefore, $(a_0 - \epsilon, a_0 + \epsilon) \subseteq V$ for some $\epsilon > 0$. This violates the assumption that $U \cap V = \varnothing$. Finally note that any connected subset of $\mathbb{R}$ is of the form

$$\varnothing,\ (a, b),\ [a, b],\ [a, b),\ (a, b],\ (a, \infty),\ [a, \infty),\ (-\infty, b),\ (-\infty, b],\ \mathbb{R},$$

where $a, b \in \mathbb{R}$ and $a \leq b$.

Remark 2.4.4. *If (X, τ) is path-connected then let $X = U \sqcup V$ be a disconnection of X, if possible. For $x \in U$ and $y \in V$ choose a path γ joining x to y. Now consider the equality $[0, 1] = \gamma^{-1}(U) \sqcup \gamma^{-1}(V)$. This is impossible as $[0, 1]$ is connected. Thus, path-connectivity implies connectivity. However, there are examples of connected spaces which are not path-connected.*

The space $\mathbb{R}^2 - \mathbb{Q}^2$ is path-connected. Euclidean spaces minus the origin is path-connected if the dimension is at least 2. However, there is a difference between punctured Euclidean spaces in terms of *higher connectivity*.

Yet another key notion is that of compactness. Recall that an open cover of a space X is a way of expressing X as the union of a collection of open sets. A finite subcover (of a given cover) simply means a finite subcollection whose union is X. We shall work with the following.

Definition 2.4.5. A topological space (X, τ) is called *sequentially compact* if every sequence $\{x_n\}$ has a convergent subsequence. A topological space (X, τ) is called *compact* if every open cover has a finite subcover.

It is a known fact from elementary real analysis that for subsets of $\mathbb{R}^n$ the two notions of compactness agree. The proof generalizes to metric spaces. Before we present some examples we note that closed subsets of compact sets are compact.

Example 2.4.6. (1) The set $(0, 1]$ is non-compact; just consider the cover given by $\{(1/n, 1]\}_{n\geq 1}$.
(2) The compact subsets of $\mathbb{R}^n$ are precisely the closed and bounded sets. This is known as the Heine-Borel Theorem.
(3) The construction outlined in Example 2.4.2 is an example of what is known as the *one-point compactifcation* of a locally compact, non-compact Hausdorff space.

Compactness is preserved under continuous maps, i.e., the image of a compact set is compact. A very useful observation (involving all of the notions we have encountered so far) is the following result.

Proposition 2.4.7. *A continuous bijection from a compact space to a Hausdorff space is a homeomorphism.*

Proof. We need only show that the map f is closed. Now any closed subset of a compact set is compact and f takes compact sets to compact sets. Finally note that a compact subset of a Hausdorff space is closed. □

2.5 Quotient topology

Suppose we have an equivalence relation $\sim$ on X. If X is equipped with a topology τ then consider the set $X/\sim$ of equivalence classes. It is natural to want the projection map $q : X \to X/\sim$ to be continuous, i.e., to find a topology τ_q on $X/\sim$ such that $q^{-1}(\tau_q) \subseteq \tau$.

Definition 2.5.1. Given an equivalence relation $\sim$ on (X, τ), the collection of sets $U \subseteq X/\sim$ such that $q^{-1}(U) \in \tau$ defines a topology called the *quotient topology.*

Observe that if τ' is a topology on $X/\sim$ such that q is continuous then τ' is a subset of the quotient topology. We say that the quotient topology is the finest topology with respect to which q is continuous.

Example 2.5.2. Let us illustrate with a few basic examples.
(1) Consider $X = [0, 1]$ and the equivalence relation $\sim$ that identifies 0 with 1, i,e., every $x \neq 0, 1$ is its own equivalence classes while $[0] = [1]$. The quotient space is Hausdorff. We define a bijective map

$$f : S^1 \to [0, 1]/\sim,\ e^{2\pi it} \mapsto [t].$$

We can check that this is continuous. It follows from Proposition 2.4.7 that f is a homeomorphism.

The example above generalizes to higher dimensions. If $\mathbb{D}^n$ denotes the

closed unit disk in $\mathbb{R}^n$ then the equivalence relation $\sim$ that identifies all its boundary points together induces a homeomorphism between S^n and $\mathbb{D}^n/\sim$ (cf. Example 2.1.2 (6)).
(2) The previous example is a special case of the following identification. Given a subset A of (X,τ) we define the equivalence relation $\sim_A$ that identifies all points of A together. We shall denote the quotient space by X/A. In general, this space may fail to be Hausdorff even if X is. For instance, let $X=\mathbb{R}$ with $A=\mathbb{R}-\{0\}$. Then $X/A=\{[0],[1]\}$ has $[0]$ as closed and $[1]$ as open. Perhaps a stranger example arises out of $X=\mathbb{R}$ and $A=\mathbb{Q}$. The quotient space is an uncountable set with a distinguished point $*$ where every point other than $*$ is closed. Moreover, any open set containing $\alpha\neq *$ must be the whole set.
(3) Consider $X=[0,1]\times[0,1]$ with the equivalence relation that identifies $(0,t)$ with $(1,1-t)$. The quotient space can be realized as a subspace inside $\mathbb{R}^3$ and is called the *Möbius strip* (cf. Figure 2.2).

We note that if (X,τ) is connected (respectively path-connected) then $X/\sim$ is connected (respectively path-connected). Compactness is also preserved under taking quotients. As observed in Example 2.5.2 a quotient of a Hausdorff space need not be Hausdorff. Groups acting on spaces provide a plethora of examples of quotient topology.

To discuss actions of groups let us recall a few notions.

Definition 2.5.3. A *topological group* G is a topological space (G,τ) such that the underlying set G is a group, and the inverse and multiplication maps are continuous.

Any group is a topological group with the discrete topology. The first examples of (non-discrete) topological groups are the Euclidean spaces with vector addition as the group operation. The set of positive real numbers, under multiplication, is also a (topological) group. It can be seen that S^1 and $\mathbb{C}^\times$ are both topological groups. In fact, $\mathbb{C}^\times$ is isomorphic (as topological groups) to $S^1\times(0,\infty)$.

Example 2.5.4. Consider the set $GL_n(\mathbb{R})$ of invertible real $n\times n$ matrices. This is a (discrete) group with several distinguished subgroups including $SL_n(\mathbb{R})$ and $O_n(\mathbb{R})$. We may equip $GL_n(\mathbb{R})$ with the subspace topology induced from $M_n(\mathbb{R})$ which can be identified with $\mathbb{R}^{n^2}$. The determinant, being a polynomial in entries, is continuous. Thus, the set $GL_n(\mathbb{R})=\det^{-1}(\mathbb{R}-0)$ is open in the space of all $n\times n$ matrices. The multiplication of matrices is a polynomial map on $\mathbb{R}^{n^2}$. This implies that $GL_n(\mathbb{R})$ is actually a topological group.

Since the topology on $M_n(\mathbb{R})$ is induced by the Euclidean metric on $\mathbb{R}^{n^2}$, the group $GL_n(\mathbb{R})$ is metrizable. In fact, $GL_n(\mathbb{R})$ is open and dense in $M_n(\mathbb{R})$. It is known that $SL_n(\mathbb{R})$ is non-compact and is evident from the sequence of diagonal matrices A_k with $a_{11}=k, a_{22}=k^{-1}$ and $a_{ii}=1$. However, $O_n(\mathbb{R})$ can be described as the common zeroes of $n(n+1)/2$ polynomials of degree 2.

Thus, $O_n(\mathbb{R})$ is a closed subset and is contained in

$$\big\{(x_{11},\dots,x_{1n},x_{21},\dots,x_{2n},\dots,x_{n1},\dots,x_{nn})\in\mathbb{R}^{n^2}\;\big|\;\sum_i x_{ji}^2=1 \text{ for any } j\big\}.$$

The space above is the n-fold product $S^{n-1}\times\cdots\times S^{n-1}$. Thus, $O_n(\mathbb{R})$ is compact.

Definition 2.5.5. We say a (topological) group G acts on a topological space X (from the left) if there exists a continuous map $\varphi : G\times X\to X$ satisfying $\varphi(e,\cdot)=\mathrm{id}_X$ and $\varphi(gh,x)=\varphi(g,\varphi(h,x))$ for any $g,h\in G$ and $x\in X$.

Note that when G is discrete then a group action is equivalent to a group homomorphism from G to $\mathrm{Homeo}(X)$, the group of all homeomorphisms of X. In general, a group action induces a natural equivalence relation on X; the quotient space is denoted by X/G.

Example 2.5.6. (1) Let $\mathbb{Z}$ act on the real line by translations. Then $\mathbb{R}/\mathbb{Z}$ is homeomorphic to S^1. A word of caution: we have earlier used the notation X/A in Example 2.5.2 (2). If we think of $\mathbb{Z}$ as a subset of $\mathbb{R}$ then $\mathbb{R}/\mathbb{Z}$, in our old notation, is not S^1. In a similar way, we can define the action of $\mathbb{Z}\times\mathbb{Z}$ on $\mathbb{R}^2$ and the quotient is the torus $S^1\times S^1$ (cf. Figure 2.1).

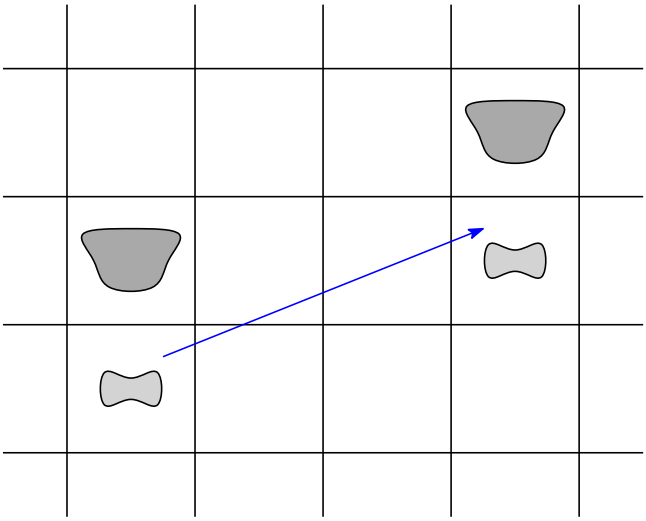


Figure 2.1: The action of the element $(3,1)\in\mathbb{Z}^2$ on the plane.

(2) Consider the $\mathbb{Z}/2\mathbb{Z}$-action on $X\times X$ given by $(x,y)\mapsto(y,x)$. The diagonal $\Delta:=\{(x,x)\in X\times X\,|\,x\in X\}$ is fixed pointwise by this action. When $X=\mathbb{R}$ we observe that the quotient space is homeomorphic to the upper half plane $\{(x,y)\in\mathbb{R}^2\,|\,y\geq 0\}$. When $X=S^1$ we can check that the quotient space is the Möbius strip (cf. Figure 2.2).

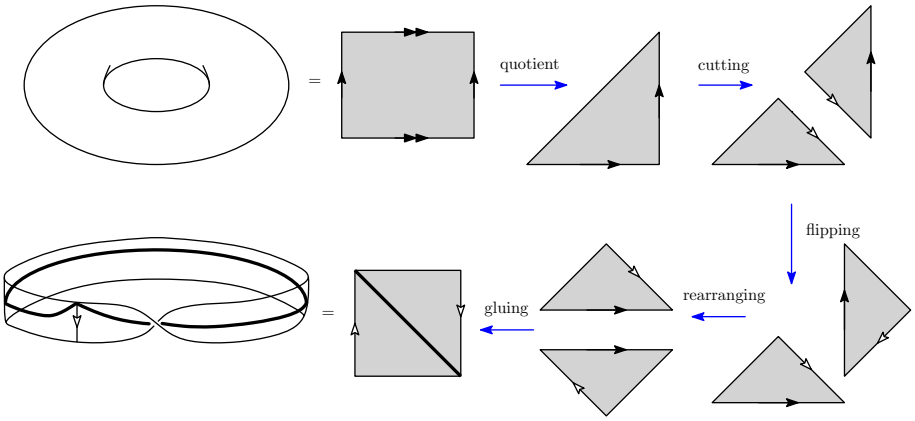


Figure 2.2: The quotient of the torus under the reflection map.

(3) The group $\mathbb{Z}/n\mathbb{Z}$ acts on S^1 via the map $([k], z) \mapsto e^{2\pi ik/n}z$. It is clear that $S^1/(\mathbb{Z}/n\mathbb{Z})$ is again the circle.
(4) The group $\mathbb{Z}/2\mathbb{Z}$ acts on S^n via identifying a point with its antipode. The quotient space is denoted by $\mathbb{R}P^n$ and called the *real projective space* of dimension n.
(5) Consider the action of $SO_2(\mathbb{R})$ on $\mathbb{R}^2$. The origin is fixed by the group while all points at a distance d away from the origin are identified with each other via elements of $SO_2(\mathbb{R})$. The quotient space is $[0, \infty)$.
(6) Consider the group $PSL_2(\mathbb{Z}) = SL_2(\mathbb{Z})/\pm I$ acting on the upper half-plane

$$\mathbb{H} := \{x + iy \in \mathbb{C} \,|\, y > 0\}$$

by $z \mapsto \frac{az+b}{cz+d}$. Here an element $A \in SL_2(\mathbb{Z})$ has entries a, b, c, d. This group and its action is important in many branches of mathematics. This group is called the *modular group* and is generated by $T(z) = z + 1$ and $S(z) = -z^{-1}$. The quotient space can be shown to be a 2-sphere minus one point.

2.6 Topological manifolds

We are ready to discuss a class of topological spaces which form the most prevalent family in algebraic and differential topology. We shall need a few preliminaries.

Definition 2.6.1. A *base* for a topology τ on X is a subcollection β such that any $U \in \tau$ can be written as a union of elements from β. A space with a countable base is called *second countable*.

Euclidean spaces are second countable. In fact, most spaces we encounter in analysis and geometry are second countable.

Definition 2.6.2. A *topological manifold* M is a topological space which is Hausdorff, second countable and for each point there exists an open set which is homeomorphic to an open ball in $\mathbb{R}^n$ for some n. A pair (U, φ) consisting of such an open set and a homeomorphism $\varphi : U \to B \subset \mathbb{R}^n$ is called a *chart*.

Remark 2.6.3. *What we have defined is a manifold without boundary. We often discuss spaces which are locally like $\mathbb{R}^n$ or the upper half-plane in $\mathbb{R}^n$. The points where a local chart looks like the latter is called a boundary point. Closed disks and closed upper half-planes are examples of such spaces, also called manifolds with boundary. A manifold without boundary may be thought of as a manifold with empty boundary.*

Proposition 2.6.4. *A connected topological manifold is also path-connected.*

Proof. Fix $x \in M$ and consider the set S of all points that can be joined by a path to x. As $x \in S$, it is non-empty. If $y \in S$ then a small open path-connected set containing y is also in S, whence S is open. If $S^c \neq \varnothing$ then a similar argument proves that S^c is open, contradicting the connectedness of M. □

If M is connected then n is constant and called the *dimension* of the manifold. This follows from proving *invariance of dimension*, i.e., that $\mathbb{R}^m$ and $\mathbb{R}^n$ cannot be homeomorphic if $m \neq n$. It is one of the basic and non-trivial facts of (algebraic) topology.

The subset of $\mathbb{R}^2$ given by the union of the x and y-axis is not a manifold as the origin does not admit a neighbourhood which looks like $\mathbb{R}$. However, there are examples of spaces which are locally Euclidean but fail to be non-Hausdorff.

Example 2.6.5. We may consider $X = \mathbb{R} \sqcup \{*\}$ equipped with the topology

$$\tau = \tau_d \cup \{(U - 0) \cup \{*\} \,|\, U \text{ is an open set in } \mathbb{R} \text{ containing } 0\}.$$

This space is called the *line with two origins*. Every point in $\mathbb{R} \subset X$ has an open set, homeomorphic to $(-1, 1)$, containing it. As for $*$, the set $((-1, 1) - 0) \cup \{*\}$ is homeomorphic to $(-1, 1)$. However, 0 and $*$ cannot be separated by open sets.

Most of what we had seen in earlier sections are examples of manifolds - spheres, real projective spaces, $GL_n(\mathbb{R})$, $O_n(\mathbb{R})$. Configuration spaces are examples of manifolds which are useful in applications as well as in theory. For example, the configuration space of a rod rotating in $\mathbb{R}^3$ about a fixed hinge at the midpoint of the rod is $\mathbb{R}P^2$. In practice, one finds the following result quite useful.

Proposition 2.6.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. If c is a non-critical value of f then $f^{-1}(c)$ is a manifold of dimension $n - 1$.*

This can be proven using the Implicit Function Theorem from several variable calculus. Note that $f^{-1}(c)$ can be empty. The empty set is also considered

a manifold of any given dimension. The above result implies, for instance, that $SL_n(\mathbb{R})$ is a manifold.

Example 2.6.7. The Grassmannians are manifolds which appear in several branches of mathematics and are of considerable importance. For a fixed n and $1 \leq k \leq n$ the *Grassmann manifold* $G_{k,n}$ (over $\mathbb{R}$) is defined to be the set of linear k-planes in $\mathbb{R}^n$. The topology is induced by the metric

$$d(L, L') = \|P_L - P_{L'}\|$$

where P_L denotes the orthogonal projection of $\mathbb{R}^n$ to the subspace L and $\|A\|$, for a linear map $A : \mathbb{R}^n \to \mathbb{R}^n$, is defined to be

$$\|A\| = \max_{v \in S^{n-1}} \|Av\|.$$

In other words, $\|A\|$ is the maximum possible size of the image of any unit vector under A. In fact, it is known that $G_{k,n}$ is homeomorphic to $O_n(\mathbb{R})/(O_k(\mathbb{R}) \times O_{n-k}(\mathbb{R}))$. Observe that when $k = 1$ the Grassmannian is the real projective space $\mathbb{R}P^{n-1}$ (cf. Example 2.5.6 (4)). In fact we had encountered this even earlier in Example 2.1.2 (4) as $\mathbb{P}(\mathbb{R}^n)$.

We finally come back a full circle to metric spaces. The question of when a topological space is metrizable, i.e., it admits a metric with the induced topology being the given topology, has been answered in various forms. For us, the following is of relevance.

Theorem 2.6.8 (Urysohn's metrization theorem)**.** *Every Hausdorff, second-countable, regular space is metrizable.*

Recall that a space (X, τ) is called *regular* if points and closed sets can be separated by open sets. More precisely, given x and a closed set F not containing x there exists $U, V \in \tau$ such that $x \in U, F \subseteq V$ and $U \cap V = \varnothing$. Since a manifold can be shown to be regular, it is metrizable.

II A Tutorial on Equivalence Relations and Quotient Sets

Equivalence relation and quotient sets

In English, we say that two objects are *related* if there is some rule connecting them together. For example, two persons are said to be related by 'blood' if they belong to the same family, by 'friendship' if they are friends of each other, by 'nationality' if they come from the same country. Thus one way to describe a *relation* would be to collect all possible pairs of objects connected by the same relation. Mathematically a relation on a set X can be viewed as a way of identification between elements of X and can be described by a collection of ordered pairs of elements of X. The ordering is important as two pairs (a, b) and (b, a) do not necessarily represent the same relation; e.g., if a is the mother of b, then the relation between (a, b) and (b, a) are different.

Definition 2.6.1. Binary relations: Let X be a non-empty set. A *binary relation* R on X is defined as a subset of the cartesian product $X \times X$ or equivalently a collection of ordered pair of elements of X. Two elements $x, y \in X$, are said to be related by R if $(x, y) \in R$. In that case, one writes $x\ R\ y$.

Example 2.6.2. (a) On the set of real numbers $\mathbb{R}$, the following well known (binary) relations, e.g., *greater than* ($>$), *greater than or equal to* ($\geq$), *less than* ($<$), *less than or equal to* ($\leq$), *is equal to* ($=$) and *divides*, can be represented by the respective subsets $R_>$, $R_\geq$, $R_<$, $R_\leq$, $R_=$ and $R_/$ of $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ as described below.

$$\begin{aligned}
R_< &= \{(x, y) \in \mathbb{R}^2 : x > y\},\\
R_\geq &= \{(x, y) \in \mathbb{R}^2 : x \geq y\},\\
R_< &= \{(x, y) \in \mathbb{R}^2 : x < y\},\\
R_\leq &= \{(x, y) \in \mathbb{R}^2 : x \leq y\},\\
R_= &= \{(x, x) \in \mathbb{R}^2 : x \in \mathbb{R}\} \text{ and}\\
R_/ &= \{(x, y) \in \mathbb{R}^2 : x \text{ divides } y\}.
\end{aligned}$$

(b) Let $P(X)$ denote the power set of a set X (i.e., $P(X)$ is the collection of all subsets of X). On the set $P(X)$, one can define the binary relation $R_\subset (\subset P(X) \times P(X))$ by

$$R_\subset = \{(A, B) \in P(X) \times P(X) : A \subset B\}.$$

(c) For a given positive integer n, one can define the binary relation called *congruence modulo* n on the set of integers $\mathbb{Z}$, as follows. Two integers a and b are said to be *congruent modulo* n, often denoted by $a \equiv b \pmod{n}$, if $a - b$ is divisible by n. For example, 23 and 73 are congruent modulo 10. Thus one can describe the relation as the subset $R_{\equiv(n)}$ of $\mathbb{Z} \times \mathbb{Z}$ defined by

$$R_{\equiv(n)} = \{(a, b) \in \mathbb{Z} \times \mathbb{Z} : a \equiv b \pmod{n}\}.$$

(d) Let $(V, \langle .,. \rangle)$ be an inner product space. Given two vectors $v, w \in V$, v is said to be *orthogonal to* w, denoted by $v \perp w$, if $\langle v, w \rangle = 0$. This defines the binary relation *orthogonal to* (on V,) described by the subset $R_\perp$ of $V \times V$ defined by

$$R_\perp = \{(v, w) \in V \times V : v \perp w\}.$$

Definition 2.6.3. Equivalence relations: A binary relation R on a set X is said to be an *equivalence relation* if the following conditions hold:

- R is *reflexive*, i.e., $(x, x) \in R$ for each $x \in X$.
- R is *symmetric*, i.e., if $(x, y) \in R$, then $(y, x) \in R$.
- R is *transitive*, i.e., if $(x, y), (y, z) \in R$, then $(x, z) \in R$.

Example 2.6.4. (a) $R_=$ and $R_{\equiv(n)}$ described in Example 2.6.2, are both examples of equivalence relations whereas none of the relations $R_>$, $R_\geq$, $R_<$, $R_\leq$, $R_/$, $R_\subset$ or $R_\perp$ in Example 2.6.2 is. Clearly, $R_>$, $R_<$ are neither reflexive nor symmetric; $R_\perp$ is neither reflexive nor transitive and $R_\geq$, $R_\leq$, $R_/$, $R_\subset$ are not symmetric.

(b) Consider a (real or complex) vector space V and let $X = V \setminus \{0\}$. Define the binary relation R_p by

$$R_p = \{(v, w) \in X \times X : w = \lambda v \text{ for some } \lambda \in \mathbb{K}\}$$

where $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ if V is a real or complex vector space respectively. Naturally, if $(v, w) \in R_p$, then $w = \lambda v$ and $\lambda (\in \mathbb{K}) \neq 0$ as both v and w are non-zero vectors. Then one can re-write R_p as

$$R_p = \{(v, \lambda v) : v \in X \text{ and } \lambda \in \mathbb{K} \setminus \{0\}\}$$

and it is easy to see that R_p defines an equivalence relation on X.

(c) Let V be a real vector space and W be any linear subspace of V. Then one can define the binary relation R_W on V by

$$R_W = \{(v_1, v_2) \in V \times V : v_1 - v_2 \in W\}.$$

Using the fact that W is also a vector space, it follows that R_W is an equivalence relation.

(d) We say that a group $(G, .)$ *acts* on a non-empty set X (or X is a G-*set*) if there is a map $\rho : G \times X \to X$ satisfying

$$\begin{aligned} \rho(g_1.g_2, x) &= \rho(g_1, \rho(g_2, x)) \ \forall x \in X \text{ and } g_1, g_2 \in G; \\ \rho(e, x) &= x \ \forall x \in X. \end{aligned} \tag{2.1}$$

e is the identity element of G. Then ρ is called an action of G on X. Define the binary relation R_G on X by

$$R_G = \{(x, \rho(g, x)) \in X \times X : x \in X \text{ and } g \in G\}.$$

It can be shown (using the properties of ρ) that R_G is an equivalence relation.

For example, the group of integers $(\mathbb{Z}, +)$ acts on the set of real numbers $\mathbb{R}$ by translation. More generally, for all $n \in \mathbb{N}$, the product group $\mathbb{Z}^n$ acts on the Euclidean n-space $\mathbb{R}^n$ by translation i.e., there is a natural map $\rho : \mathbb{Z}^n \times \mathbb{R}^n \to \mathbb{R}^n$ defined by $\rho(a, x) = a + x$ for all $a \in \mathbb{Z}^n$ and $x \in \mathbb{R}^n$; satisfying the conditions (2.1). Using this action of $\mathbb{Z}^n$ on $\mathbb{R}^n$, one obtains the equivalence relation $R_{\mathbb{Z}^n}$ given by

$$R_{\mathbb{Z}^n} = \{(x, a + x)) : x \in \mathbb{R}^n \text{ and } a \in \mathbb{Z}^n\}.$$

Definition 2.6.5. Equivalence classes: Let (X, R) denote a set X together with an equivalence relation R (on X). Given any element x_0, the *equivalence class* of x_0 denoted by $[x_0]$ is a subset of X defined by

$$[x_0] = \{x \in X : (x, x_0) \in R\}$$

i.e., $x \in [x_0]$ if and only if x is related to x_0 under the relation R.

From the definition of equivalence class, it follows that given a set X together with an equivalence relation R and any two elements $x, y \in X$, $[x] = [y]$ if and only if x is related to y by R. In such a case, the elements x and y are said to be *equivalent* under the relation R.

Example 2.6.6. (a) For the equivalence relation $R_=$ as in Example 2.6.4(a), the equivalence class $[x_0]$ for any $x_0 \in \mathbb{R}$, is the singleton subset $\{x_0\}$ of X.

For the equivalence relation $R_{\equiv(n)}$, check that there are only n distinct equivalence classes, namely, $[0], [1], ..., [n-1]$.

(b) Let V and X be as in Example 2.6.4(b). Given any vector $v \in X$, its equivalence class corresponding to the relation R_p is the subset

$$[v] = \{\lambda v : \lambda \in \mathbb{R}, \lambda \neq 0\}.$$

Thus if the vector space V is $\mathbb{R}^n$ or $\mathbb{C}^n$, the equivalence class $[v]$ is the line (real or complex) in $\mathbb{R}^n$ or $\mathbb{C}^n$ respectively, passing through the origin and the vector v with the point 0 removed. Moreover, it is easy to see that in this case, two vectors $v, w \in X$ will be equivalent under R_p if and only if they both lie on the same straight line (real or complex) passing through the origin in $\mathbb{R}^n$ or $\mathbb{C}^n$ respectively.

(c) Consider the equivalence relation R_W on a vector space V corresponding to a given subspace W of V described in Example 2.6.4(c). It is easy to see that with respect to the relation R_W, the equivalence class $[0]$ of the null vector $0 \in V$, is the subspace W, as using the definitions of R_W and the equivalence class $[0]$, it follows that $v \in [0]$ if and only if $(v - 0) = v \in W$. Also each vector

in W is equivalent to the null vector 0 under R_W and hence $[w] = [0] = W$ for all $w \in W$. For a vector $v \in V$ which lies outside W, the equivalence class $[v]$ is different from W (as v is not equivalent to any vector in W). In this case, one can check that the equivalence class is $[v] = v + W = \{v + w : w \in W\}$ which is an affine subspace of V.

(d) As discussed in Example 2.6.4(d), consider a G-set X for some group G and let ρ be the corresponding group action. Then for any x_0, its equivalence class $[x_0]$ with respect to the relation R_G, is the following subset of X described by

$$[x_0] = \{\rho(g, x_0) : g \in G\}.$$

Thus when $X = \mathbb{R}^n$, $G = \mathbb{Z}^n$ and the group action ρ is just the usual translation, then for any x_0, its equivalence class is

$$[x_0] = \{a + x_0 : a \in \mathbb{Z}^n\}$$

i.e., $[x_0]$ consists of all translates of x_0 by the elements of $\mathbb{Z}^n$. Also in this case, any two elements $x_0, y_0 \in \mathbb{R}^n$, will be equivalent or $[x_0] = [y_0]$ if and only if $(x_0 - y_0) \in \mathbb{Z}^n$.

Definition 2.6.7. Quotient sets: Let (X, R) denote a set X together with an equivalence relation R (on X). The set of all equivalence classes of X with respect to the relation R, denoted by X/R and defined by $X/R := \{[x] \mid x \in X\}$, is said to be the *quotient set* of X by the relation R.

If X has some additional (algebraic/topological/geometric) properties, there are standard techniques of transferring these properties to the quotient set X/R (details of which we avoid in this note). For example, if X is a topological space, one can transform X/R also into a topological space in a canonical manner.

Example 2.6.8. (a) For the equivalence relation $R_=$ (cf. Example 2.6.4(a)), the quotient set X/R (here $X = \mathbb{R}$) is the same as the set of real numbers (one can easily see that in this case, the map which sends each real number to its equivalence class under $R_=$, defines a bijection of sets).

The quotient set for the relation $R_{\equiv(n)}$, commonly denoted by $\mathbb{Z}/n\mathbb{Z}$ or $\mathbb{Z}/n$ is the finite set $\mathbb{Z}/n\mathbb{Z} = \{[0], [1], .., [n-1]\}$. This quotient set $\mathbb{Z}/n\mathbb{Z}$ has several important mathematical properties that are related to many branches of Mathematics.

(b) In Example 2.6.4(b), the quotient set for the equivalence relation R_p is

$$X/R = \{[v] : v \in X\}$$

which consists of all one dimensional subspaces of V minus the null vector. Thus if V is $\mathbb{R}^n$ (respectively $\mathbb{C}^n$), the corresponding quotient set is the set of all real (complex) lines minus the origin in $\mathbb{R}^n$ ($\mathbb{C}^n$) passing through the origin. The quotient space has very rich topological and geometric structures and are known as the real (complex) projective space.

(c) For the equivalence relation R_W on a vector space V, (cf. Example 2.6.4(c)), it is easy to see that the quotient set V/R_W (commonly denoted by V/W), is a vector space of dimension $\dim V - \dim W$ with addition and scalar multiplication of vectors in V/W given by

$$\begin{aligned} [v_1] + [v_2] &= [v_1 + v_2] \ \forall\ v_1, v_2 \in V, \text{and} \\ c[v_1] &= [cv_1] \ \forall\ v_1 \in V,\ c \in \mathbb{K} \end{aligned}$$

($\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ if V is a real or complex vector space respectively).

If V is equipped with an inner product $\langle .,. \rangle$, then there is a nice description for the quotient space V/W as follows. Recall that in this case, for any subspace W of V, $(V, \langle .,. \rangle)$ can be decomposed as the *orthogonal direct sum* $V = W \oplus W^{\perp}$ (i.e., each $v \in V$ has a unique expression $v = v_0 + v_1$ such that $v_0 \in W$ and $v_1 \in W^{\perp}$) where $W^{\perp} = \{v \in V : \langle v, w \rangle = 0 \ \ \forall\ w \in W\}$ is the subspace of V called the *orthogonal complement* of W in V and $\dim W^{\perp} = \dim V - \dim W$. Also recall the *orthogonal projection* map $P : V \to W^{\perp}$ which is a surjective linear map sending any vector $v = v_0 + v_1 \in V$ (as described above), to v_1. It can be checked that there is a canonical isomorphism between the vector spaces V/W and $W^{\perp}$. As both the vector spaces are of the same dimension, it suffices to construct an injective linear map between them. Define $L : V/W \to W^{\perp}$ by $L([v]) = P(v)$, for $v \in V$.

Then L is well defined (i.e., for any two vectors $u, v \in V$ with [u]=[v] (or $(u-v) \in W$), one has $L([u]) = L([v])$): In fact, $(u-v) = (u_1-v_1)+(u_0-v_0) \in W$ where $u = u_0+u_1$ and $v = v_0+v_1$ with $u_0, v_0 \in W$ and $u_1, v_1 \in W^{\perp}$ respectively, implies that $u_1 = v_1$ and thus $P(u) = P(v)$.
It is easy to see that L is a linear map. Injectivity of L follows from the fact that $L([v]) = P(v) = 0$ if and only if $v \in W$ (property of an orthogonal projection) or equivalently, $[v] = 0$.

(d) Consider a G-set X for some group G (cf. Example 2.6.4(d)) and let ρ be the corresponding group action. The quotient space consisting of all equivalence classes with respect to the relation R_G, is commonly denoted by X/G.

In particular, when $X = \mathbb{R}$, $G = \mathbb{Z}$ and the group action ρ is given by $\rho(a, x) = a + x$, $x \in \mathbb{R}$ and $a \in \mathbb{Z}$; then the quotient space $\mathbb{R}/\mathbb{Z}$ is the unit circle $S^1 = \{z \in \mathbb{C} : |z| = 1\}$ in the complex plane via the bijection $f : \mathbb{R}/\mathbb{Z} \to S^1$ defined by $f([x]) = e^{2\pi i x}$, for $x \in \mathbb{R}$. Clearly, f is well defined i.e., for any two equivalent elements $x, y \in \mathbb{R}$ (i.e., [x]=[y] and $(x - y) \in \mathbb{Z}$,), one has $f([x]) = f([y])$.

f is surjective by construction. Finally f is injective as $f([x]) = f([y])$ implies that $e^{2\pi i(x-y)} = 0$ i.e., $(x - y) \in \mathbb{Z}$ and hence $[x] = [y]$. Similarly, for $X = \mathbb{R}^n$, $G = \mathbb{Z}^n$ $(n \geq 1)$ and the group action ρ given by $\rho(a, x) = a + x$, $x \in \mathbb{R}^n$ and $a = (a_1, ...a_n) \in \mathbb{Z}^n$; the quotient space $\mathbb{R}^n/\mathbb{Z}^n$ is the n-torus $T^n = S^1 \times S^1 \times .... \times S^1$($n$-times) via the bijection $F : \mathbb{R}^n/\mathbb{Z}^n \to T^n$ defined by $f([x]) = (e^{2\pi i x_1}, .., e^{2\pi i x_n})$ for $x = (x_1, .., x_n) \in \mathbb{R}^n$. It turns out that these quotient sets have very special structures from algebraic, topological as well as geometric points of view.

References

[1] A. Hatcher, *Algebraic topology*. (Cambridge University Press, Cambridge, 2002).

[2] J. R. Munkres, *Topology: a first course*. (Prentice-Hall, Inc., Englewood Cliffs, N.J., 1975).

[3] M. Nakahara, *Geometry, topology and physics*. (Taylor & Francis, Boca Raton, FL, USA, 2003).

3

Homotopy theory

Samik Basu and Soma Maity

In the first part, the fundamental group is defined using loops in topological spaces, which is the first of a series of invariants called homotopy groups. Unlike other homotopy groups, these groups are non-Abelian. However, these are computable in many examples. In this chapter, we discuss some properties of Fundamental groups and some computations.
Higher dimensional analogues of the above involve maps out of higher dimensional spheres and the resulting invariants are called homotopy groups. In the second part, we define homotopy groups and list some of the main computations.
In a tutorial section, the fundamental groups of spheres and real projective spaces are worked out and a few examples of group action discussed.

I The Fundamental Group

3.1 Introduction

In the basic topology section, we have seen a large list of examples of topological spaces. These include the Euclidean space $\mathbb{R}^n$, the n-disk $\mathcal{D}^n$, the n-sphere S^n, the n-dimensional real projective space $\mathbb{R}P^n$, surfaces (of genus g), their products, subspaces, and quotients under equivalence relations (as a special case, orbit spaces of group actions). In topology, we try to answer natural questions about topological spaces such as these; for example, we would like show that two different spaces from the list above are not homeomorphic. This

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_3

involves the careful definition of invariants and their computation in important cases. There are two families of invariants of particular importance –

1. Homotopy groups : π_n for $n \geq 1$. For $n = 1$, π_1 is referred to as the fundamental group and is non-abelian in general. For $n \geq 2$, the group $\pi_n(X)$ is abelian.
2. Homology and cohomology groups : denoted by H_n, H^n for $n \geq 0$. These groups are always abelian.

The first family of invariants are easy to define but very hard to compute, especially, when $n \geq 2$. The second family of invariants have an involved description but their computation is easy. This note is a brief introduction to the fundamental group.

3.2 Paths and loops in a topological space

A path in a topological space is a continuous curve between two points as in Fig. 3.1 (see below). We think of this as a path in the space X from the point x_0

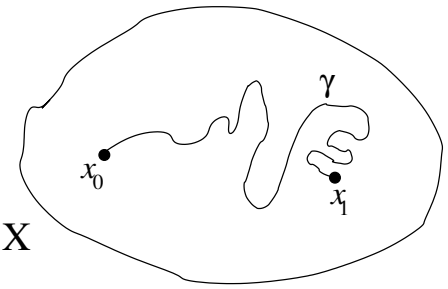


Figure 3.1: A path joining points in a space.

to the point x_1. More precisely, the path shown in the diagram is the image of a continuous function from an interval to X and the end points of the interval are mapped to x_0 and x_1. For certain technical reasons, we define a "path in a topological space" as the underlying function as opposed to its image.

Definition 3.2.1.
- A path in a topological space X is a continuous function $\gamma : [0, 1] \to X$.
- The initial point of γ is defined as $\gamma(0)$. The final point of γ is defined as $\gamma(1)$.
- If $\gamma(0) = x_0$ and $\gamma(1) = x_1$, γ is said to be a path from x_0 to x_1.

Note that this implies a path means not just the figure above but also contains the information about how one moves along the path, that is, one may define velocity, acceleration along the path. However, for our purposes in this lecture we consider different properties of paths.

Similar to a path, a loop in a space may be viewed as a path from x_0 to x_0 as shown in Fig. 3.2 below.

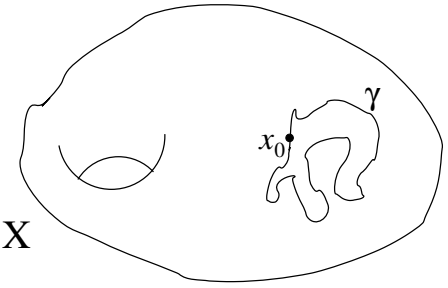


Figure 3.2: Loop based at x_0

Definition 3.2.2. A loop based at x_0 is a path γ from x_0 to x_0, that is, $\gamma : [0,1] \to X$ is a continuous function with $\gamma(0) = \gamma(1) = x_0$.

Example 3.2.3. *The circle $x^2 + y^2 = 1$ may be described as a loop based at 1 by $\gamma : [0,1] \to \mathbb{R}^2$ defined according to the equation $\gamma(t) = e^{2\pi it}$ identifying the 2-plane with the complex numbers.*

Example 3.2.4. *The formula $\gamma(t) = (a \cos 2\pi t, b \sin 2\pi t)$ for positive real numbers a and b defines a loop based at $(a,0)$ which traces out the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$.*

Example 3.2.5. *Since constant functions to any topological space are continuous the formula $c(t) = x_0$ defines a loop based at x_0. This loop is called the* constant loop.

The basic idea of the fundamental group is to consider loops in a space upto continuous deformation. As an example, consider the unit circle γ in the example above. As a loop based at $(1,0)$ in $\mathbb{R}^2$, γ may be continuously deformed to c, the constant loop based at $(1,0)$, as demonstrated in Fig 3.3. The value of the deformation at time s may be taken as $\gamma_s(t) = ((1-s) \cos 2\pi t + s, (1-s) \sin 2\pi t)$ which is γ at $s = 0$ and the constant loop at $s = 1$.

Now consider the space $\mathbb{R}^2 - 0$, and note that both γ and c are well-defined loops based at $(1,0)$. However the deformation above cannot be defined in $\mathbb{R}^2 - 0$ because some intermediate loop would have to pass through 0. For the formula above this happens at $s = \frac{1}{2}$ and $t = \frac{1}{2}$. We will see later that γ cannot be deformed to the constant loop. In this way we notice how removing the point 0 from the plane changes the topology.

The idea of continuous deformations is encapsulated in the concept of homotopy. (See Fig. 3.4.)

Definition 3.2.6. Two paths γ, γ' from x_0 to x_1 are said to be homotopic (denoted $\gamma \simeq \gamma'$) if there exists a continuous function $H : [0,1] \times [0,1] \to X$ such

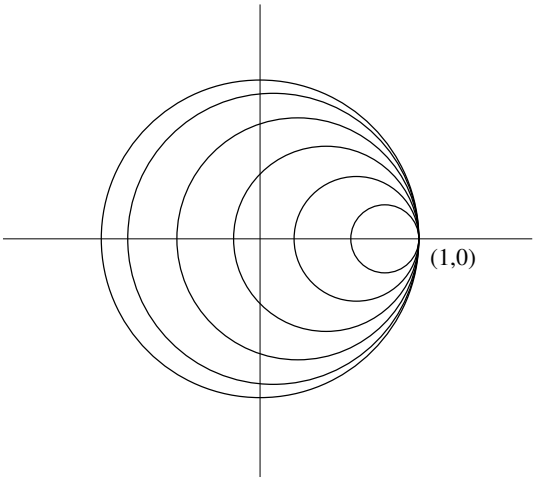


Figure 3.3: Homotopy of based loops.

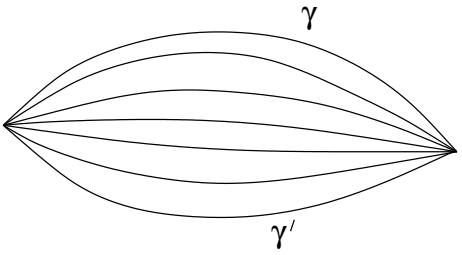


Figure 3.4: Homotopy of paths.

that

$$H(0,s) = \gamma(s), \quad H(1,s) = \gamma'(s)$$

$$H(t,0) = x_0, \quad H(t,1) = x_1$$

Two loops γ, γ' based at x_0 are said to be homotopic if they are homotopic as paths from x_0 to x_0.

Problem 3.2.1: Verify that homotopy is an equivalence relation.

3.3 Operations on paths and loops

We consider the set

$$\pi_1(X, x_0) = \{\text{Loops based at } x_0\}/\text{homotopy}$$

An element in $\pi_1(X, x_0)$ is written as $[\gamma]$ for loops γ in X based at x_0. Given two loops γ, σ we may form a new loop : γ followed by σ, denoted as $\gamma * \sigma$. (See Fig. 3.5.)

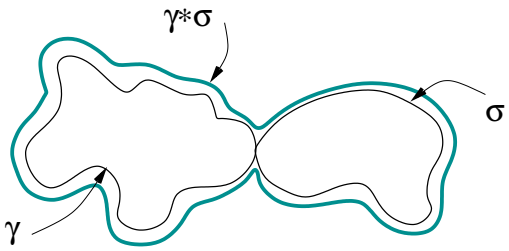


Figure 3.5: Multiplication of loops.

We use the following formula

$$\gamma * \sigma(t) = \begin{cases} \gamma(2t) & \text{if } \ 0 \leq t \leq 1/2 \\ \sigma(2t-1) & \text{if } \ 1/2 \leq t \leq 1 \end{cases}$$

Remark 3.3.1. *Note that this operation is well defined as long as the starting point of σ matches the end-point of γ.*

Using this operation, one proves that

1. The operation $*$ induces a binary operation on $\pi_1(X, x_0)$ which we also denote by $*$.
2. Under the operation $*$, $\pi_1(X, x_0)$ has a group structure with the constant loop as the identity element.

For 1), one has to show

$$\gamma \simeq \gamma', \ \sigma \simeq \sigma' \implies \gamma * \sigma \simeq \gamma' * \sigma'$$

For 2), one has to first prove the associative condition :

$$(\gamma_1 * \gamma_2) * \gamma_3 \simeq \gamma_1 * (\gamma_2 * \gamma_3)$$

In order to prove that the class of the constant loop c is the identity element, one needs to verify

$$\gamma * c \simeq \gamma, \ c * \gamma \simeq \gamma$$

for all loops γ based at x_0. Finally for a loop γ based at x_0, one verifies that its inverse is given by the loop that traverses along γ in the opposite direction : $\bar{\gamma}(t) = \gamma(1-t)$.

Let us verify the associative condition. The rest we leave as an exercise for the reader. First observe that $\alpha_{12,3} = (\gamma_1 * \gamma_2) * \gamma_3$ and $\alpha_{1,23} = \gamma_1 * (\gamma_2 * \gamma_3)$

are given by the formulas

$$\alpha_{12,3}(t) = \begin{cases} \gamma_1(4t) & \text{if } 0 \le t \le \frac{1}{4} \\ \gamma_2(4t-1) & \text{if } \frac{1}{4} \le t \le \frac{1}{2} \\ \gamma_3(2t-1) & \text{if } \frac{1}{2} \le t \le 1 \end{cases}$$

$$\alpha_{1,23}(t) = \begin{cases} \gamma_1(2t) & \text{if } 0 \le t \le \frac{1}{2} \\ \gamma_2(4t-2) & \text{if } \frac{1}{2} \le t \le \frac{3}{4} \\ \gamma_3(4t-3) & \text{if } \frac{3}{4} \le t \le 1 \end{cases}$$

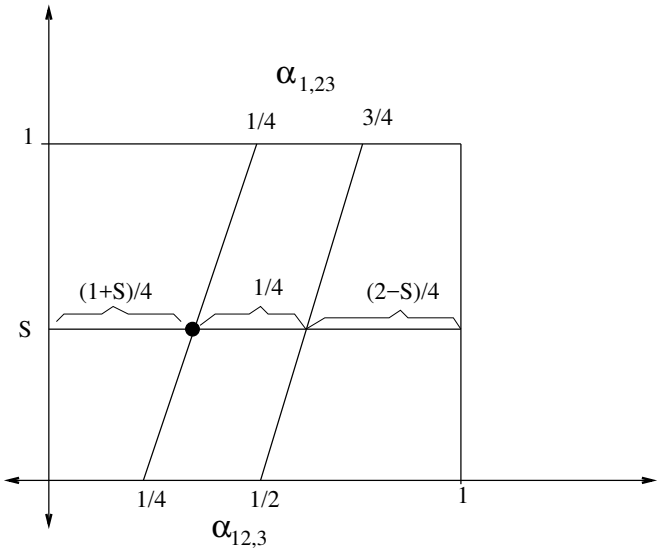


Figure 3.6: Associativity of loop multiplication.

Note that $\alpha_{12,3}$ and $\alpha_{1,23}$ are not the same loops although their images are the same. The difference lies in their speeds. The loop $\alpha_{12,3}$ traverses along γ_1 for time $1/4$, then γ_2 for time $1/4$ and finally along γ_3 for time $1/2$. The loop $\alpha_{1,23}$ traverses along γ_1 for time $1/2$, then γ_2 for time $1/4$ and γ_3 for time $1/4$. We may construct an explicit homotopy to be read off from Fig. 3.6. Now draw a horizontal line at time s and read off the time spent by the loop α_s along γ_1, γ_2 and γ_3. The loop α_s traverses through γ_1 for time $\frac{1+s}{4}$, then γ_2 for time $1/4$ and then γ_3 for time $\frac{2-s}{4}$. This leads to the following formula

$$\alpha_s(t) = \begin{cases} \gamma_1(\frac{4t}{1+s}) & \text{if } 0 \le t \le \frac{1+s}{4} \\ \gamma_2(4t-1-s) & \text{if } \frac{1+s}{4} \le t \le \frac{2+s}{4} \\ \gamma_3(\frac{4t-2-s}{2-s}) & \text{if } \frac{2+s}{4} \le t \le 1 \end{cases}$$

which is easily verified to be a homotopy between $\alpha_{12,3}$ and $\alpha_{1,23}$.

Problem 3.3.1: Using ideas as above write down formulas verifying the other required properties above.

Once all the above properties are verified, $\pi_1(X, x_0)$ has the structure of a group. This is called the *fundamental group* of the topological space X at the point x_0.

3.4 The fundamental group

We start with an example

Example 3.4.1. *Suppose X is a convex subset of $\mathbb{R}^n$. That is, for any two points $x_1, x_2 \in X$, the line segment $\{tx_1 + (1-t)x_2 | t \in [0,1]\}$ is a subset of X. Now for any loop γ in X*

$$\gamma_t(s) = (1-t)\gamma(s) + tx_0$$

is a homotopy $\gamma \simeq c$. Hence every loop at x_0 lies in the class of the identity loop and as a consequence, the fundamental group is the trivial group 0. Therefore, for a convex subset X of $\mathbb{R}^n$, $\pi_1(X, x_0) = 0$.

Example 3.4.2. *Note that the above example can be fitted into an abstract framework. For writing down the argument, we only used that there is a natural way of moving every point to x_0. This is achieved by a continuous deformation of the identity map to the constant map. A space X is said to be contractible if there is a continuous map $H : X \times [0,1] \to X$ such that $H(x,0) = x$ and $H(x,0) = x_0$. Under the additional assumption $H(x_0, t) = x_0$ for all t, the formula*

$$\alpha_s(t) = H(\gamma(t), s)$$

gives a homotopy between γ and the constant loop. Therefore we obtain $\pi_1(X, x_0) = 0$. (In fact, the statement is true without the additional assumption but the proof requires more work.)

Next we list some properties of the fundamental group :

Proposition 3.4.3. *If there is a path from x_0 to x_1, $\pi_1(X, x_0) \cong \pi_1(X, x_1)$.*

Proof. A path γ from x_0 to x_1 induces a group homomorphism $c(\gamma) : \pi_1(X, x_0) \to \pi_1(X, x_1)$ which sends a loop α based at x_0 to the loop $\bar{\gamma} * \alpha * \gamma$ ($\bar{\gamma}(t) = \gamma(1-t)$ is the path from x_1 to x_0 obtained by moving backwards along γ). This is easily verified to be well-defined and that the homomorphism depends only on the homotopy class of the path γ. It follows that the inverse of $c(\gamma)$ is given by $c(\bar{\gamma})$ through the following chain of implications

$$\begin{aligned} c(\gamma) \circ c(\bar{\gamma}) &= c(\gamma * \bar{\gamma}) \\ &= c(\text{const}) \quad \text{as } \gamma * \bar{\gamma} \simeq \text{const} \\ &= id. \end{aligned}$$

and similarly, $c(\bar{\gamma}) \circ c(\gamma) = id$. □

Remark 3.4.4. *It follows from Proposition 3.4.3 that for a path connected space X, the fundamental group $\pi_1(X, x_0)$ does not depend on the point x_0 upto isomorphism. Following this fact, when X is path connected we write $\pi_1(X)$ for the fundamental group and omit the point x_0 from the notation.*

Now define

$$\pi(X; x_0, x_1) = \{\text{set of paths from } x_0 \text{ to } x_1\}/\text{homotopy}.$$

A similar proof to Proposition 3.4.3 may be used to verify that the set $\pi_1(X; x_0, x_1)$ is in bijection with the fundamental group $\pi_1(X, x_0)$ whenever x_0 can be joined to x_1 by a path.

Let $f : X \to Y$ be a continuous function such that $f(x_0) = y_0$. If γ is a loop in X based at x_0, the composition $f \circ \gamma$ is a loop in Y based at y_0. This operation induces a group homomorphism $f_* : \pi_1(X, x_0) \to \pi_1(Y, y_0)$.

Problem 3.4.1: Verify this fact.

One may formulate a condition about when $f_* = g_*$ for two different continuous maps between X and Y. This involves deforming the map f continuously to the map g a definition that will be made explicit in the section on higher homotopy groups. The idea is that in such a case the continuous deformation may be used to write down a homotopy between $f \circ \gamma$ and $g \circ \gamma$. We deal with an explicit case below.

Definition 3.4.5. A subspace $A \subset X$ is said to be a deformation retract, if there exists $H : X \times [0, 1] \to X$ (called deformation retraction) satisfying

$$H(x, 0) = x,\ H(x, 1) \in A,\ H(a, t) = a\ \forall\ t \in [0, 1].$$

Example 3.4.6. *The space $\mathcal{D}^2 - 0$ deformation retracts to the unit circle S^1. In fact $\mathcal{D}^n - 0$ deformation retracts to S^{n-1}. To see this, think of $\mathcal{D}^n - 0$ to be made of an elastic material with a rigid boundary. Now linearly pressurize the material onto its boundary. Such a process gives a continuous deformation. In terms of formulas, we may write :*

$$H(t, x) = (1 - t)x + t\frac{x}{|x|}$$

to deform $\mathcal{D}^n - 0$ to the unit sphere S^{n-1}. Therefore, $\pi_1(\mathcal{D}^n - 0) \cong \pi_1(S^{n-1})$.

Proposition 3.4.7. *Let $A \subset X$ be a deformation retract. Then for a point x_0 in A, $\pi_1(X, x_0) \cong \pi_1(A, x_0)$. [The proof involves deforming a loop in X continuously down to a loop in A using the deformation of X to A]*

Proof. Let H be the deformation retraction. Then, H describes a homotopy between the identity map of X and the function $h_1 : X \to A$ which sends $x \mapsto H(x, 1)$. Let $i : A \to X$ denote the inclusion of the subspace. Then we have

$$\pi_1(A, x_0) \underset{h_{1*}}{\overset{i_*}{\rightleftarrows}} \pi_1(X, x_0)$$

which satisfies

$$h_{1*} \circ i_* = (h_1 \circ i)_* = id_* = id$$
$$i_* \circ h_{1*} = (i \circ h_1)_* = id$$

as $i \circ h_1 \simeq id$ via the homotopy H. □

A special class of topological spaces are the ones whose fundamental group is the trivial group.

Definition 3.4.8. A space X is called *simply connected* if X is path connected and for every point x_0 of X, $\pi_1(X, x_0) = 0$.

We may deduce from the discussion above that if a space is simply connected any two paths with the same endpoints are homotopic. From Example 3.4.1 we have seen that a convex set is simply connected. In fact, a convex set deformation retracts to a point. From Proposition 3.4.7, we know that any space which deformation retracts to a point is simply connected. Observe this is precisely the Example 3.4.2.

So far we have computed the fundamental groups of only those spaces which deformation retracts to a point. Examples of these are $\mathbb{R}^n$ and $\mathcal{D}^n$. Next we consider $X = S^1$ in which case the fundamental group turns out to be non-trivial. This computation passes through a mathematical argument involving " covering spaces". We only outline the main ideas in this case.

Let $p : \mathbb{R} \to S^1$ be the map given by $p(t) = e^{2\pi it}$. Observe that $p(t+1) = p(t)$ and that p is a local homeomorphism, that is, for every t in $\mathbb{R}$ there is an ϵ such that

$$p : (t - \epsilon, t + \epsilon) \to U_t = \text{open neighbourhood of } p(t)$$

is a homeomorphism for some open neighbourhood U_t. The map is described in the Fig. 3.7 below.

Let $\gamma : [0, 1] \to S^1$ be a loop so that $\gamma(0) = \gamma(1) = (1, 0)$. One can prove that the loop γ may be *lifted to a unique loop* $\tilde{\gamma}$ *in* $\mathbb{R}$ *starting at* 0 : that is, there is a unique $\tilde{\gamma} : [0, 1] \to \mathbb{R}$ such that $p \circ \tilde{\gamma} = \gamma$ and $\tilde{\gamma}(0) = 0$. To observe this, first write $\tilde{\gamma}(0) = 0$. Now there exists ϵ such that

$$p : (-\epsilon, \epsilon) \to U_1$$

is a homeomorphism. Since U_1 is open, from the continuity of γ we deduce that there is a δ such that $\gamma : [0, \delta) \to U_1$. Now we readily see that $\tilde{\gamma}$ is uniquely defined on $[0, \delta)$ as $p^{-1} \circ \gamma$. Now iterate the above logic with δ instead of 0 which defines $\tilde{\gamma}$ uniquely on a bigger interval. This can be carried on until we reach 1 and the entire path $\tilde{\gamma}$ is uniquely defined. In this way we deduce

Path-lifting : Any loop γ in S^1 based at the point 1 can be uniquely lifted to a path in $\mathbb{R}$ starting at 0.

We remark that the reader may easily notice that the statement "This can be carried on" needs a rigorous mathematical proof. Such a proof requires the property that $[0, 1]$ is compact, which implies that the process needs to be repeated only a finite number of times to reach 1.

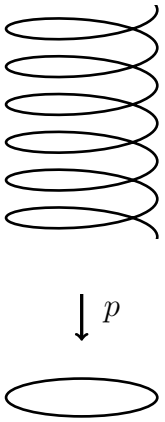


Figure 3.7: Mapping the real line onto the circle.

Example 3.4.9. *Consider the loop γ in S^1 based at 1 given by $\gamma(t) = e^{2\pi it}$. Observe that the unique lift is given by $\tilde{\gamma}(t) = t$.*

The example demonstrates that the unique lift need not be a loop. One makes a similar argument with homotopies to deduce
Homotopy-lifting: If $\gamma \simeq \gamma'$, then $\tilde{\gamma} \simeq \tilde{\gamma}'$ as paths.

Now we prove that $\pi_1(S^1) \cong \mathbb{Z}$. Construct

$$\phi : \pi_1(S^1) \to \mathbb{Z}$$

by defining $\phi([\gamma]) = \tilde{\gamma}(1)$ ($\tilde{\gamma}(1) \in p^{-1}(1) = \mathbb{Z}$). Note that this is well defined by homotopy-lifting:

$$\gamma \simeq \gamma' \implies \tilde{\gamma} \simeq \tilde{\gamma}' \implies \tilde{\gamma}(1) = \tilde{\gamma}'(1)$$

ϕ is easily verified to be a group homomorphism by the following formula of the unique lift $\widetilde{\gamma * \gamma'}$ of $\gamma * \gamma'$.

$$\widetilde{\gamma * \gamma'}(t) = \begin{cases} \tilde{\gamma}(2t) & \text{if } \ 0 \le t \le 1/2 \\ \tilde{\gamma}(1) + \tilde{\gamma}'(2t-1) & \text{if } \ 1/2 \le t \le 1 \end{cases}$$

so that $\widetilde{\gamma * \gamma'}(1) = \tilde{\gamma}(1) + \tilde{\gamma}'(1) = \phi([\gamma]) + \phi([\gamma'])$. We prove that ϕ is an isomorphism as follows

- ϕ is injective : If $\phi([\gamma]) = 0$, then $\tilde{\gamma}$ is a loop in $\mathbb{R}$ based at 0. Since $\mathbb{R}$ is simply connected, $[\tilde{\gamma}] = 0$ which in turn implies $[\gamma] = p_*[\tilde{\gamma}] = 0$.

- ϕ is surjective : Let $\sigma_n : [0,1] \to \mathbb{R}$ be defined as $\sigma_n(t) = nt$ for $n \in \mathbb{Z}$. Then, $p_*\sigma_n$ is a loop based at $1 \in S^1$. The lift is same as σ by the uniqueness so that $\phi([p_*\sigma]) = n$.

Thus we have observed that $\pi_1 S^1 \cong \mathbb{Z}$. A generator is given by a loop $[\gamma]$ with $\phi(\gamma) = 1$ so the loop in Example 3.4.9 is a generator. As a loop this traverses around the circle once in anti-clockwise direction.

Next we turn our attention to the spheres S^n for $n \geq 2$. This relies on the following theorem, a special case of Van-Kampen Theorem.

Theorem 3.4.10. *Suppose U, V are simply connected open subsets of X such that $U \cap V$ is path-connected and $X = U \cup V$. Then X is simply connected.*

Proof. Suppose γ is a loop in X written as $\gamma : [0,1] \to X$. Divide $[0,1]$ as $0 = s_0 < s_1 \cdots < s_n = 1$ so that $\gamma([s_i, s_{i+1}])$ lies entirely in U or V. As an example see Fig. 3.8.

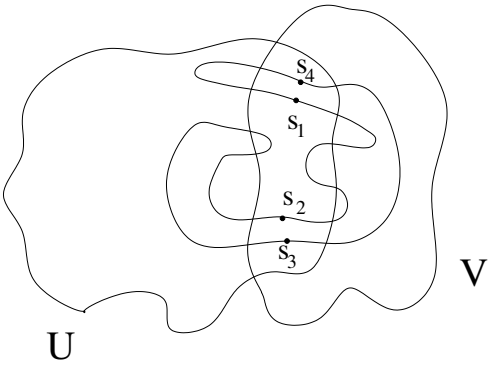


Figure 3.8: Homotopy of a loop in a union.

If $\gamma([s_i, s_{i+1}]$ lies in V join $\gamma(s_i)$ to $\gamma(s_{i+1})$ by a path σ in $U \cap V$. Then from the simply connectedness of V we have that $\gamma|_{[s_i, s_{i+1}]}$ is homotopic to σ. In this way we may homotope γ by removing segments mapping to V by those mapping to $U \cap V$. Therefore, γ is homotopic to a loop in U which is homotopic to the constant loop as U is simply connected. □

Now we may apply the above theorem to S^n by writing the sphere as a union of neighbourhoods of the upper and lower hemispheres. These are homeomorphic to $\mathcal{D}^n$ which is simply connected. The intersection deforms down to the equator S^{n-1} which is path connected if $n \geq 2$. It follows that S^n is simply connected if $n \geq 2$.

We end the discussion pointing out two further computations of the fundamental group.

Example 3.4.11. *We may compute the fundamental group of $\mathbb{R}P^n$ from the fundamental group of S^n. For $n = 1$, the space $\mathbb{R}P^1$ is homeomorphic to S^1 and thus the fundamental group is $\mathbb{Z}$. For $n \geq 2$ we consider the map $q : S^n \to \mathbb{R}P^n$. Analogous to the case of $p : \mathbb{R} \to S^1$, we may observe that every point has two inverse images and that the map q is a local homeomorphism. Results such as path lifting and homotopy lifting follow.*

We may follow the proof of the fundamental group of S^1 for $\mathbb{R}P^n$. By fixing a point $[x] \in \mathbb{R}P^n$ for $x \in S^n$, we see that for a homotopy class $[\gamma] \in \pi_1(\mathbb{R}P^n, [x])$ the element $\tilde{\gamma}(1)$ is well defined and belongs to the set $\{x, -x\}$. We may use the simply connectedness of S^n for $n \geq 2$ analogous to the circle case, to prove that the association above is a bijection. Therefore the fundamental group of $\mathbb{R}P^n$ is a group of order 2 for $n \geq 2$. Hence $\pi_1(\mathbb{R}P^n) \cong \mathbb{Z}_2$.

Example 3.4.12. *Let X be the torus. (See Fig. 3.9.)*

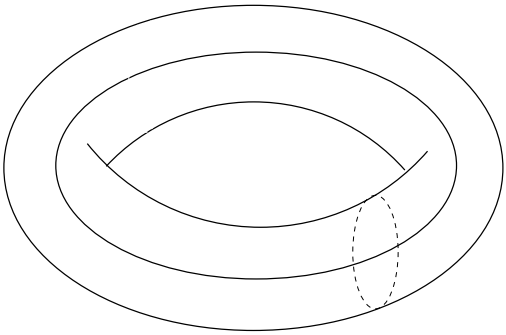


Figure 3.9: Generators of the Fundamental group of a Torus.

Note that the torus is homeomorphic to $S^1 \times S^1$. Thus $\pi_1(X) \cong \pi_1(S^1 \times S^1)$. A loop in the product space $S^1 \times S^1$ is a pair of loops in S^1 and the homotopies also work out pairwise. Therefore $\pi_1(S^1 \times S^1) \cong \mathbb{Z} \oplus \mathbb{Z}$. Note that the loops in figure give generators of either factor.

II Higher homotopy Groups

3.5 Definition of homotopy groups

In the section on the fundamental group, we have seen how considering loops upto homotopy leads to an invariant of topological spaces with values in groups. Following the ideas of the fundamental group. we try to define

$$\pi_n(X) = \{\text{Maps} : S^n \to X\}/\text{Continuous deformations.}$$

We rigorize this concept below.

Definition 3.5.1.

- A based space is a pair (X, x_0) where X is a topological space and $x_0 \in X$. (We will use the notation X for a based space to mean the pair (X, x_0))
- A map of based spaces $X \to Y$ is a continuous function $f : X \to Y$ such that $f(x_0) = y_0$.
- Homotopy of maps : Suppose f, g are two based maps from $X \to Y$. A homotopy from f to g is a continuous function $H : X \times [0,1] \to Y$ such that $H(x,0) = f(x)$ and $H(x,1) = g(x)$ and each $H_t(x) = H(x,t)$ is a based map. For this relation we use the notation $f \simeq g$.

Problem 3.5.1: Verify that homotopy is an equivalence relation.

Notation : We write $[X, Y]$ to denote the homotopy classes of maps between the based spaces X and Y.

Recall that the n-sphere S^n is the space of all unit vectors in $\mathbb{R}^{n+1}$. We fix a base point $*$ in S^n throughout so that S^n is a based space.

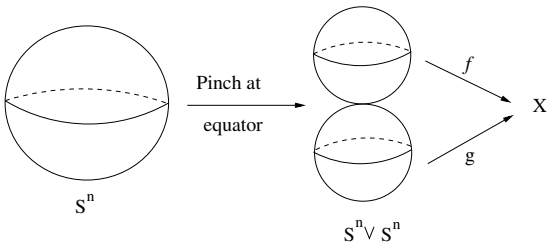


Figure 3.10: Multiplying maps from spheres.

Definition 3.5.2. Let X be a based space. Define $\pi_n(X) = [S^n, X]$.

Next we introduce a binary product on $\pi_n(X)$. Given two maps $f, g : S^n \to X$ we may form the product $f \cdot g : S^n \to X$ as demonstrated in Fig. 3.10.

We may rewrite this in an equivalent fashion. Let I^n denote the n-cube. Recall that $I^n/\partial I^n \cong S^n$. Therefore a based map $S^n \to X$ is equivalent to a continuous function $f : I^n \to X$ such that ∂I^n maps to x_0. We represent such an object by Fig. 3.11.
In this notation the above product can be written schematically as

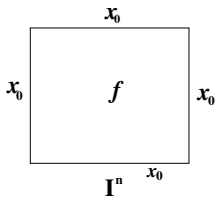


Figure 3.11

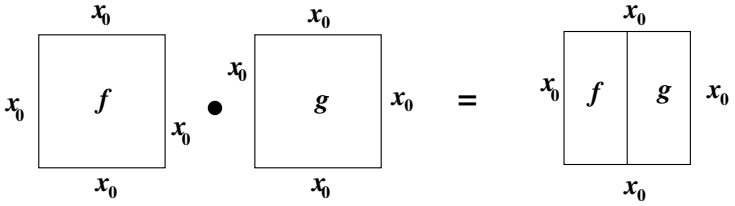


Figure 3.12: Multiplying based maps from cubes.

Problem 3.5.2: Verify that the above operation (Fig. 3.12) induces a group structure on $\pi_n(X)$ and for $n = 1$ this is the fundamental group as defined in the previous section. [HINT : The last equivalent formulation can be written in terms of explicit formulas similar to the fundamental group.]

Next we note that $\pi_n(X)$ is abelian if $n \geq 2$. The following sequence of diagrams (Fig. 3.13) demonstrates a proof of this statement

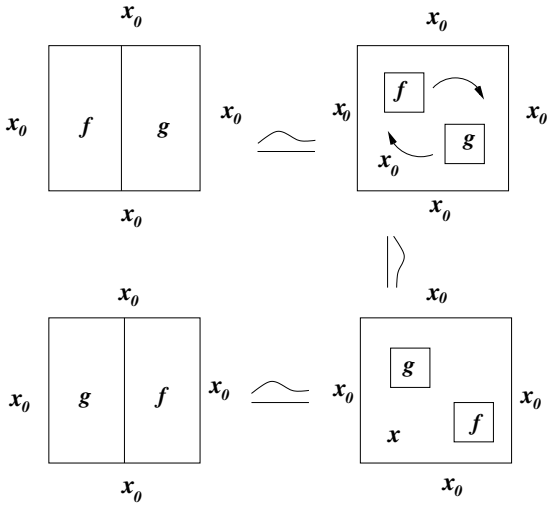


Figure 3.13: Commutativity of multiplication of higher homotopy groups.

Note that the above manoeuvre is not possible if $n = 1$.

3.6 Computing homotopy groups

Computations of homotopy groups are difficult in general and in most cases these are still unknown. We list some important cases below.
1) For contractible spaces X (like convex subsets of $\mathbb{R}^n$ for example), $\pi_n(X) = 0$.
2) For " covering spaces" $p : E \to X$ (like $p : \mathbb{R} \to S^1$ or $q : S^n \to \mathbb{R}P^n$), $\pi_n(E) \cong \pi_n(X)$ for $n \geq 2$.

The properties 1) and 2) together imply that $\pi_n(S^1) \cong \pi_n(\mathbb{R}) = 0$ for $n \geq 2$. Applying 2) for $\mathbb{R}P^n$ we have $\pi_k(\mathbb{R}P^n) \cong \pi_k(S^n)$ if $k \geq 2$.

Next we look at the homotopy groups of spheres $\pi_k(S^n)$ for various k and n. Suppose $f : S^k \to S^n$ is a continuous map. Then, we have
1) Every continuous map is homotopic to a smooth map. (that is, all the partial derivatives are continuous)
2) A smooth map from a space of low dimension cannot map surjectively onto a space of high dimension.[1]

From the two properties above we deduce that a map $f : S^k \to S^n$ for $k < n$ can be homotoped so that it misses some point of S^n. Now using stereographic projection we know that the space $S^n - pt$ is homeomorphic to $\mathbb{R}^n$ which is contractible. Therefore f factors through some contractible space and is homotopic to a constant map. We conclude that $\pi_k(S^n) = 0$ if $k < n$.

Next consider the case $k = n$. The property 1) above still holds. The analogue of property 2) states that for most points (outside a measure zero set to be precise) the inverse image is a finite set. This enables us to construct an invariant. Let $f : S^n \to S^n$ and y be such that $f^{-1}(y) = \{x_1, \cdots, x_r\}$. The restriction of f to sufficiently small neighbourhoods of x_i is a homeomorphism onto its image. Define the sign of x_i to be ± 1 in accordance with whether f is orientation preserving at x_i or not. Define

$$d_y(f) = \sum_i sgn(x_i)$$

If we change the point y to a different point y' we may compare $d_y(f)$ and $d_{y'}(f)$ as follows. Choose a path from y to y' and consider its inverse image under f. It is a subset of S^n consisting of a union of intervals and circles. As an example see Fig. 3.14 below.

From this we may read off the equality $d_y(f) = d_{y'}(f)$. Thus, the number $d_y(f)$ does not depend on the choice of y. This number is called the degree of the map f which is an invariant of the homotopy class. The Hopf degree theorem states that the map $\pi_n(S^n) \to \mathbb{Z}$ given by $f \mapsto deg(f)$ is an isomorphism. We summarize the computations in the following theorem

Theorem 3.6.1.
- *For $k < n$, $\pi_k(S^n) = 0$.*
- *$\pi_n(S^1) = 0$ for $n \geq 2$.*
- *$\pi_n(S^n) \cong \mathbb{Z}$.*

[1] This is an application of Sard's Theorem in multivariable calculus

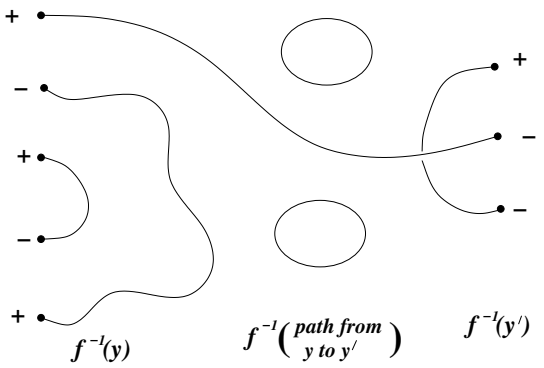


Figure 3.14: Inverse image of a regular path.

For $k > n$ the computation of $\pi_k(S^n)$ is extremely hard. The complete list is unknown for even $n = 2$. Calculation of all homotopy groups for a fixed sphere S^n is a very difficult problem when $n \geq 2$. In comparison, it is easier to compute $\pi_{n+k}(S^n)$ for $k \ll n$ which turn out to be independent of n. We list some of the values below.

k	0	1	2	3	4	5	6	7	8	9
$\pi_{n+k}(S^n)$	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{24}$	0	0	$\mathbb{Z}_2$	$\mathbb{Z}_{240}$	$\mathbb{Z}_2 \oplus \mathbb{Z}_2$	$\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$

We end by stating some more computational results. There is one very important fact which arises while computing homotopy groups. For spaces such as spheres, manifolds and in general objects which are finite dimensional, homotopy groups are hard to compute. On the other hand there are certain infinite dimensional objects[2] for which the computations are slightly easier and more results are known. One such example is the space $\mathbb{C}P^\infty = \cup_n \mathbb{C}P^n$. We have the following result

$$\pi_k(\mathbb{C}P^\infty) = \begin{cases} \mathbb{Z} & \text{if } k = 2 \\ 0 & \text{if } k \neq 2 \end{cases}$$

and similarly for $\mathbb{R}P^\infty = \cup_n \mathbb{R}P^n$

$$\pi_k(\mathbb{R}P^\infty) = \begin{cases} \mathbb{Z}_2 & \text{if } k = 1 \\ 0 & \text{if } k \neq 1 \end{cases}$$

There are computations for $U = \cup_n U(n)$ and $O = \cup_n O(n)$

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

[2]More precisely, infinite loop spaces.

$$\pi_k(O) = \begin{cases} \mathbb{Z}_2 & k \equiv 0 \ (mod\ 8) \\ \mathbb{Z}_2 & k \equiv 1 \ (mod\ 8) \\ 0 & k \equiv 2 \ (mod\ 8) \\ \mathbb{Z} & k \equiv 3 \ (mod\ 8) \\ 0 & k \equiv 4 \ (mod\ 8) \\ 0 & k \equiv 5 \ (mod\ 8) \\ 0 & k \equiv 6 \ (mod\ 8) \\ \mathbb{Z} & k \equiv 7 \ (mod\ 8) \end{cases}$$

The last two computations are part of the famous Bott periodicity theorem. Using these computations $\pi_k(U(n))$ and $\pi_k(O(n))$ are computed in a small range of k.

III A Tutorial on Fundamental groups and group actions

Let X be a topological space and $\pi_1(X)$ denote its fundamental group. We assume that we know $\pi_1(\mathbb{R}^n) = 1$ and $\pi_1(S^1) = \mathbb{Z}$.

3.7 Fundamental groups of Spheres

$S^n = \{(x_1, x_2, ..., x_{n+1}) \in \mathbb{R}^n : \sum_{i=1}^{n+1} x_i^2 = 1\}$ denotes the n-dimensional sphere. We prove that S^n is simply connected, i.e., $\pi_1(S^n) = 1$ for $n > 1$. Let γ be a loop in S^n at a base point a, i.e., $\gamma : [0,1] \in S^n$ is a continuous function such that $\gamma(0) = \gamma(1) = a$. Suppose γ is not surjective. Choose a point p in S^n from outside of the image of γ. Since $S^n - \{p\}$ is diffeomorphic to $\mathbb{R}^n$, choose a diffeomorphism $\phi : \mathbb{R}^n \to S^n - \{p\}$. Let $\phi^{-1}(a) = b$. $\phi^{-1} \circ \gamma$ be a loop in $\mathbb{R}^n$ at a base point b. Since $\mathbb{R}^n$ is simply connected, $\phi^{-1} \circ \gamma$ is homotopic to the constant loop at b. If H is a homotopy between them, $\phi \circ H$ is a homotopy between γ and the constant loop at a. Therefore, any loop which is not surjective is homotopic to the identity in S^n.

Let $\gamma[0,1] = S^n$. Choose a point $p \neq a$ in S^n. Consider a sufficiently small closed disc D centered at p such that D does not contain a. Using compactness of $[0,1]$ one can prove that γ enters and exits D passing through p finitely many times. Let $\gamma_i : [a_i, b_i] \to S^n$ be the components of γ which enter and exits D at a_i and b_i and pass through p at c_i, i.e., $\gamma_i(a_i), \gamma_i(b_i)$ are points in the boundary of D and $\gamma_i(c_i) = p$. $[a_i, b_i] \subset [a, b]$. Since D is simply connected, one can choose a curve $\sigma_i : [a_i, b_i] \to D$ with same end points as γ_i such that σ_i is homotopic to γ_i and σ_i does not pass through p. Then one can replace γ_i's by σ_i's in γ and construct a curve σ which is homotopic to γ but does not pass through p. Hence, σ is not surjective. Therefore, it is homotopic to the identity.

3.8 Fundamental group of Real Projective spaces

The n dimensional real projective space is denoted by $\mathbb{R}P^n$. It can be viewed as a quotient space of S^n. Consider the equivalence relation $\sim$ on S^n defined by $x \sim (-x)$. $\mathbb{R}P^n$ is the set of all equivalence classes with the quotient topology on it. Let $P : S^n \to \mathbb{R}P^n$ be the quotient map defined by $P(x) = [x]$. One can prove that P is a local homeomorphism. Let γ be a loop at a fixed base point $[a] \in \mathbb{R}P^n$. Then there exists a unique curve $\tilde{\gamma}$ in S^n starting at a such that $P \circ \tilde{\gamma} = \gamma$. Since γ is a loop, the end point of $\tilde{\gamma}$ is either a or $-a$. So, there are two types of loops in $\mathbb{R}P^n$ at $[a]$.

Type 1: A Type 1 loop is a loop γ such that $\tilde{\gamma}$ is a loop in S^n at a . Since S^n is simply connected, there is a homotopy H between $\tilde{\gamma}$ and the constant loop at a. Then $P \circ H$ is a homotopy between γ and identity in $\mathbb{R}R^n$. So, Type 1 loops are homotopic to the identity.

Type 2: A Type 2 loop is a loop γ such that the end point of $\tilde{\gamma}$ is $-a$. Consider any two Type 2 loops γ_1 and γ_2. $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ are loops with same end points a and $-a$. Since S^n is simply connected, they are homotopic. If we compose that homotopy with P then we get a homotopy between γ_1 and γ_2 in $\mathbb{R}P^n$.

Therefore, the fundamental group of $\mathbb{R}P^n$ consists of two elements. Hence, it is $\mathbb{Z}_2$.

3.9 Group actions

Let G be a group and X be a topological map. A *group action* is a function $F : G \times X \to X$ such that

1. $F_g : X \to X$ defined by $F_g(x) = F(g, x)$ is a homeomorphism of X for all $g \in G$.
2. $F_g \circ F_h = F_{g \circ h}$ for all $g, h \in G$.
3. $F_e = Id$ where e is the identity of G and Id is identity function of X.

A group G is said to act freely if $F_g(x) = F_h(x)$ for any $x \in X$ implies $g = h$. If a group G acts on X hen one can define an equivalence relation $\sim$ on X by $x \sim y$ if and only if $y = F(g, x)$ for some $g \in G$. The quotient space X/G is the set of all equivalence classes with quotient topology on it. As in the case of $\mathbb{R}P^n$ we have the following result.

Theorem 3.9.1. *If a group G acts on a simply connected topological space X freely and "nicely" then $\pi_1(X/G) = G$.*

For a detailed proof we refer to Ref. 1.

Example 1: $T^n = \mathbb{R}^n/\mathbb{Z}^n$. Let $\tilde{m} = (m_1, m_2, ..., m_n) \in \mathbb{Z}^n$ and $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$. The group action is defined by,

$$F(m, x) = (m_1 + x_1, m_2 + x_2, ..., m_n + x_n).$$

T^n is homeomorphic to $S^1 \times S^1 \times S^1 ... n$-times. For $n = 1$, one gets S^1. The fundamental group of T^n is $\mathbb{Z}^n$.

Example 2: $\mathbb{R}P^n = S^n/\mathbb{Z}_2$. $\mathbb{Z}_2 = \{1, -1\}$. In this case F_1 is the identity function of S^n and F_{-1} is the antipodal map which sends x to $-x$.

Example 3 : S^{2n-1} is embedded in $\mathbb{C}^n$.

$$S^{2n-1} = \{z = (z_1, z_2, ..., z_n) \in \mathbb{C}^n : |z_1|^2 + |z_2|^2 + ... + |z_n|^2 = 1\}$$

Let α be a p-th root of unity. $\mathbb{Z}_p$ is isomorphic to the group $(1, \alpha, \alpha^2, ..., \alpha^{p-1})$.

$$F(\alpha^k, z) = (\alpha^k z_1, \alpha^k z_2, ..., \alpha^k z_n)$$

Therefore, $\pi_1(S^{2n+1}/\mathbb{Z}_p) = \mathbb{Z}_p$.

Example 4 : Lens spaces: Consider S^3 embedded in $\mathbb{C}^2$ as in the previous example. Let α be a p-th root of unity for a prime p, and $1 \leq k \leq p-1$. Then,

$$F(\alpha, z) = (\alpha z_1, \alpha^k z_2)$$

This gives a free action of $\mathbb{Z}_p$ on S^3. The quotient space is called a Lens space and denoted by $L_{p,k}$.

References

[1] Y. Felix, S. Halperin and J. C. Thomas, *Rational homotopy theory*, Graduate Texts in Mathematics, 205. (Springer, New York, 2001).

[2] P. Griffiths and J. Morgan, *Rational homotopy theory and differential forms*. Second edition. Progress in Mathematics, 16. (Springer, New York, 2013).

[3] A. Hatcher, *Algebraic topology*. (Cambridge University Press, Cambridge, 2002).

[4] J. R. Munkres, *Topology: a first course* (Prentice-Hall, Inc., Englewood Cliffs, N.J., 1975).

[5] M. Nakahara, *Geometry, topology and physics*, (Taylor & Francis, Boca Raton, FL, USA, 2003).

[6] V. P. Mineev, *Topologically Stable Defects and Solitons in Ordered Media*, Classic Reviews in Physics, (Book 1), (CRC Press, 1998)

[7] N. D. Mermin, "The topological theory of defects in ordered media", Rev. Mod. Phys. **51**, 591 (1979)

4

Homology

Dheeraj Kulkarni

In this chapter, we discuss simplicial homology theory. We explain, with examples, the main motivation behind various homology theories. Simplicial Homology involves the ideas of simplicial complexes and triangulations for topological spaces, which are introduced at appropriate places. We discuss concrete examples of triangulations for surfaces. Further the chain complex, the boundary operator and the simplicial homology groups are defined. We compute simplicial homology for several examples explicitly. At the end of the chapter, we see the usefulness of homology groups in distinguishing topological spaces.

4.1 Introduction

After formally defining notions of topological spaces, continuity of maps between topological spaces and homeomorphisms, the following natural question arises :

Question 4.1.1. *Given two topological spaces X and Y, are they homeomorphic?*

Recall that a homeomorphism is a continuous map with a continuous inverse. Homeomorphisms preserve the topology (i.e., collection of open sets). In more mundane language, homeomorphisms ignore stretching, bending, shear of the space. For example, the graph of the *sine* function is homeomorphic to the real line.

Let us pose this question for familiar examples of topological spaces – the Euclidean spaces. Is $\mathbb{R}^1$ homeomorphic to $\mathbb{R}^2$? A little thought reveals that they cannot be. The reason is as follows. Suppose they are homeomorphic. Let us remove a point from the real line $\mathbb{R}$ and the image of this point from $\mathbb{R}^2$.

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_4

The resulting spaces must still be homeomorphic. However, the real line with a point removed is disconnected whereas the plane minus a point is connected. Thus, we get a contradiction. Further, we observe that the argument is just the same if we want to show that $\mathbb{R}$ and $\mathbb{R}^k$ can not be homeomorphic for any $k \geq 2$.

Now the next question is : Is $\mathbb{R}^2$ homeomorphic to $\mathbb{R}^3$? We see that we can no longer use the above argument successfully to arrive at a contradiction. We need a new idea to deal with this problem. Homology is a tool developed to resolve this classification problem completely and partial resolutions to other classification questions.

In homology theory, the main idea is to associate abelian groups to topological spaces in such a way that these groups encode information about topological properties of the space. Further, to every map (map means a continuous function) between two topological spaces we associate group homomorphisms between associated groups. Overall, we translate the problems in the world of topological spaces to the world of abelian groups. We understand abelian groups better than abstract topological spaces so we expect to solve topological problems transformed into group theoretic problems. More often than not, this approach helps us to answer negatively the questions raised above.

In these lectures, we will focus on Simplicial Homology theory which is easier to formulate than other homology theories. Besides that, the simplicial homology theory has strong geometric intuition behind it.

4.2 Motivating Examples

Before we begin defining the simplicial homology, let us look at a few simple and instructive examples. We will deal with symmetric objects in the Euclidean spaces namely the convex polyhedrons. We will focus on boundaries of these objects in Fig. 4.1.

Finding "Holes" in the space.

First observe from Fig. 4.1 that the boundary of a boundary of a convex polyhedron is empty. This is a very important geometric observation on which homology theory rests.

Next there may be objects without boundaries. We call objects without boundary as "cycles" (the term is justified if we look at the examples in Fig. 4.1). The figure suggests that there could be cycles in topological spaces that are not boundaries. In other words, existence of cycles which are not boundaries suggests that there are "holes" in the space. This is the second important observation for defining homology theory.

4.3 Simplicial Complex

Let us define what do we mean by a symmetric polyhedron in m-dimensional Euclidean space.

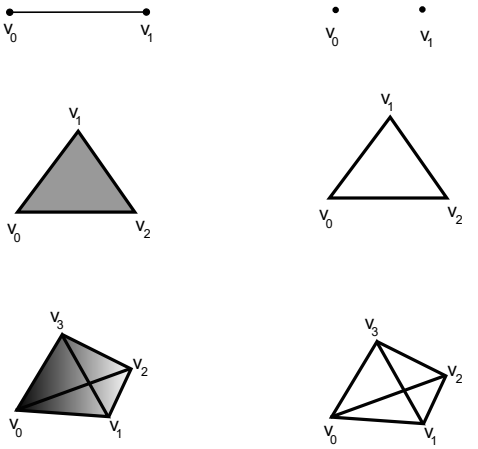


Figure 4.1: Simplices and their boundaries

Definition 4.3.1. *Let $\{v_0, v_1, v_2, \ldots, v_n\}$ be $n+1$ points in $\mathbb{R}^m$ such that they are not contained in a hyperplane of dimension less than n. Then an n-*simplex *is the smallest convex set containing $\{v_0, v_1, v_2, \ldots, v_n\}$. We denote it by $[v_0, v_1, \ldots, v_n]$. Thus, we have*

$$[v_0, v_1, \ldots, v_n] = \left\{ \sum_{i=0}^{n} t_i v_i \,|\, 0 \leq t_i \leq 1 \text{ and } \sum_{i=0}^{n} t_i = 1 \right\}.$$

Thus, every point in n-simplex receives coordinates given by $(t_0, t_1, \ldots, t_n)$. They are called **barycentric coordinates** .

Remark 4.3.1. The points $\{v_0, v_1, \ldots, v_n\}$ do not lie in a hyperplane of dimension less than n is equivalent to saying that the set of vectors $\{v_1 - v_0, v_2 - v_0, \ldots, v_n - v_0\}$ is a linearly independent set in $\mathbb{R}^m$.

Remark 4.3.2. By an n-simplex, we really mean an ordered set of points $[v_0, v_1, v_2, \ldots, v_n]$. This naturally induces an order on the subsets of points by writing them in increasing order of subscripts. For example $[v_0, v_1], [v_1, v_2]$ and $[v_0, v_2]$ are sub-simplices of a 3-simplex $[v_0, v_1, v_2]$. One more consequence of ordering on the vertices is that there is a canonical linear homeomorphism between any two n-simplices preserving the order of points.

Definition 4.3.2. *If we delete one of the $n+1$ points from n-simplex $[v_0, v_1, \ldots, v_n]$ while keeping the increasing order of subscripts of points, then the remaining n vertices form an $n-1$ simplex which is called a* **face** *of $[v_0, v_1, \ldots, v_n]$.*

Now we are ready to define a *finite* simplicial complex.

Definition 4.3.3. *Let K be a finite collection of simplices. We call K a* **simplicial complex** *if the following conditions hold*

1. *If σ is a simplex in K then every face of σ is also in K.*
2. *If σ and τ are simplices in K then either $\sigma \cap \tau = \emptyset$ or $\sigma \cap \tau$ is a common face of both σ and τ.*

Fig. 4.2 gives an illustration of a simplicial complex and an example which is not a simplicial complex.

We can define infinite simplicial complexes by modifying the definition suitably but we will restrict our attention to finite simplicial complexes.

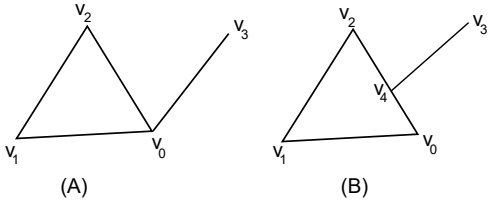


Figure 4.2: Part(A) shows allowed intersection in Simplicial Complex. Part(B) is not a simplicial complex

Notice that a simplicial complex is *not* a topological space as defined above. It is only a collection of simplices that satisfy some properties. We introduce the following notion of the geometric carrier of a simplicial complex which *is* a topological space.

Definition 4.3.4. *Let K be a (finite) simplicial complex. We define a topology on the set*

$$|K| := \bigcup_{\sigma \in K} \sigma$$

A subset U of $|K|$ is open if $U \cap \sigma$ is open in σ for every $\sigma \in K$. Recall that each σ, a simplex from Euclidean space, is a topological space in its own right. Thus, the above definition makes sense. We denote the collection of open sets by $\mathcal{T}$. It is easy to see that $\mathcal{T}$ defines a topology on $|K|$.

We call $|K|$ with topology $\mathcal{T}$ the **geometric carrier** *of the simplicial complex K.*

Definition 4.3.5. *Let X be a topological space. We say that X is triangulated if there is a simplicial complex K and a homeomorphism $h : |K| \to X$. Further, we say that the pair (K, h) is a triangulation of X. In this case, we say that X is* triangulable.

In more mundane words, triangulation of a space is a decomposition of the space into nicely fitting triangles (more generally polyhedrons). A triangulation of space X gives a good model to understand the topology of the space. However, showing existence of triangulations for spaces can be very challenging. We will be dealing with nice topological spaces, namely, manifolds. There are deep theorems that guarantee the existence of triangulations for *smooth* manifolds. We will assume that we are already given at least one triangulation and work with it.

4.4 Review of Abelian Groups

Recall that a group $(G, +)$ is abelian if any two elements commute, i.e., $a+b = b+a$ for all $a, b \in G$. We say that G is **finitely generated** if there is a finite generating set for G. More explicitly, there is a set $\{a_1, \ldots, a_k\}$ such that for any element g in G we have $g = n_1a_1 + n_2a_2 + \cdots + n_ka_k$ for some integers n_i.

We say that a set $\{g_1, g_2, \ldots, g_r\}$ is linearly independent if $n_1g_1 + n_2g_2 + \cdots + n_rg_r = 0$ implies that each $n_i = 0$.

Definition 4.4.1. *We say that an abelian group G is a* **free abelian group** *of* **rank** *r if there is a linearly independent generating set with r elements.*

Example 4.4.2. *$\mathbb{Z}$ is a free abelian group of rank 1 but $\mathbb{Z}_n$ for $n \neq 0$ is not a free abelian group. In general if G is a free abelian group of rank r then it is isomorphic to*

$$\underbrace{\mathbb{Z} \oplus \mathbb{Z} \oplus \cdots \oplus \mathbb{Z}}_{r \text{ copies}}.$$

Now let us recall examples of finite abelian groups. First we have the cyclic group C_n of order n which is isomorphic to $\mathbb{Z}/n\mathbb{Z}$. We can further take direct sums of cyclic abelian groups, for example, $C_n \oplus C_m$. The following theorem asserts that essentially these are the only finite abelian groups.

Theorem 4.4.1 (Structure of Finite Abelian Groups)**.** *Let G be a finite abelian group. Then G is isomorphic to*

$$\frac{\mathbb{Z}}{n_1\mathbb{Z}} \oplus \frac{\mathbb{Z}}{n_2\mathbb{Z}} \oplus \cdots \oplus \frac{\mathbb{Z}}{n_r\mathbb{Z}}.$$

Now, we are ready to state the fundamental theorem of abelian groups.

Theorem 4.4.2. *Let G be a finitely generated abelian group. Then G is isomorphic to*

$$\underbrace{\mathbb{Z} \oplus \mathbb{Z} \oplus \cdots \oplus \mathbb{Z}}_{r \text{ copies}} \oplus \frac{\mathbb{Z}}{n_1\mathbb{Z}} \oplus \frac{\mathbb{Z}}{n_2\mathbb{Z}} \oplus \cdots \oplus \frac{\mathbb{Z}}{n_r\mathbb{Z}}.$$

Moreover r is the same for any decomposition of G given as above.

Remark 4.4.3. In the view of the above theorem, we can define rank for any *finitely generated* abelian group as the number of copies of $\mathbb{Z}$ in any decomposition of G given as above.

4.5 Chain Groups, Boundary Maps and Homology Groups

Let K be a simiplicial complex that triangulates the topological X. Now consider the set consisting of (finite) formal sums

$$C_k = \left\{ \sum_{\sigma^k \in K} n_i \sigma_i^k \mid n_i \in \mathbb{Z} \right\}.$$

Clearly, C_k is a free abelian group generated by all k-simplices in K. We call C_k as k-th **chain group** and elements in C_k are called **k-chains**.

Now we define the boundary map which is a homomorphism $\partial_k : C_k \to C_{k-1}$. We define it on the basis elements of C_k end extend it linearly as follows. Let $[v_0, v_1, \ldots, v_k]$ be a k-simplex. By $[v_0, v_1, \ldots, \hat{v}_i, \ldots, v_k]$ we denote the $(k-1)$-simplex obtained by deleting i-th vertex. Now define the boundary map as

$$\partial_k([v_0, v_1, \ldots, v_k]) = \sum_{i=0}^{k} (-1)^i [v_0, v_1, \ldots, \hat{v}_i, \ldots, v_k].$$

It does not take much effort to show that $\partial_k \circ \partial_{k+1} = 0$. (The essential point in the proof is that every face occurs exactly twice but with the opposite signs).

Now, observe that $\mathrm{Im}\partial_{k+1} \subset \ker\partial_k$ as $\partial_k \circ \partial_{k+1} = 0$.

We call the elements of $\ker(\partial_r)$ as *r-cycles* and denote the subgroup r-cycles by Z_r. The elements in $\mathrm{Im}(\partial_{r+1})$ are called as *r-boundaries*. We denote the subgroup of r-boundaries by B_r. The terminology has its origin in geometry as we see that 1-cycles are closed loops and 1- boundaries are boundaries of some 2-dimensional region. Thus, r-boundaries are always r-cycles. However, there could be r-cycles that are not r-boundaries. This motivates the following definition of homology groups.

Definition 4.5.1 (Homology Groups)**.** *We define the homology groups $H_r(K)$, associated to simplicial complex K, to be the quotient of subgroups of r-cycles by r-boundaries. Formally, we define*

$$H_r(K) := \frac{Z_r(K)}{B_r(K)}.$$

The elements of $H_r(K)$ are called homology classes. In the quotient group $H_r(K)$, two r-cycles c_1 and c_2 represent the same homology class if and only if $c_1 - c_2 = \partial_{r+1} b$, where b is a $(r+1)$-chain. In other words, two cycles represent the same homology class if and only if they differ by some r-boundary. In that case, we say that r-cycles c_1 and c_2 are *homologous*.

Now the following fact states that the homology groups $H_r(K)$ are independent of choice of the simplicial complex K that triangulates X. In other words, the homology groups depend only on the underlying topological space X. The reader is encouraged to look at Chapter 1 and 2 of [3].

Theorem 4.5.1. *Let X be a topological space. Suppose X is triangulable and K is a simplicial complex that triangulates X. Then the homology groups $H_r(K)$ are independent of the choice of K. In other words, the homology groups $H_r(K)$ depend only on the underlying space X.*

In the light of the above theorem, we refer to groups $H_r(K)$ as the homology groups associated to X and denote them by $H_r(X)$.

The following theorem says that that to a continuous function $f : X \to Y$ there is a natural way of associating a group homomorphism $f_{*r} : H_r(X) \to$

$H_r(Y)$ for each r such that composition of maps is associated to composition of homomorphisms. The identity map gets associated to the identity homomorphism. In Category theoretic language, this is summarized by saying that homology is a *functor* from the category of topological spaces to the category of abelian groups.

Theorem 4.5.2. *Let X, Y and Z be triangulable spaces. Let $f : X \to Y$ and $g : Y \to Z$ be continuous functions. Then there are group homomorphisms f_{*r}, g_{*r} such that $(g \circ f)_{*r} = g_{*r} \circ f_{*r}$. Moreover, for the identity map $id_X : X \to X$ the associated group homomorphisms $id_{X_{*r}}$ are identity group homomorphisms.*

Frequently, the above mentioned theorem helps us prove that spaces can not be homeomorphic by showing that the corresponding homology groups can not be isomorphic.

4.6 Computation of Homology Groups Of Surfaces

In this section, we see some computations of homology groups associated to surfaces. These examples illustrate the power of homology theory. The reader is encouraged to do more computations as it will help develop the geometric intuition and insight into ideas of homology theory.

4.6.1 The Cylinder

We consider the following triangulation given in Fig. 4.3 for the compact cylinder $S^1 \times [0, 1]$. Let us denote it by K.

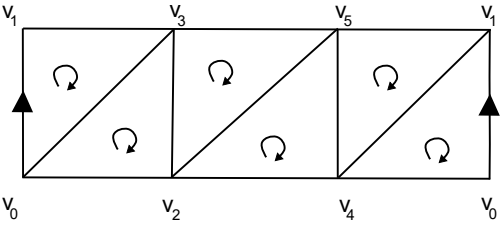


Figure 4.3: A triangulation K of the cylinder.

Now we observe that the 0-th chain group $C_0(K)$ is the abelian group generated by the vertices $v_0, v_1, v_2, v_3, v_4, v_5$. Thus, we formally write

$$C_0(K) = \quad [v_0]\mathbb{Z} \oplus [v_1]\mathbb{Z} \oplus [v_2]\mathbb{Z} \oplus [v_3]\mathbb{Z} \oplus [v_4]\mathbb{Z} \oplus [v_5]\mathbb{Z}.$$

In other words $C_0(K) \cong \mathbb{Z}^5$. Similarly, we note that the chain group $C_1(K)$ is generated by all the edges shown in the figure. Thus, we have

$$\begin{aligned} C_1(K) = \quad & [v_0, v_1]\mathbb{Z} \oplus [v_0, v_2]\mathbb{Z} \oplus [v_0, v_3]\mathbb{Z} \oplus [v_0, v_4]\mathbb{Z} \\ & \oplus [v_1, v_3]\mathbb{Z} \oplus [v_1, v_5]\mathbb{Z} \oplus ... \oplus [v_4, v_5]\mathbb{Z}. \end{aligned}$$

The group of 2-chains is generated by six triangles as shown in the above diagram.

$$\begin{aligned} C_2(K) = \quad & [v_0, v_1, v_3]\mathbb{Z} \oplus [v_0, v_2, v_3]\mathbb{Z} \oplus [v_0, v_4, v_1]\mathbb{Z} \\ & \oplus [v_2, v_3, v_5]\mathbb{Z} \oplus [v_2, v_5, v_4]\mathbb{Z} \oplus [v_1, v_4, v_5]\mathbb{Z}. \end{aligned}$$

Let us now compute $H_0(K)$ which by definition $C_0(K)$ modulo $B_0(K)$. By definition $B_1(K)$ is the subgroup of vertices that are boundaries of 1-chains. We observe the following few examples

$$\begin{aligned} \partial_1([v_0, v_1]) &= [v_1] - [v_0], \\ \partial_1([v_3, v_5]) &= [v_5] - [v_3], \\ \partial_1([v_4, v_2]) &= [v_2] - [v_4]. \end{aligned}$$

To summarize, the boundary of an oriented edge is the terminal vertex minus the initial vertex of the oriented edge. Further, observe that if there is an oriented path given by a sequence of oriented edges then the boundary of this path is also the terminal vertex of the last edge in the sequence minus the initial vertex of the first edge in the sequence as every other vertex appears twice but with opposite sign.

Now, we see that the vertex v_0 can be joined by an oriented path to any other vertex in K. Thus, $H_0(K) = [v_0]\mathbb{Z}$. Later, we will see that H_0 is a measure of connectedness of the underlying space. Let us now turn our attention to $H_1(K)$. By definition, H_1 is the group of 1-cycles modulo 1-boundaries. Now, 1-cycles are elements in $C_1(K)$ that get mapped to zero under the boundary map. Thus, formally we set the following equation

$$\partial_1(m_1[v_0, v_1] + m_2[v_0, v_2] + m_3[v_0, v_3] + m_4[v_0, v_4] + \cdots + m_{12}[v_5, v_4]) = 0, \quad (4.1)$$

where $m_1, m_2, \ldots, m_{12}$ are integers. There are twelve edges in K (see Fig. 4.3). Reading off the coefficients of each vertex and setting it equal to zero, we get a system of five linear equations. For examples, the coefficient of the vertex v_0 gives the following equation

$$-m_1 - m_2 - m_3 - m_4 = 0.$$

Solutions of the above system of equations gives all possible 1-cycles. There is a more geometric way to find out 1-cycles. In Fig. 4.4, we observe that the oriented path $[v_0, v_2] + [v_2, v_4] + [v_4, v_0]$ gives a closed loop (dotted lines). It is easy to see that

$$\partial_1([v_0, v_2] + [v_2, v_4] + [v_4, v_0]) = 0.$$

There are many more loops formed by sequence of edges, for instance $[v_0, v_3] + [v_3, v_2] + [v_2, v_5] + [v_5, v_4] + [v_4, v_1] + [v_1, v_0]$ which gives a "zig-zag" loop as shown in Fig. 4.4.

Now, we need to see which 1-cycles differ by 1-boundaries. We observe that the 1-cycles given by $[v_0, v_2] + [v_2, v_4] + [v_4, v_0]$ and $[v_0, v_3] + [v_3, v_2] +$

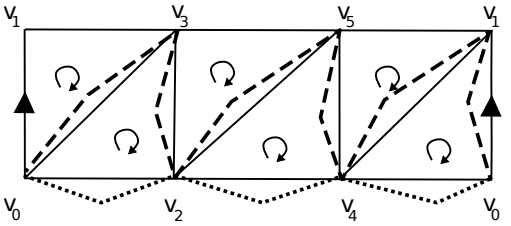


Figure 4.4: Two possible closed paths for the triangulation of Fig. 4.3: one along the dotted lines, and one along the zig-zag dashed lines.

$[v_2, v_4] + [v_4, v_0]$ (check that these are indeed 1-cycles!) differ by a 1-boundary, viz., $\partial_2[v_0, v_3, v_2]$. We write

$$\begin{aligned}([v_0, v_3] + [v_3, v_2] + [v_2, v_4] + [v_4, v_0]) - ([v_0, v_2] + [v_2, v_4] + [v_4, v_0])\\ = [v_0, v_3] + [v_3, v_2] - [v_0, v_2] = \partial_2([v_0, v_3, v_2]).\end{aligned}$$

Further observation suggests that any two 1-cycles will differ by 1-boundaries. To see this, one needs two find out triangles in K that form the difference of given two 1-cycles. Thus, there is only one 1-cycle upto 1-boundaries. The reader is encouraged to verify the claim that $H_1(K) \cong \mathbb{Z}$.

Now, to compute $H_2(K)$ we need to see only 2-cycles as there are no non-trivial 2-boundaries (There are no non-trivial 3-chains). We notice that there are six triangles that generate $C_2(K)$. We set the equation for finding out 2-cycles as before

$$\begin{aligned}\partial_2\big(n_1[v_0, v_1, v_3] + n_2[v_0, v_3, v_2] + n_3[v_2, v_3, v_5]\\ +n_4[v_2, v_5, v_4] + n_5[v_4, v_5, v_1] + n_6[v_0, v_4, v_1]\big) = 0.\end{aligned}$$

Again reading off the coefficients of edges and equating them to zero, we get twelve equations. The solution space gives all possible 2-cycles. However, there is a more geometric way to find out 2-cycles. First observe that if there is an edge shared by *exactly* two triangles then the edge gets opposite orientations from the two triangles then it is cancelled by the boundary map if and only if both the triangles appear with the same coefficient. For example, the edge $[v_0, v_1]$ is shared by *exactly* two (oriented) triangles namely, $[v_0, v_1, v_3]$ and $[v_1, v_0, v_4]$. Thus, we have $n_4 = n_6$ from the above equation.

Observe that each of the edges $[v_1, v_3], [v_3, v_5], [v_5, v_1], [v_0, v_2], [v_2, v_4], [v_4, v_0]$ in K belongs to only one triangle. Each of the six triangles in K contains precisely one edge from this list. Thus, in the boundary computation, the coefficient of each triangle must be equal to zero in order to get a 2-cycle. Thus, the element 0 is the only 2-cycle. Hence $H_2(K) = 0$.

4.6.2 Torus

The torus T can be obtained as the quotient space of a rectangle with opposite sides identified in a manner that preserves orientations on them. Thus, we can triangulate the torus as shown in the Fig. 4.5.

Instead of going through the tedious computations, we argue geometrically to compute the homology groups of torus T.

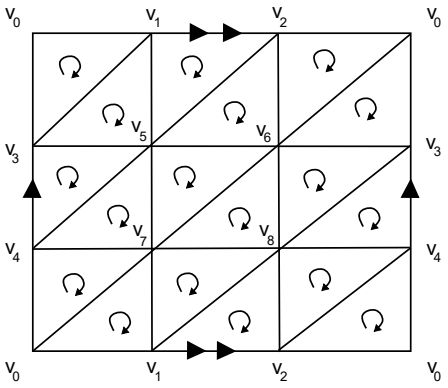


Figure 4.5: A triangulation of the torus T.

We first observe that all the homology groups for $H_r(T) = 0$ for $r \geq 3$ as there are no non-trivial r-chains for $r \geq 3$. Next, recall that $H_0(T) = C_0/B_0$. Again any two vertices can be joined by an oriented path. Therefore, there is only one 0-cycle upto 0-boundaries. Hence, $H_0(T) \cong \mathbb{Z}$.

There are eighteen triangles in the above diagram. To get a 2-cycle, we observe that each edge is shared by *precisely* two adjacent triangles. Thus, the coefficients of these triangles must be the same for cancellations to take place. Any two triangles can be joined by a sequence of adjacent triangles. Therefore, the coefficients of all triangles must be the same. This shows that the subgroup of 2-cycles is generated by the cycle that is obtained by taking the sum of all (oriented) triangles as shown in the diagram. Thus, $H_2(T) \cong \mathbb{Z}$.

To compute $H_1(T)$, first we look for closed loops formed by sequences of oriented edges. Evidently, $[v_0, v_1]+[v_1, v_2]+[v_2, v_0]$ and $[v_0, v_3]+[v_3, v_4]+[v_4, v_0]$ are 1-cycles. We denote them by α and β respectively. We claim that these two 1-cycles are linearly independent and do not differ by a 1-boundary. A geometric way of arriving at this statement is to observe that the two 1-cycles do not form the boundary (as in the usual sense) of any combination of triangles. The reader can verify this fact and may support it with a rigorous argument.

There are many more (in fact, infinitely many!) 1-cycles that can be seen in Fig. 4.5. For example, $[v_0, v_7] + [v_7, v_6] + [v_6, v_0]$ is another 1-cycle. We notice that any 1-cycle containing the vertex v_0 differs from a linear combination of α and β by a 1-boundary. More explicitly, if a 1-cycle effectively (considering orientations) covers m horizontal and n vertical loops then it differs by a 1-

boundary with the 1-cycle $m\alpha + n\beta$. For example the 1-cycle $[v_0, v_7] + [v_7, v_6] + [v_6, v_0]$ covers the horizontal and vertical loops exactly once. Hence, it differs by a 1-boundary with $\alpha + \beta$. The 1-boundary in this case is the image, under ∂_2, of sum of all triangles in upper half portion above the 1-cycle $[v_0, v_7] + [v_7, v_6] + [v_6, v_0]$.

From the discussion above it follows that $H_1(T) = \alpha\mathbb{Z} \oplus \beta\mathbb{Z}$.

4.6.3 The Projective Plane

The projective plane is obtained by taking a closed unit disk in the plane and identifying diametrically opposite points on the boundary circle. The resulting topological space can *not* be given an orientation. We consider the triangulation shown in Fig. 4.6 for the projective plane.

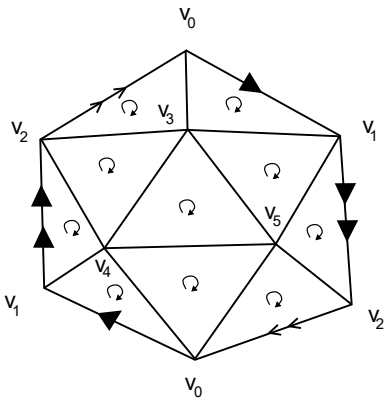


Figure 4.6: A triangulation of $\mathbb{R}P^2$

By now, an alert reader might have acquired a geometric feel for guessing the homology groups and supporting the guesses with rigorous arguments.

First, observe that $H_r(\mathbb{R}P^2) = 0$ for $r \geq 3$ as there are no non-trivial chains in dimension greater than or equal to three. Now $H_0(\mathbb{R}P^2) \cong \mathbb{Z}$ as all the vertices can be connected by a path of edges. Next, we note that $H_2(\mathbb{R}P^2) = 0$ as any edge is shared by *precisely* two adjacent triangles. Further, any two triangles can be joined by a sequence of adjacent triangles. Therefore, the coefficients must be the same for cancellations to happen along non-peripheral edges. The coefficients along the peripheral edges $[v_0, v_1], [v_1, v_2]$ and $[v_2, v_0]$ get added instead of cancelling. This is due to the identification along the boundary. Therefore, one may argue that the only 2-cycle is the trivial 2-cycle (the element 0). Hence $H_2(\mathbb{R}P^2) = 0$.

Now for $H_1(\mathbb{R}P^2)$, first observe that $[v_0, v_1] + [v_1, v_2] + [v_2, v_0]$ is a 1-cycle. We denote it by α. Notice that $2([v_0, v_1] + [v_1, v_2] + [v_2, v_0])$, the geometric boundary of the disk is the boundary (image under the boundary map ∂_2) of sum of all triangles in the figure. Thus, $2\alpha = 0$ in H_1. Any loop formed by a

sequence of non-peripheral edges (or 1-cycle) differs by a 1-boundary with the l-cycle 2α. Thus, it must be the 0 element in $H_1(\mathbb{R}P^2)$. The only non-trivial element in H_1 is α with the property that $2\alpha = 0$. Hence, $H_1(\mathbb{R}P^2) = \mathbb{Z}/2\mathbb{Z}$.

4.7 Some Remarks and Conclusion

As the reader may have noticed some common features and some differences in the above computations. We state the following facts without proof.

Theorem 4.7.1. *Let X be connected triangulable topological space. Then $H_0(X) \cong \mathbb{Z}$. More generally, if X is not connected then $H_0(X)$ has rank equal to the number of connected components of X.*

Theorem 4.7.2. *If topological spaces X and Y are homeomorphic then $H_r(X) \cong H_r(Y)$ for all r.*

As a corollary, we note that if the homology groups of the two given spaces are not isomorphic then the spaces can not be homeomorphic. This is useful in the following instance.

Theorem 4.7.3. *Let S^n denote the n-dimensional unit sphere centered at the origin in Euclidean space $\mathbb{R}^{n+1}$. Then $H_i(S^n) = \mathbb{Z}$ for $i = 0, n$, and $H_i(S^n) = 0$ otherwise.*

Proof. Here we give a sketch of a proof. We leave the details to be filled by the reader. First observe that S^n has a triangulation given by the boundary of the n-dimensional regular polyhedron. Using this triangulation, one concludes that $H_i(S^n) = \mathbb{Z}$ when $i = 0, n$. If $0 < i < n$, observe that every i-chain is an i-boundary. One can "see" that this is an n-dimensional regular polyhedron. □

Corollary 4.7.4. *$\mathbb{R}^n$ is homeomorphic to $\mathbb{R}^m$ if and only if $m = n$.*

Proof. When $m = n$ the statement is obviously true. To prove the converse assume that $\mathbb{R}^n$ and $\mathbb{R}^m$ are homeomorphic via a homeomorphism h. Then it follows that $\mathbb{R}^n - \{0\}$ is homeomorphic to $\mathbb{R}^m - \{0\}$ (to get a homeomorphism bring the point $h(0)$ to the origin 0 in $\mathbb{R}^m$ by an appropriate translation). We now use that fact that S^{k-1} is a *deformation retract*[1] of $\mathbb{R}^k - \{0\}$ for $k \geq 1$. Therefore, $H_i(S^{k-1}) \cong H_i(\mathbb{R}^k)$. The reader is encouraged to look at the notion of deformation retract and its relation to homology groups from Refs. [2], [3] and [1]. Now, it follows from Theorem 4.7.3 that $m = n$. □

The above corollary completely answers Question 4.1.1 when X and Y are Euclidean spaces.

In conclusion, homology theory is a useful tool to distinguish topological spaces. Moreover, the power of homology theory lies in the fact that homology groups can be computed in many situations.

[1]We may recall (see Example 3.4.6) that a map $h(t, x) = (1 - t)x + tx/|x|, t \in [0, 1]$ takes any point $x \in \mathbb{R}^k - \{0\}$ to a point on the unit sphere S^{k-1}. Here $|x|$ is the length of vector x.

The author would like to point out that these notes are merely an introduction to (simplicial) homology theory. Interested reader may refer to Refs. [2] and [1].

References

[1] M. Greenberg and J. Harper, *Algebraic Topology : A First Course(Mathematics Lecture Notes Series)*, (Westview Pr (Short Disc); Revised edition, 1982),

[2] A. Hatcher, *Algebraic Topology*, (Cambridge University Press, Cambridge, 2002).

[3] J. R. Munkres, *Elements of algebraic topology*, (Addison-Wesley Publishing Company, Menlo Park, CA, 1984).

5

Differential Topology and Differential Geometry

Kingshook Biswas and Soma Maity

In the first part of this chapter, we give a brief introduction to Smooth Manifolds and Differential Forms following mainly the text of Arnold ("Mathematical Methods of Classical Mechanics").
In the second part, we start with the definitions of Riemannian metrics, connections and curvatures on open sets of Euclidean spaces, and then give a brief introduction of those on smooth manifolds.

I A Brief Introduction to Manifolds and Differential Forms

5.1 Introduction: Smooth Manifolds, Tangent Spaces, Derivatives

A manifold is a topological space formed by patching together pieces of $\mathbb{R}^n$ by differentiable maps, to which the usual concepts from multivariable calculus, such as vector fields, line and surface integrals generalize. The formal definition is:

Definition 5.1.1. A C^∞ (or smooth) n-manifold is a second countable, Hausdorff topological space M together with a collection of pairs $\{(U_i, \phi_i)_{i\in I}\}$ called *charts*, where the sets $\{U_i\}_{i\in I}$ form an open cover of M, each ϕ_i is a homeomorphism from U_i onto an open set V_i in $\mathbb{R}^n$, and whenever U_i and U_j intersect, the map $\phi_j \circ \phi_i^{-1} : \phi_i(U_i \cap U_j) \subset \mathbb{R}^n \to \phi_j(U_i \cap U_j) \subset \mathbb{R}^n$ is C^∞ or "smooth" (partial derivatives of all orders exist).

The collection of charts is called an *atlas*, while the real-valued functions $x_1, \ldots, x_n$ such that $\phi_i = (x_1, \ldots, x_n)$ are called *local coordinates*.

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_5

A manifold is locally homeomorphic to $\mathbb{R}^n$, and can be thought of as formed by patching together the open subsets $V_i \subset \mathbb{R}^n$ smoothly by the maps $\phi_j \circ \phi_i^{-1}$.

A function on a manifold $f : M \to \mathbb{R}$ is called smooth if near each point it is smooth as a function of the local coordinates near the point (the function $f \circ \phi^{-1}$ is smooth for any chart ϕ on M). The smooth functions on a manifold form a ring, denoted by $C^\infty(M)$, which is useful in defining many concepts from calculus in an intrinsic manner on manifolds without having to resort to local coordinates.

A map between manifolds $f : M \to N$ is said to be smooth if when expressed in local coordinates on M and N it is smooth as a map between $\mathbb{R}^n$ and $\mathbb{R}^m$ (all maps $\psi \circ f \circ \phi^{-1}$ are smooth, where ϕ, ψ are charts on M, N respectively). A smooth map $f : M \to N$ is called a *diffeomorphism* if it has a smooth inverse; we say then that M and N are *diffeomorphic*. The set of diffeomorphisms from a manifold M to itself forms a group denoted by $\mathrm{Diff}(M)$.

5.1.1 Examples of manifolds

1. $\mathbb{R}^n$, with atlas consisting of the single chart $(\mathbb{R}^n, id)$.
2. The unit circle $S^1 \subset \mathbb{R}^2$, with atlas consisting of two charts giving the usual angular coordinate,

$$\theta_1 : S^1 - \{(-1,0)\} \to (-\pi, \pi), \theta_2 : S^1 - \{(1,0)\} \to (0, 2\pi).$$

3. The unit sphere $S^2 \subset \mathbb{R}^3$, with atlas consisting of two charts given by stereographic projections from the North Pole $N = (0,0,1)$ and South Pole $S = (0,0,-1)$ respectively,

$$\phi_1 : S^2 - \{N\} \to \mathbb{R}^2, \phi_2 : S^2 - \{P\} \to \mathbb{R}^2.$$

The same construction works for the n-sphere $S^n \subset \mathbb{R}^{n+1}$.
4. The two-torus $T^2 = S^1 \times S^1 \subset \mathbb{R}^4$, with atlas consisting of 4 charts

$$\begin{aligned} U_1 \times U_1 &\to (-\pi, \pi) \times (-\pi, \pi), \\ U_1 \times U_2 &\to (-\pi, \pi) \times (0, 2\pi), \\ U_2 \times U_2 &\to (0, 2\pi) \times (0, 2\pi), \\ U_2 \times U_1 &\to (0, 2\pi) \times (-\pi, \pi), \end{aligned}$$

giving the two angular coordinates on the torus, where

$$U_1 = S^1 - \{(-1,0)\}, U_2 = S^1 - \{(1,0)\}.$$

Similarly the Cartesian product $M \times N$ of two manifolds M, N is a manifold with charts $\phi \times \psi : U \times V \to \mathbb{R}^m \times \mathbb{R}^n$, where $(U, \phi), (V, \psi)$ are charts on M, N respectively.
Remark. While all the examples above are embedded in Euclidean spaces, and it is a well-known theorem (Whitney Embedding Theorem) that in fact any

smooth n-manifold may be smoothly embedded into $\mathbb{R}^{2n+1}$, in practice manifolds often arise as abstract quotient spaces, as in the following two examples.
5. The two-torus may also be described as a quotient space $\mathbb{C}/L$, where $L = \mathbb{Z}\cdot\omega_1 \oplus \mathbb{Z}\cdot\omega_2$ is a discrete subgroup of $\mathbb{C}$ generated by two $\mathbb{R}$-linearly independent vectors ω_1, ω_2. We can think of this torus as obtained by identifying by translations opposite sides of a parallelogram with vertices $0, \omega_1, \omega_1+\omega_2, \omega_2$. If $q : \mathbb{C} \to \mathbb{C}/L$ is the quotient map, then q is a local homeomorphism and charts on $\mathbb{C}/L$ are given by local inverses $s : U \subset \mathbb{C}/L \to \mathbb{C}$ of q, with changes of charts being given by translations of the form $z \mapsto z + \omega, \omega \in L$.
6. If M is a manifold and Γ is a discrete subgroup of $\mathrm{Diff}(M)$ such that the quotient map $q : M \to M/\Gamma$ is a covering map, then M/Γ is a manifold. Charts on M/Γ are given by composing charts on M with local inverses of q, $U \subset M/\Gamma \to V \subset M \to \mathbb{R}^n$.

5.1.2 Manifolds as configuration spaces of mechanical systems

The configuration spaces of many mechanical systems are often manifolds. We give some examples below, taken from Arnold's book.
Example 1. The configuration space of a planar pendulum is the circle S^1.
Example 2. The configuration space of the "spherical" mathematical pendulum is the two-sphere S^2.
Example 3. The configuration space of a "planar double pendulum" is the two torus T^2.

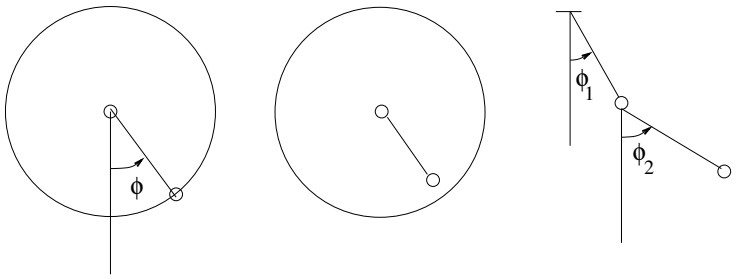


Figure 5.1: Planar, spherical and double planar pendulums

Example 4. The configuration space of a spherical double pendulum is the Cartesian product of two spheres $S^2 \times S^2$.
Example 5. The configuration space of a rigid line segment in the (q_1, q_2)-plane is the manifold $\mathbb{R}^2 \times S^1$ with coordinates q_1, q_2, q_3.
Example 6. A rigid right-angled triangle OAB moves around the vertex O in $\mathbb{R}^3$. The position of the triangle is completely described by an orthogonal right-handed frame $\vec{e_1} = OA/|OA|, \vec{e_2} = OB/|OB|, \vec{e_3} = \vec{e_1} \times \vec{e_2}$, or equivalently by the 3×3 orthogonal matrix $[\vec{e_1}|\vec{e_2}|\vec{e_3}]$ with determinant $+1$. The configuration space of the triangle OAB is the group $SO(3)$ of such matrices, which is a 3-manifold.

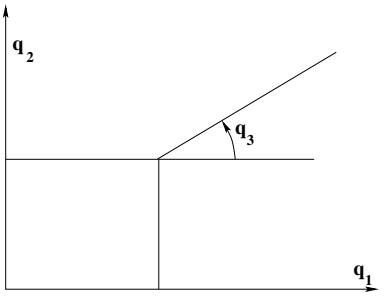


Figure 5.2: Configuration space of a segment in the plane

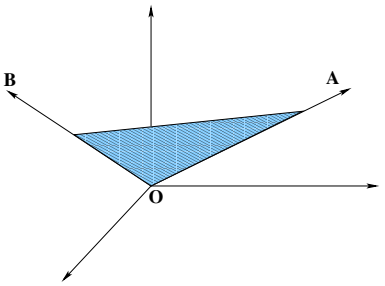


Figure 5.3: Configuration space of a right-angled triangle OAB

5.1.3 Tangent spaces and derivatives

Just as smooth curves in $\mathbb{R}$ or surfaces in $\mathbb{R}^3$ have at each point a tangent line or tangent plane, to each point x of a k-manifold M embedded in $\mathbb{R}^n$ one can associate a k-dimensional subspace T_xM of $\mathbb{R}^n$, called the *tangent space to M at p*. The vector space T_xM can be described as the set of velocities $\dot{\gamma}(0)$ of smooth curves γ in M passing through x at time 0.

For an abstract manifold without a given embedding into Euclidean space, such as those defined as quotient spaces, it is not immediately clear however how to define tangent spaces. While the Whitney Embedding Theorem ensures that any smooth manifold may be embedded into some Euclidean space, it is useful (and aesthetically satisfying) to have a definition of the tangent space T_pM as an abstract vector space defined intrinsically independent of the choice of an embedding into Euclidean space. We give two definitions, one geometric and the other algebraic.

Geometric definition of tangent spaces and differentials

The geometric definition of tangent spaces is based on the intuitive notion of tangent vectors to a point x being velocities of smooth curves passing through x.

We say that two curves γ_1, γ_2 in $\mathbb{R}^n$ passing through the same point x at time $t = 0$ have *first-order contact at* x if $\gamma_1(t) - \gamma_2(t) = o(t)$ as $t \to 0$. Similarly one could define a relation of having *kth-order contact at* x by requiring that $\gamma_1(t) - \gamma_2(t) = o(t^k)$ as $t \to 0$.

Then it is easy to see that the relation of first-order contact is an equivalence relation on the set of curves passing through x at time $t = 0$, which, in the case where the curves are smooth, is equivalent to both curves having the same velocity at time $t = 0$, $\dot{\gamma_1}(0) = \dot{\gamma_2}(0)$ (while k-th order contact is equivalent to the first k derivatives at time $t = 0$ being equal).

For curves γ_1, γ_2 passing through a point x of an n-manifold M at time $t = 0$, we say they have first-order contact at x if this is true when viewed in some chart near p (the curves $\phi \circ \gamma_1, \phi \circ \gamma_2$ have first-order contact at $\phi(x) \in \mathbb{R}^n$, where ϕ is a chart near x). This gives an equivalence relation on curves passing through x which is independent of the choice of chart (because the changes of charts are smooth maps).

Definition 5.1.2 (Geometric definition of tangent space)**.** Let M be a smooth manifold and let $x \in M$. The tangent space to x at M, denoted by T_xM, is the set of equivalence classes $[\gamma]$ of smooth curves γ passing through x under the relation of first-order contact at x.

While this defines the tangent space as a set, to give it the structure of a vector space, we can choose any chart chart ϕ near x, then the map $d\phi_x := ([\gamma] \mapsto \phi \circ \gamma(0))$ gives a bijection between T_xM and the vector space $\mathbb{R}^n$, so we can use the addition and scalar multiplication on $\mathbb{R}^n$ to define the addition and scalar multiplication in T_xM (for example, $[\gamma_1] + [\gamma_2] := [\gamma_3]$ if $d\phi_x([\gamma_1]) + d\phi_x([\gamma_2]) = d\phi_x([\gamma_3])$ in $\mathbb{R}^n$). This vector space structure is independent of the choice of chart ϕ because the changes of charts are smooth; if ψ is another chart near x, then the two bijections $d\phi_x : T_xM \to \mathbb{R}^n$ and $d\psi_x : T_xM \to \mathbb{R}^n$ are related by the linear map $d\eta : \mathbb{R}^n \to \mathbb{R}^n$ where η is the change of charts.

With this vector space structure the map $d\phi_x : T_xM \to \mathbb{R}^n$ becomes a linear isomorphism, so T_xM is a vector space of dimension n.

We can give a geometric definition of the derivative df_x of a smooth map $f : M \to N$ between manifolds at a point x:

Definition 5.1.3 (Geometric definition of derivative)**.** Given a smooth map between manifolds $f : M^m \to N^n$, the derivative of f at a point x of M, denoted by df_x or f_{*x}, is the map

$$\begin{aligned} df_x : T_xM &\to T_yN \\ [\gamma] &\mapsto [f \circ \gamma] \end{aligned}$$

where $y = f(x) \in N$.

If ϕ, ψ are charts near x, y respectively, and $\hat{f}$ is the smooth map $\psi \circ f \circ \phi^{-1}$ from $\mathbb{R}^m$ to $\mathbb{R}^n$ given by the expression of f in these charts, then $d\psi_y \circ df_x \circ (d\phi_x)^{-1}$ is given by the linear map $d\hat{f}_{\phi(x)}$ from $\mathbb{R}^m$ to $\mathbb{R}^n$. From the definition

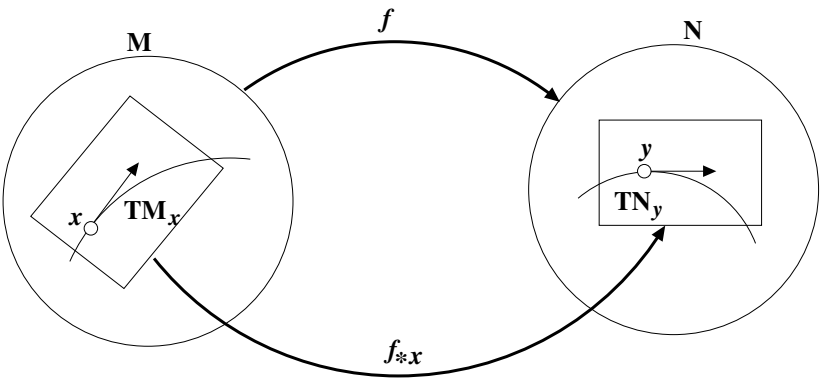


Figure 5.4: Derivative of a smooth map

of the vector space structures on T_xM, T_yN it follows that df_x is a linear map. Moreover it is easy to see that the Chain Rule $d(g \circ f)_x = dg_y \circ df_x$ holds for smooth maps between manifolds.

While the geometric definitions of tangent spaces and derivatives of smooth maps are aesthetically satisfying, for practical purposes the algebraic definition, which we give in the next section, is the one commonly used for computations.

Algebraic definition of tangent spaces

The algebraic definition of tangent spaces is based on the idea of tangent vectors to a point p being directions along which one can compute derivatives of smooth functions near p.

For a point p in $\mathbb{R}^n$, let $C^\infty(p)$ be the set of smooth functions f defined in some neighbourhood of p, modulo the equivalence relation $f \sim g$ if $f = g$ on some neighbourhood of p. This forms a ring called the ring of *germs of smooth functions at p*. Any vector v in $\mathbb{R}^n$ gives a map $\partial_v : C^\infty(p) \to \mathbb{R}$ by taking the directional derivative of a function f in the direction v,

$$(\partial_v f)(p) = \lim_{t \to 0} \frac{f(p + tv) - f(p)}{t}$$

This map is linear and satisfies the Leibniz (or product) rule,

$$(\partial_v f \cdot g)(p) = (\partial_v f)(p)g(p) + f(p)(\partial_v g)(p)$$

Conversely any map $l : C^\infty(p) \to \mathbb{R}$ which is linear and satisfies Leibniz rule is of the form ∂_v for a unique vector v.

For a point p of a manifold M, we can define similarly the ring $C^\infty(p)$ of germs of smooth functions near p.

Definition 5.1.4 (Algebraic definition of tangent space)**.** Let M be a manifold and let $p \in M$. The tangent space to p at M, denoted T_pM, is defined to be the set of linear maps $l : C^\infty(p) \to \mathbb{R}$ which satisfy Leibniz rule, $l(f \cdot g) = l(f)g(p) + f(p)l(g)$.

In this case it is clear that T_pM forms a vector space, and from the remarks of the previous paragraph, $v \mapsto \partial_v$ gives an isomorphism between $\mathbb{R}^n$ and $T_p\mathbb{R}^n$ for any $p \in \mathbb{R}^n$.

Definition 5.1.5 (Algebraic definition of derivative)**.** Let $f : M \to N$ be a smooth map between manifolds and let $p \in M, q = f(p) \in N$. Then f induces a linear map $f^* : C^\infty(q) \to C^\infty(p), u \mapsto u \circ f$. The derivative of f at p is the map

$$\begin{aligned} df_p : T_pM &\to T_qN \\ l &\mapsto l \circ f^* \end{aligned}$$

It is clear in this case that df_p is a linear map. Choosing a chart $\phi = (x_1, \ldots, x_n)$ near p with $\phi(p) = q$, the map $d\phi_p : T_pM \to T_q\mathbb{R}^n \simeq \mathbb{R}^n$ is an isomorphism, so that T_pM is an n-dimensional real vector space, with basis given by $\{\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}\}$, where

$$\begin{aligned} \frac{\partial}{\partial x_i} : C^\infty(p) &\to \mathbb{R} \\ u &\mapsto \frac{\partial}{\partial x_i} u \circ \phi^{-1}(q) \end{aligned}$$

In this case as well the Chain Rule holds, and there is a natural isomorphism between the geometric and algebraic definitions of tangent spaces, $[\gamma] \to l_\gamma$, where $l_\gamma : C^\infty(p) \to \mathbb{R}, u \mapsto \frac{d}{dt}_{|t=0}(u \circ \gamma)$. For the remainder of these notes we shall only be using the algebraic definition of tangent spaces however.

In what follows, the dual space to the tangent space, T_p^*M, called the *cotangent space to M at p*, will play an important role. Note that for any real valued function f on M, its derivative at p $df_p : T_pM \to T_{f(p)}R \simeq \mathbb{R}$ gives an element of the cotangent space T_p^*M. If $x_1, \ldots, x_n$ are local coordinates near p, then the dual basis to the basis $\{\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}\}$ of T_pM is given by the derivatives $\{dx_1, \ldots, dx_n\}$ of the coordinate functions, and the derivative of a smooth function f is given by

$$df = \sum_i \frac{\partial f}{\partial x_i} dx_i$$

5.1.4 Tangent bundle, vector fields and flows

The notion of *vector fields* on manifolds formalizes the idea of force fields acting on configuration spaces of mechanical systems.

Definition 5.1.6 (Tangent bundle)**.** Let M be a smooth manifold. The tangent bundle of M, denoted by TM, is given as a set by the disjoint union of the tangent spaces,

$$TM = \bigcup_{p \in M} T_p M$$

The tangent bundle comes equipped with a map $\pi : TM \to M$ which sends a tangent vector v to the point p at which it is based, so that $\pi^{-1}(p) = T_p M$. The tangent bundle can be given the structure of a smooth vector bundle of rank n as follows:

Given a chart $(U, \phi = (x_1, \ldots, x_n))$ of an atlas for M, we have a bijection

$$T_\phi : \pi^{-1}(U) \to U \times \mathbb{R}^n$$
$$v = \sum_i v_i \left(\frac{\partial}{\partial x_i} \right)_p \mapsto (p = \pi(v), d\phi_p(v) = (v_1, \ldots, v_n))$$

The bijections T_ϕ define bundle charts on TM. If (V, ψ) is another chart for M with $U \cap V = W$, let $\eta = \psi \circ \phi^{-1}$ be the change of charts map. Then by the Chain Rule $d\eta \circ d\phi = d\psi$, so that the bijections T_ϕ and T_ψ are related on $\pi^{-1}(W)$ by

$$T_\psi = (id_W \times d\eta) \circ T_\phi$$

so that the transition function for the bundle TM is given by $p \in W \mapsto d\eta_{\phi(p)}$ which is smooth. Thus TM has the structure of a smooth vector bundle of rank n over M, and in particular is a smooth manifold of dimension $2n$.

Definition 5.1.7 (Vector field)**.** A vector field on a manifold M is a smooth section X of the vector bundle $TM \to M$, in other words a smooth map $X : M \to TM$ such that $X(p) \in T_p M$ for all p (or $\pi \circ X = id_M$).

Concretely, a vector field is given by a smooth assignment to each point p of a tangent vector $X(p)$ based at that point. In terms of local coordinates $x_1, \ldots, x_n$ it can be written locally as

$$X(p) = \sum_i a_i(p) \left(\frac{\partial}{\partial x_i} \right)_p$$

where $a_1, \ldots, a_n$ are smooth functions. A vector field on M determines integral curves and a local flow (1-parameter group of diffeomorphisms) on M:

Definition 5.1.8 (Integral curves)**.** A smooth curve γ in M is an integral curve of a vector field X if

$$\frac{d}{dt}\gamma = X(\gamma(t)) \in T_{\gamma(t)} M$$

for all t.

In terms of local coordinates if γ is given by n functions of t, $\gamma(t) = (x_1(t), \ldots, x_n(t))$, then the above equation is equivalent to a system of first-order ODEs:

$$x_i'(t) = a_i(x_1(t), \ldots, x_n(t)) \quad i = 1, \ldots, n$$

and hence for any p in M, by the Existence-Uniqueness Theorem for ODEs for $\epsilon > 0$ small enough there is a unique integral curve $\gamma : (-\epsilon, \epsilon) \to M$ satisfying the initial condition $\gamma(0) = p$. A vector field is called *complete* if all integral curves are defined for all time. For example any vector field on a compact manifold is complete.

Definition 5.1.9 (Flow of a vector field)**.** The flow of a complete vector field X is the 1-parameter family of maps $\phi_t : M \to M, p \mapsto \gamma(t)$, where γ is the unique integral curve of X satisfying $\gamma(0) = p$.

If $t \mapsto \gamma(t)$ is an integral curve then so is $t \mapsto \gamma(s+t)$ for any fixed s, and it then follows easily from the uniqueness of integral curves that $\phi_t \circ \phi_s = \phi_{s+t}$ for all t, s, thus the maps ϕ_t form a 1-parameter group of diffeomorphisms of M. Elements of the flow are often denoted by $exp(tX)$.

5.2 Differential forms, wedge products and exterior derivative

Differential forms on manifolds provide a way of formalizing the traditional 'infinitesimal' quantities of calculus such as 'line elements' dx, 'surface elements' $dxdy$, 'volume elements' $dxdydz$, etc. On a manifold, 1-forms are to represent infinitesimal quantities which one should be able to integrate over smooth curves in the manifold, 2-forms infinitesimal quantities which one should be able to integrate over smooth surfaces in the manifold, and so on. For example work done by a force field along a curve γ should be given by the integral of a 1-form over γ, flux of a fluid through a surface S should be given by the integral of a 2-form over S.

A k-form ω, supposed to represent a notion of infinitesimal oriented k-dimensional volume, should be a function $(v_1, \ldots, v_k) \mapsto \omega(v_1, \ldots, v_k)$ of 'infinitesimal parallelepipeds' with sides $v_1, \ldots, v_k$ which are tangent vectors based at a point p, which should be linear as a function of each side v_j when the others are fixed (being a notion of volume), and which should change sign if the order of two sides v_i, v_j is interchanged (being a notion of *oriented* volume). This should remind the reader of k-dimensional determinants, which are functions of k vectors satisfying these properties.

5.2.1 Multilinear forms, alternating forms, tensor products, wedge products

Let V be an n-dimensional real vector space, with a basis $\{e_1, \ldots, e_n\}$. Let V^* be the dual space of V with dual basis $\{e_1^*, \ldots, e_n^*\}$ (the example to keep in mind for what follows is $V = T_pM$ and $V^* = T_p^*M$ with bases $\{e_1 = \frac{\partial}{\partial x_1}, \ldots, e_n = \frac{\partial}{\partial x_n}\}$ and $\{e_1^* = dx_1, \ldots, e_n^* = dx_n\}$).

Definition 5.2.1 (Multilinear and alternating k-forms)**.** For $k \geq 0$, a multilinear k-form on V is a real-valued function on the Cartesian product of V with itself k times, $\omega : V \times \cdots \times V \to \mathbb{R}$, such that ω is linear as a function of each variable

$v_i \in V$ when the other $(k-1)$ are fixed (if $k = 0$ then ω is a scalar, and if $k = 1$ then ω is simply a linear functional on V).

An alternating k-form on V is a multilinear k-form ω which changes sign when the order of two vectors v_i, v_j is reversed:

$$\omega(v_1, \ldots, v_i, \ldots, v_j, \ldots, v_k) = -\omega(v_1, \ldots, v_j, \ldots, v_i, \ldots, v_k)$$

The set of multilinear and alternating k-forms on V form vector spaces which we shall denote by $V^{*\otimes k}$ and $\wedge^k V^*$ (these spaces are special cases for the vector space V^* of general constructions for vector spaces, namely k-fold tensor products and wedge products, but we will not be requiring these notions in full generality).

There is a natural bilinear product $(\alpha, \beta) \mapsto \alpha \otimes \beta$ on the direct sum of the spaces $V^{*\otimes k}$ which gives a $(k+l)$-multilinear form $\alpha \otimes \beta$ from multilinear k and l-forms α, β, and turns this direct sum into an $\mathbb{R}$-*algebra*. We recall that a k-algebra A is a vector space over a field k together with a multiplication that makes it a ring, such that the ring multiplication and scalar multiplication are compatible ($\alpha \cdot (v \cdot w) = (\alpha v) \cdot w = v \cdot (\alpha w)$, for all vectors $v, w \in A$ and scalars $\alpha \in k$), such as the algebra of polynomials $k[x]$ or the algebra of $n \times n$ matrices $M_n(k)$.

Definition 5.2.2 (Tensor product and tensor algebra of multilinear forms)**.** The tensor product of a multilinear k-form α and a multilinear l-form β is the multilinear $(k+l)$-form $\alpha \otimes \beta$ defined by:

$$(\alpha \otimes \beta)(v_1, \ldots, v_k, w_1, \ldots, w_l) := \alpha(v_1, \ldots, v_k)\beta(w_1, \ldots, w_l)$$

The tensor algebra of multilinear forms, denoted by $T(V^*)$, is the direct sum of the spaces of multilinear k-forms, $T(V^*) := \oplus_{k=0}^{\infty} V^{*\otimes k}$, together with the multiplication given by the tensor product $\otimes : V^{*\otimes k} \times V^{*\otimes l} \to V^{*\otimes(k+l)}$ extended to $T(V^*)$ so as to be bilinear.

The symmetric group S_k of permutations σ of $\{1, \ldots, k\}$ acts linearly on the space of multilinear k-forms by permuting the arguments of a multilinear k-form ω:

$$(\sigma \cdot \omega)(v_1, \ldots, v_k) := \omega(v_{\sigma(1)}, \ldots, v_{\sigma(k)})$$

Since any permutation is a product of transpositions, the alternating k-forms are precisely those which satisfy $(\sigma \cdot \omega) = sgn(\sigma)\omega$, where $sgn(\sigma)$ denotes the sign of the permutation σ. An alternating k-form may thus also be viewed as a common fixed point for the group of linear maps $T_\sigma : V^{*\otimes k} \to V^{*\otimes k}, \omega \mapsto sgn(\sigma)(\sigma \cdot \omega)$. A standard way to obtain a common fixed point for a finite group of linear maps is to average over an orbit of the group. We thus define a linear map

$$\pi_{alt} : V^{*\otimes k} \to \wedge^k V^*$$
$$\omega \mapsto \frac{1}{k!} \sum_{\sigma \in S_k} T_\sigma \omega = \frac{1}{k!} \sum_{\sigma \in S_k} sgn(\sigma)(\sigma \cdot \omega)$$

The map π_{alt} satisfies $\pi_{alt}^2 = \pi_{alt}, (\pi_{alt})_{|\wedge^k V^*} = id$, and is in fact the projection of $V^{*\otimes k}$ onto the subspace $\wedge^k V^*$ along the complementary subspace of *symmetric k-forms* (the multilinear k-forms which are invariant under the action of S_k). This allows us to define a bilinear product on alternating forms, the *wedge product*:

Definition 5.2.3 (Wedge product and exterior algebra of alternating forms)**.** The wedge product of an alternating k-form α with an alternating l-form is the alternating $(k+l)$-form $\alpha \wedge \beta$ defined by

$$\begin{aligned}(\alpha \wedge \beta)(v_1, \ldots, v_k, v_{k+1}, \ldots, v_{k+l}) \\ &:= (k+l)!\ \ \pi_{alt}(\alpha \otimes \beta)(v_1, \ldots, v_k, v_{k+1}, \ldots, v_{k+l}) \\ &= \sum_{\sigma \in S_{k+l}} sgn(\sigma)\alpha(v_{\sigma(1)}, \ldots, v_{\sigma(k)})\beta(v_{\sigma(k+1)}, \ldots, v_{\sigma(k+l)})\end{aligned}$$

The exterior algebra of alternating forms, denoted by $A(V^*)$, is the direct sum of the spaces of alternating k-forms, $A(V^*) := \oplus_{k=0}^{\infty} \wedge^k V^*$, with the multiplication given by the wedge product $\wedge : \wedge^k V^* \times \wedge^l V^* \to \wedge^{k+l} V^*$ extended to $A(V^*)$ so as to be bilinear.

The reason for the factor $(k+l)!$ is so that the wedge product may be interpreted as a sum of oriented volumes. For example, given k 1-forms $\omega_1, \ldots, \omega_k$, the wedge product $\omega_1 \wedge \cdots \wedge \omega_k$ acting on a k-tuple of vectors $(v_1, \ldots, v_k)$ turns out to be given by the following determinant:

$$(\omega_1 \wedge \cdots \wedge \omega_k)(v_1, \ldots, v_k) = \begin{vmatrix} \omega_1(v_1) & \omega_1(v_2) & \ldots & \omega_1(v_k) \\ \omega_2(v_1) & \omega_2(v_2) & \ldots & \omega_2(v_k) \\ \vdots & \vdots & \ddots & \vdots \\ \omega_k(v_1) & \omega_k(v_2) & \ldots & \omega_k(v_k) \end{vmatrix}$$

which is the oriented volume of the parallelepiped with sides $T(v_1), \ldots, T(v_k)$ in $\mathbb{R}^k$, where $T : V \to \mathbb{R}^k$ is the linear map defined by $T(v) = (\omega_1(v), \ldots, \omega_k(v))$.

For the remainder of these notes we shall refer to alternating k-forms simply as k-forms. Any k-form is determined by its value on k-tuples of the form $(e_{i_1}, \ldots, e_{i_k})$ where $1 \le i_1 \le \cdots \le i_k \le n$. It follows that any k-form ω can be written as

$$\omega = \sum_{1 \le i_1 \le \cdots \le i_k \le n} \omega(e_{i_1}, \ldots, e_{i_k}) e_{i_1}^* \wedge \cdots \wedge e_{i_k}^*$$

since both sides above take the same value on such k-tuples. Moreover a basis for $\wedge^k V^*$ is given by $\{e_{i_1}^* \wedge \cdots \wedge e_{i_k}^* : 1 \le i_1 \le \cdots \le i_k \le n\}$, so the dimension of $\wedge^k V^*$ is $\binom{n}{k}$. We note that $\wedge^n V^*$ is 1-dimensional, spanned by $e_1^* \wedge \cdots \wedge e_n^*$; when $V = \mathbb{R}^n$ and $e_1, \ldots, e_n$ is the standard basis of $\mathbb{R}^n$ then $e_1^* \wedge \cdots \wedge e_n^*$ is

just the determinant, which is an alternating multilinear function of n column vectors. We also note that $\wedge^k V^* = \{0\}$ for $k > n$.

The wedge product is associative, distributive and *skew-commutative*: $\alpha \wedge \beta = (-1)^{kl}\beta \wedge \alpha$, where α is a k-form and β an l-form.

5.2.2 Differential forms on manifolds

Just as the tangent bundle TM has the structure of a smooth vector bundle over M, so does the *cotangent bundle* $T^*M := \cup_{p\in M} T_p^* M$, as well as the *kth exterior power* of the cotangent bundle, $\wedge^k T^*M := \cup_{p\in M} \wedge^k T_p^* M$, for $1 \leq k \leq n$.

Definition 5.2.4 (k-form)**.** A k-form on a manifold M is a smooth section of the bundle $\wedge^k T^*M$, in other words a smooth map $\omega : M \to \wedge^k T^*M$ such that $\omega(p) \in \wedge^k T_p^* M$ for all p.

For example if f is a smooth function on M then its derivative df is a 1-form on M. In local coordinates, a k-form can be expressed as

$$\omega = \sum_{1\leq i_1 \leq \cdots \leq i_k \leq n} a_{i_1 \ldots i_k} dx_{i_1} \wedge \cdots \wedge dx_{i_k}$$

where the functions $a_{i_1 \ldots i_k}$ are smooth.

The wedge product of two forms is defined by taking the wedge product pointwise in each exterior algebra $A(T_p^* M), p \in M$. An important operation on forms is the *pull-back of forms*:

Definition 5.2.5 (Pull-back of forms)**.** Let $f : M \to N$ be a smooth map between manifolds and let ω be a k-form on N. The pull-back of ω by f is the k-form $f^*\omega$ on M whose action on vectors $v_1, \ldots, v_k \in T_p M$ is given by

$$(f^*\omega)_p(v_1, \ldots, v_k) := \omega_{f(p)}(df_p(v_1), \ldots, df_p(v_k))$$

Note for a function u on N, $f^*du = d(u \circ f)$ by the Chain Rule. Pull-back of forms commutes with wedge product, in the sense that $f^*(\alpha \wedge \beta) = f^*\alpha \wedge f^*\beta$.

5.2.3 Integration of differential forms

We follow closely Chapter 7, section 35 of Arnold's book in this section.

To motivate the definition of *integration of a k-form over a k-chain*, we first consider the integrals of 1-forms and 2-forms over curves and surfaces.

Integrals of 1-forms and 2-forms

Let ω be a 1-form on a manifold M and let $\gamma : [a, b] \to M$ be a smooth curve in M. The integral $\int_\gamma \omega$ can be defined as a limit of sums of the 1-form ω acting on tangent vectors ξ_i tangent to the curve γ, obtained as follows:

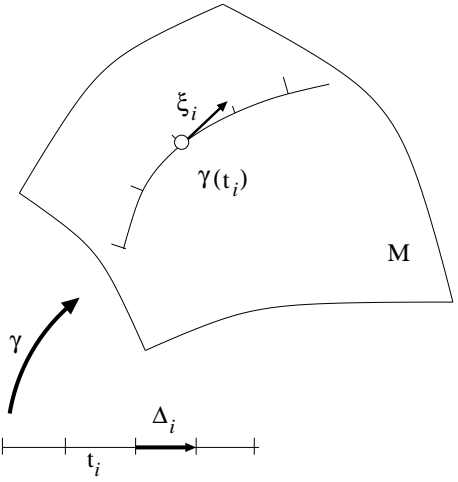


Figure 5.5: Integral of a 1-form over a curve

Given a partition $P = \{t_i\}$ of $[a, b]$, each increment $\Delta_i = t_{i+1} - t_i$ can be viewed as a tangent vector in $\mathbb{R}$ based at the point t_i, giving a tangent vector $\xi_i := d\gamma_{t_i}(\Delta_i) \in T_{\gamma(t_i)}M$. The integral of ω over γ is then defined by

$$\int_\gamma \omega := \lim_{||P||\to 0} \sum_i \omega(\xi_i)$$

(where $||P|| = \max_i \Delta_i$). The integral can also be viewed as a limit of Riemann sums for the integral $\int_a^b \phi(t)dt$ of the function ϕ on $[a, b]$ such that $\gamma^*\omega = \phi(t)dt$.

Similarly the integral $\int_\sigma \omega$ of a 2-form ω over a smooth parametrized surface in M, $\sigma : [a, b] \times [c, d] \to M$, can be defined as a limit of sums of the values of ω on 'infinitesimal parallelograms' formed by pairs of tangent vectors $\xi_{ij}^{(1)}, \xi_{ij}^{(2)} \in T_{\sigma(s_i,t_j)}M$ where $\xi_{ij}^{(1)}, \xi_{ij}^{(2)}$ are images under $d\sigma_{(s_i,t_j)}$ of increments $\Delta_{ij}^{(1)} = (s_{i+1} - s_i, 0), \Delta_{ij}^{(2)} = (0, t_{j+1} - t_j)$, thought of as tangent vectors based at the points (s_i, t_j) where $||P|| = \{s_i\}, ||Q|| = \{t_j\}$ are partitions of $[a, b], [c, d]$ respectively:

$$\int_\sigma \omega := \lim_{||P||,||Q||\to 0} \sum_{i,j} \omega(\xi_{ij}^{(1)}, \xi_{ij}^{(2)})$$

As in the case of the integral of a 1-form over a curve, the integral above can be viewed as a limit of Riemann sums for the integral $\int_{[a,b]\times[c,d]} \phi(s, t)dsdt$ where ϕ is the function such that $\sigma^*\omega = \phi(s, t)dsdt$.

Here the parametrized surface σ plays the role of "path of integration". In the following section we describe how to generalize these definitions to the case of k-forms on a manifold, where the "path of integration" will be given by

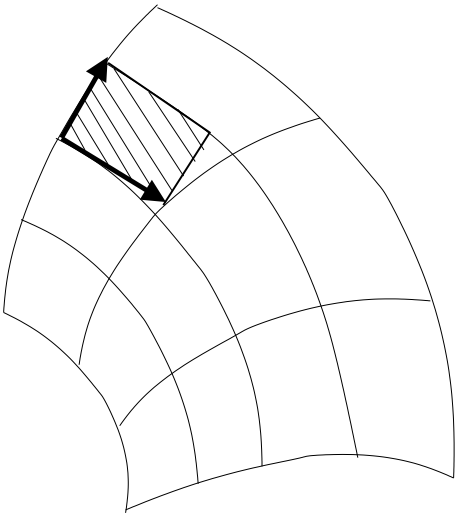


Figure 5.6: Integral of a 2-form over a surface

a smooth map $\sigma : D \subset \mathbb{R}^k \to M$ where D will be a k-dimensional, bounded, convex polyhedron in $\mathbb{R}^k$.

Integral of a k-form over a k-chain

We first need to recall the notion of *orientation* on a finite-dimensional real vector space. The simplest example is that of $\mathbb{R}^2$, where an ordered frame $(\vec{v}, \vec{w})$ is said to be *positively oriented* if the angle going anti-clockwise from $\vec{v}$ to $\vec{w}$ is less than π, and *negatively oriented* if the angle is in between π and 2π. These two cases correspond to the determinant of the matrix $[\vec{v}|\vec{w}]$ being positive and negative respectively. Equivalently, the ordered frame being positively (respectively negatively) oriented corresponds to the case when the matrix $[\vec{v}|\vec{w}]$ can (respectively cannot) be deformed into the identity matrix through a continuous path $([\vec{v_t}|\vec{w_t}])_{t\in[0,1]}$ of matrices.

In general, for an n-dimensional real vector space V, the set of ordered frames of V can be identified with the group $GL(V)$ of invertible linear maps of V to itself after choosing one distinguished ordered basis $(\vec{e_1}, \ldots, \vec{e_n})$, via the bijective correspondence which associates to $T \in GL(V)$ the ordered basis $(T(\vec{e_1}), \ldots, T(\vec{e_n}))$. The group $GL(V)$ has two connected components, which are given by $\det^{-1}(1), \det^{-1}(-1)$, where $\det : GL(V) \to \mathbb{R}$ is the determinant function. The space of ordered frames thus has two connected components, and an orientation corresponds to a choice of one of these.

Definition 5.2.6 (Orientation)**.** Let V be an n-dimensional real vector space. Two ordered frames $(\vec{v_1}, \ldots, \vec{v_n}), (\vec{w_1}, \ldots, \vec{w_n})$ are declared equivalent if they lie in the same connected component of the space of ordered frames, or equivalently

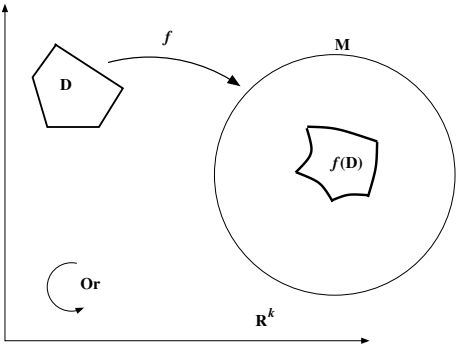


Figure 5.7: A k-cell

if the invertible linear map $T \in GL(V)$ such that $T(v_i) = w_i, i = 1, \ldots, n$ has positive determinant. An orientation on V is a choice of an equivalence class of ordered frames.

On $\mathbb{R}^n$ the orientation given by the standard ordered basis $(\vec{e_1}, \ldots, \vec{e_n})$ will be called the standard orientation on $\mathbb{R}^n$.

A *convex polyhedron* D in a real vector space V is a closed convex set with nonempty interior given by the intersection of finitely many half-spaces $H_1, \ldots, H_m$, where each half-space is a set of the form $H_i = \{v \in V | l_i(v) \leq 0\}$, $l_i : V \to \mathbb{R}$ being a non-zero linear functional on V. We assume that no half-space H_i is contained in any other half-space H_j, and that the *faces* $D_i = D \cap V_i$ of D are also polyhedra, where V_i is the subspace $\{v \in V | l_i(v) = 0\}$.

Definition 5.2.7 (Cell)**.** Let M be an n-manifold and let $1 \leq k \leq n$. A k-cell is a triple $\sigma = (D, f, Or)$ where $D \subset V$ is a bounded, convex polyhedron in a k-dimensional real vector space V, $f : D \to M$ is a smooth map, and Or is an orientation on V.

Definition 5.2.8 (Integration of a form over a cell)**.** Given a k-form ω and a k-cell $\sigma = (D, f, Or)$, the pull-back form $f^*\omega$ is a k-form on $D \subset V$ and is hence of the form $\phi(x)\, dx_1 \wedge \cdots \wedge dx_k$ for some smooth function ϕ on D, where $x_1 = e_1^*, \ldots, x_k = e_k^* : V \to \mathbb{R}$ are linear coordinates on V dual to an ordered basis $(e_1, \ldots, e_n)$ of V giving the orientation Or. We define the integral of ω over the cell σ to be

$$\int_\sigma \omega := \int_D \phi dx_1 \ldots dx_k$$

(where the integral on the right-hand side is given by the usual limit of Riemann sums).

The integral as defined above is independent of the choice of coordinate system $x_1, \ldots, x_n$ on V (this can be seen using the Change of Variables Formula from multivariable calculus).

Note that for the integral over a cube $C = [0,1]^k$ in $\mathbb{R}^k$, taking the partition $\{i/N : 0 \leq i \leq N\}$ of $[0,1]$ divides the cube into N^k small cubes $C_{\underline{i}}$ where $\underline{i} \in \{0, \ldots, N-1\}$, and the integral can be written as a limit as $N \to \infty$ of sums of the form

$$\sum_{\underline{i}} \omega(\xi^1_{\underline{i}}, \ldots, \xi^k_{\underline{i}})$$

where the tangent vectors $\xi^1_{\underline{i}}, \ldots, \xi^k_{\underline{i}}$ are the images under df of the sides of $C_{\underline{i}}$ at the vertex $p_{\underline{i}} = \underline{i}/N$ (the sides being thought of as tangent vectors based at the point $p_{\underline{i}}$).

It will also be useful to consider *chains*, given by finite linear combinations of cells.

Definition 5.2.9 (Chains)**.** A k-chain c on M consists of a finite collection of k cells $\sigma_1, \ldots, \sigma_k$ together with associated multiplicities $m_1, \ldots, m_k$ which are integers, denoted by $c = m_1 \cdot \sigma_1 + \cdots + m_k \cdot \sigma_k$, where for any cell $\sigma = (D, f, Or)$ the chain $-1 \cdot \sigma$ is identified with the cell $\sigma^{op} = (D, f, Or^{op})$, Or^{op} being the opposite orientation to Op. In other words the set of k-chains is the free abelian group C_k on the set of k-cells, modulo the relations $\sigma^{op} = -1 \cdot \sigma$ for all cells σ.

The integral of a k-form ω on a k-chain $c = \sum_i m_i \sigma_i$ is defined by

$$\int_c \omega := \sum_i m_i \int_{\sigma_i} \omega$$

This defines a *k-cochain*, or a homomorphism from the group of k-chains to $\mathbb{R}$,

$$\begin{aligned} \hat{\omega} : C_k &\to \mathbb{R} \\ c &\mapsto \int_c \omega \end{aligned}$$

The cochain $\hat{\omega}$ completely determines ω:

If $\xi_1, \ldots, \xi_k$ are tangent vectors based at a point x of M, then we can define a family of 'shrinking' curvilinear k-dimensional parallelepipeds Π_ϵ based at x tangent to these vectors near x, by taking a smooth map f from $\mathbb{R}^k$ into M such that $f(0) = x$ and df_0 maps the standard basis $e_1, \ldots, e_k$ to the vectors $\xi_1, \ldots, \xi_k$, and letting Π_ϵ be the image under f of the k-dimensional cube with side length ϵ, $[0, \epsilon]^k$ in $\mathbb{R}^k$ based at 0 with sides parallel to $e_1, \ldots, e_k$. Then one can show that

$$\omega(\xi_1, \ldots, \xi_k) = \lim_{\epsilon \to 0} \frac{\int_{\Pi_\epsilon} \omega}{\epsilon^k} = \lim_{\epsilon \to 0} \frac{\hat{\omega}(\Pi_\epsilon)}{\epsilon^k} \qquad -(*)$$

(where by abuse of notation we write Π_ϵ for the k-cell given by $[0, \epsilon]^k, f$, and the standard orientation on $\mathbb{R}^k$).

5.2.4 Boundary of a chain

Given a k-cell $\sigma = (D, f, Or)$, one can define a $(k-1)$-chain $\partial\sigma$ called the *boundary of* σ, given by "adding the faces of the polyhedron D with orientations taken to be compatible with the orientation of D", as follows:

Each face D_i of the polyhedron D is a $(k-1)$-dimensional polyhedron in the vector space V_i. Restricting f to D_i gives a smooth map $f_i := f_{|D_i}$ of D_i into M. We choose an orientation Or_i on V_i which is "compatible" with the orientation Or on D in the following sense:

Let $\vec{n}$ be an "outward" pointing normal vector on the face D_i of the polyhedron (so that $l_i(\vec{n}) > 0$). Choose the orientation Or_i on V_i such that if it is given by an ordered frame $(\vec{v_1}, \ldots, \vec{v}_{k-1})$, then the ordered frame $(\vec{n}, \vec{v_1}, \ldots, \vec{v}_{k-1})$ gives the orientation Or on V. We obtain a $(k-1)$-cell $\sigma_i := (D_i, f_i, Or_i)$.

Definition 5.2.10 (Boundary of a chain)**.** The boundary of a k-cell $\sigma = (D, f, Or)$, denoted by $\partial\sigma$ is the $(k-1)$-chain given by the sum of the $(k-1)$-cells σ_i defined above,

$$\partial\sigma := \sum_i \sigma_i$$

The boundary map from k-cells to $(k-1)$-chains then extends uniquely to a homomorphism from the group C_k of k-chains to the group C_{k-1} of $(k-1)$-chains, which we also denote by ∂.

For any chain c, one can show that

$$\partial^2 c = 0$$

("each cell in $\partial^2 c$ occurs twice with opposite signs").

Chains c such that $\partial c = 0$ are called *cycles*, while chains c which can be written in the form $\partial c'$ for some chain c' are called *boundaries*. Since $\partial^2 = 0$, every boundary is a cycle. The quotient group $\{k - \text{cycles}\}/\{k - \text{boundaries}\}$ is called the k-th homology group $H_k(M, \mathbb{Z})$ of M with integer coefficients. Similarly allowing chains $\sum_i r_i \cdot \sigma_i$ with real coefficients r_i one can define the k-th homology group $H_k(M, \mathbb{R})$ of M with real coefficients. It is a fact that for a compact manifold, $H_k(M, \mathbb{Z})$ is a finitely generated abelian group, while $H_k(M, \mathbb{R})$ is a finite dimensional real vector space, whose dimension β_k is called the kth Betti number of M.

5.2.5 Exterior derivative of differential forms

Let $C_k^* = Hom(C_k, \mathbb{R})$ be the group of k-cochains, then the boundary homomorphism $\partial : C_{k+1} \to C_k$ determines a dual homomorphism on cochains

$$\begin{aligned} \partial^* : C_k^* &\to C_{k+1}^* \\ \eta &\mapsto (\partial^*\eta : c \mapsto \eta(\partial c)) \end{aligned}$$

Since any k-form ω is determined by the corresponding k-cochain $\hat{\omega}$, it is reasonable to ask whether the $(k+1)$-cochain $\partial^*\hat{\omega}$ corresponds to some $(k+1)$-form. It turns out that this is indeed the case:

Theorem 5.2.11. *(Existence of exterior derivative) Let ω be a k-form on a smooth manifold M. Then there exists a unique $(k+1)$ form called the exterior derivative of ω, denoted by $d\omega$, such that for any $\xi_1, \ldots, \xi_{k+1} \in T_x M$,*

$$d\omega(\xi_1, \ldots, \xi_{k+1}) = \lim_{\epsilon \to 0} \frac{\int_{\partial \Pi_\epsilon} \omega}{\epsilon^{k+1}} = \lim_{\epsilon \to 0} \frac{(\partial^* \hat{\omega})(\Pi_\epsilon)}{\epsilon^{k+1}} \qquad -(**)$$

(where Π_ϵ is a $(k+1)$-dimensional curvilinear parallelepiped corresponding to the vectors $\xi_1, \ldots, \xi_{k+1}$ as defined above).

In local coordinates, if

$$\omega = \sum_{1 \le i_1 \le \cdots \le i_k \le n} a_{i_1 \ldots i_k} dx_{i_1} \wedge \cdots \wedge dx_{i_k},$$

then

$$d\omega = \sum_{1 \le i_1 \le \cdots \le i_k \le n} da_{i_1 \ldots i_k} \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k}$$

Note if ω is a 0-form, in other words a function f, then the exterior derivative coincides with the derivative df of the function. Moreover, $\partial^2 = 0$ implies $\partial^{*2} = 0$, and hence

$$d^2 = 0$$

5.2.6 Stokes' Theorem

Stokes' theorem is a generalization to higher dimensions of the Fundamental Theorem of Calculus,

$$\int_a^b f'(t)dt = f(b) - f(a).$$

Stokes' Theorem asserts the integral of the exterior derivative of a form along a chain (which corresponds to the left-hand side above) equals the integral of the form along the boundary of the chain (which corresponds to the right-hand side above):

Theorem 5.2.12. *(Stokes' Theorem). For any k-form ω on a manifold M and any $(k+1)$-chain c,*

$$\int_c d\omega = \int_{\partial c} \omega$$

By linearity it suffices to prove the Theorem for cells. We briefly sketch the proof of Stokes' Theorem for a curvilinear parallelepiped Π:

We may assume the curvilinear parallelepiped Π is given by a smooth map f from the $(k+1)$-dimensional cube $[0,1]^{k+1}$ into M. Taking the partition $P = \{i/n : 0 \le i \le N\}$ of the interval $[0,1]$ gives a partition of the cube $[0,1]^{k+1}$ into N^{k+1} small cubes of side length $1/N$ each, and hence gives N^{k+1} small curvilinear parallelepipeds $\Pi^{\underline{i}}$, where $\underline{i} = (i_1, \ldots, i_{k+1})$ with $0 \le i_1, \ldots, i_{k+1} \le (N-1)$. Then

$$\int_{\partial \Pi} \omega = \sum_{\underline{i}} \int_{\partial \Pi^{\underline{i}}} \omega$$

since the integrals over faces common to two small curvilinear parallelepipeds cancel. As $N \to \infty$, the integrals over the boundaries of the small curvilinear parallelepipeds, are given, using $(**)$, by

$$\int_{\partial \Pi^{\underline{i}}} \omega = d\omega(\xi_1^{\underline{i}}, \ldots, \xi_{k+1}^{\underline{i}}) + o\left(\left(\frac{1}{N}\right)^{k+1}\right)$$

uniformly in $\underline{i}$, thus

$$\begin{aligned}\int_{\partial \Pi} \omega &= \sum_{\underline{i}} \int_{\partial \Pi^{\underline{i}}} \omega \\ &= \sum_{\underline{i}} \left(d\omega(\xi_1^{\underline{i}}, \ldots, \xi_{k+1}^{\underline{i}}) + o\left(\left(\frac{1}{N}\right)^{k+1}\right)\right) \\ &= \left(\sum_{\underline{i}} d\omega(\xi_1^{\underline{i}}, \ldots, \xi_{k+1}^{\underline{i}})\right) + o(1),\end{aligned}$$

and the first term on the right-hand side above we recognize as a Riemann sum for the integral $\int_\Pi d\omega$, so letting $N \to \infty$ gives Stokes' Theorem.

5.3 de Rham cohomology of smooth manifolds

Definition 5.3.1 (Exact and closed forms, de Rham cohomology)**.** A k-form α is called *exact* if it is of the form $\alpha = d\beta$ for some $(k-1)$-form β, and *closed* if it satisfies $d\alpha = 0$. Since $d^2 = 0$, it follows that any exact form is closed. The de Rham cohomology groups of M, denoted by $H^k_{dR}(M)$, are defined to be the quotient of the space of closed k-forms by the subspace of exact k-forms,

$$H^k_{dR}(M) := (\text{Ker } d : \Omega^k(M) \to \Omega^{k+1}(M))/(\text{Im } d : \Omega^{k-1}(M) \to \Omega^k(M))$$

5.3.1 Poincaré Lemma

Not all closed forms are exact, a well-known example being given by the 1-form

$$\omega = \frac{-y}{x^2+y^2}dx + \frac{x}{x^2+y^2}dy,$$

on the punctured plane $\mathbb{R}^2 - \{0\}$, which is closed because locally it can be written as $d\theta$ where θ is a local branch of the argument function, but it is not exact because for the closed curve $\gamma : t \in [0, 2\pi] \mapsto (\cos t, \sin t)$ the integral $\int_\gamma \omega$ equals 2π, whereas if ω were exact, equal to df for some function f, then we would have

$$\int_\gamma \omega = \int_\gamma df = f(\gamma(2\pi)) - f(\gamma(0)) = 0$$

since γ is a closed curve.

As in the above example, any form which is locally exact will clearly be closed. The Poincaré Lemma asserts that the converse is also true, namely any form which is closed is locally exact. Since any point in a manifold has a neighbourhood diffeomorphic to a ball in $\mathbb{R}^n$, it suffices to prove the following:

Theorem 5.3.2. *(Poincaré Lemma). Any closed k-form α on a ball B in $\mathbb{R}^n$ is exact (there is a $(k+1)$-form β on B such that $\alpha = d\beta$).*

We give a brief sketch of the proof, using the "cone construction" on chains:

Given a k-cell $\sigma = (D, f, Or)$ in the ball B, we construct a cell $p\sigma$ called the "cone over σ". Assume the ball B is centered at the origin in $\mathbb{R}^n$,$D \subset \mathbb{R}^k \times \{1\} \subset \mathbb{R}^{k+1}$ and Or is the standard orientation on $\mathbb{R}^k \times \{1\}$. Choose a point $x \in \mathbb{R}^k \times \{0\}$. Let $pD := \{(1-t)x + ty : t \in [0,1], y \in D\} \subset \mathbb{R}^{k+1}$ be the polyhedron given by joining all points of D to the point x by straight lines. Extend the map $f : D \to B$ to a map $pf : pD \to B$ by defining $pf : (1-t)x+ty \mapsto tf(y)$, so that $pf_{|D} = f$ and $pf(x) = 0$. Let pOr be the standard orientation on $\mathbb{R}^{k+1}$.

Definition 5.3.3 (Cone over a chain)**.** The cone over the cell σ is defined to be the $(k+1)$-cell $p\sigma := (pD, pF, pOr)$. The cone map p on cells extends uniquely to a homomorphism on chains, $p : C_k \to C_{k+1}$. For a chain c, the chain pc is called the cone over the chain c.

For example if we take a 1-cell given by a straight line segment L, the cone over L is a triangle pL with one side equal to L and the other two sides given by the cone on the boundary of L (which is the pair of endpoints of L) which occur with orientations opposite to the orientations induced from the orientation on the triangle pL, so the boundary of pL is given by $\partial pL = L - p\partial L$. Similarly, in general we have

$$\partial \circ p = id - p \circ \partial$$

on all k-chains. Then for the dual homomorphisms we have

$$p^* \circ \partial^* = id - \partial^* \circ p^*$$

The dual homomorphism $p^* : C_k^* \to C_{k-1}^*$ on cochains can also be expressed, like ∂^*, as an operation on forms, since for any $(k-1)$-cell σ and any k-form ω, the integral over the cell $p\sigma$ can be written as an iterated integral,

$$\begin{aligned} p^*(\hat{\omega})(\sigma) &= \int_{p\sigma} \omega \\ &= \int_{\sigma} P\omega \end{aligned}$$

where the form $P\omega$ is a $(k-1)$-form whose value on vectors $\xi_1, \ldots, \xi_{k-1} \in T_q\mathbb{R}^n$ (where $q \in B$) is given by

$$(P\omega)_q(\xi_1, \ldots, \xi_{k-1}) = \int_0^1 \omega_{tq}(q, t\xi_1, \ldots, t\xi_{k-1})\ dt$$

From the relation $p^* \circ \partial^* = id - \partial^* \circ p^*$, it follows that $P \circ d = id - d \circ P$ on $\Omega^k(B)$, and hence if α is a closed form in B, then

$$\alpha = d(P\alpha)$$

5.3.2 de Rham's Theorem

As noted before, any k-form ω gives a k-cochain by integration on chains, $(c \mapsto \int_c \omega)$. If the form is closed, then by Stokes' Theorem, adding a k-boundary $\partial c'$ to c does not change the integral, since

$$\int_{c+\partial c'} \omega = \int_c \omega + \int_{\partial c'} \omega = \int_c \omega + \int_{c'} d\omega = \int_c \omega$$

(as $d\omega = 0$). Thus any closed form ω defines a linear map $\hat{\omega} : H_k(M, \mathbb{R}) \to \mathbb{R}, [c] \mapsto \int_c \omega$. Furthermore, adding an exact form $d\beta$ to ω does not change the integral over a cycle either, since

$$\int_c \omega + d\beta = \int_c \omega + \int_{\partial c} \beta = \int_c \omega$$

(as $\partial c = 0$). It follows that we have a well-defined bilinear pairing $H^k_{dR}(M) \times H_k(M, \mathbb{R}) \to \mathbb{R}, ([\omega], [c]) \mapsto \int_c \omega$. Then we have:

Theorem 5.3.4. *(de Rham's Theorem) The bilinear pairing $H^k_{dR}(M) \times H_k(M, \mathbb{R}) \to \mathbb{R}$ is nondegenerate.*

References

[1] V. I. Arnold, *Mathematical methods of classical mechanics*. Graduate Texts in Mathematics, 60 (Springer, New York, 1978).

[2] R. Bott and L. W. Tu, *Differential forms in algebraic topology*. Graduate Texts in Mathematics, 82 (Springer, New York, 1982).

[3] I. Madsen and J. Tornehave, *From calculus to cohomology, de Rham cohomology and characteristic classes* (Cambridge University Press, Cambridge, 1997).

II An Introduction to Riemannian Geometry

"A *geometry* is a topological space where lines, circles, angles, triangles, distance, isometries, etc. are defined. The Euclidean Geometry we studied in school is the most elementary *geometry*. It is based on 5 axioms introduced in "Elements" by Euclid in 300 B.C.

Euclid's Axioms :

1. Given two points, there is a line segment connecting them.
2. Any line segment can be extended to a line.
3. Given a point and a positive real number, there is a circle with that point as a center and the number as radius.
4. All right angles are congruent.
5. (Parallel postulate:) Given a line and a point not on the line there is at most one line through the point that does not meet the given line.

Are these axioms mutually independent? It is easy to construct geometries violating one of the first four axioms by taking suitable subspaces of the Euclidean space. It was not easy to prove that the parallel postulate can't be derived from previous four postulates. This question leads to the construction of a Non-Euclidean geometry.

5.4 A Non-Euclidean Geometry

Let Ω be an open connected subset of $\mathbb{R}^2$. For example, we can take Ω to be the unit disc D defined by

$$D = \{(x, y) | x^2 + y^2 < 1\}.$$

A curve in Ω is a smooth function $\sigma : [0, 1] \to \Omega$

$$\sigma(t) = (\sigma_1(t), \sigma_2(t)),$$

where $\sigma_i(t) : [0, 1] \to \mathbb{R}$ are differentiable functions. To define distance between points in Ω, let us first define the length of a curve in Ω. The usual definition of length is

$$l(\sigma) = \int_0^1 \|\sigma'(t)\| \, dt,$$

where $\|v\| = \langle v, v \rangle^{\frac{1}{2}}$.

Proposition. *The formula for distance in the Euclidean space is*

$$\begin{aligned} d(p, q) &= \|p - q\| \\ &= \inf_{\sigma} \int_0^1 \|\sigma'(t)\| \, dt \\ &= \inf_{\sigma} l(\sigma). \end{aligned}$$

Let γ be another curve which intersects σ. The angle between curves at their point of intersection, $\gamma(t_0) = \sigma(t_0)$, is the angle between $\gamma'(t_0)$ and $\sigma'(t_0)$ which is

$$\cos^{-1}\left(\frac{\langle \gamma'(t_0), \sigma'(t_0)\rangle}{\|\gamma'(t_0)\|\|\sigma'(t_0)\|}\right).$$

So, the inner product between tangent vectors is an important quantity to define distance, length and angle. We now change the inner product at each point as follows.

Let $f : \Omega \to (0, \infty)$ be a positive smooth function. Then we can define a new notion of inner product at the tangent space of a point p as follows:

$$\langle v, w\rangle_f := f(p)^2\langle v, w\rangle.$$

In particular,

$$\|v\|_f = f(p)\|v\|.$$

Example 1: If we let $f \equiv 1$, we get the old inner product at each point.

Example 2: $\Omega = D$, $f(p) = \dfrac{1}{1 - \|p\|^2}$.

Let us see how the notions of length and angles change. The new length is,

$$l_f(\sigma) = \int_0^1 f(\sigma(t))\|\sigma'(t)\|dt,$$

and the angle α between v and w is

$$\begin{aligned}\alpha &= \cos^{-1}\left(\frac{\langle v, w\rangle_f}{\|v\|_f\|w\|_f}\right)\\ &= \cos^{-1}\left(\frac{\langle v, w\rangle}{\|v\|\|w\|}\right).\end{aligned}$$

Therefore, the angle between two curves does not change whereas the length of curves changes. To define distance between two points in Ω, we use Proposition 1, i.e.,

$$d_f(p, q) = \inf_\sigma l_f(\sigma),$$

where σ is any curve joining p and q. If this infimum is realized by a smooth curve, then the curve is called a *geodesic* (or a line segment). Hence, the geodesics of $\mathbb{R}^2$ are straight lines.

Proposition. *(Ω, d_f) is a metric space, i.e.,*
(i) $d_f(p, q) \geq 0$ and $d_f(p, q) = 0$, iff $p = q$,
(ii) $d_f(p, q) = d_f(q, p)$,
(iii) $d_f(p, q) \leq d_f(p, r) + d_f(r, q)$.

A *line* is an infinitely extended curve,

$$\sigma : (-\infty, \infty) \to \Omega,$$

such that σ is a geodesic between any two points on it. The *circle* with center p and radius r is the set

$$\{x \in \Omega | d_f(p, x) = r\}.$$

Next, we define *congruence*. In Euclidean space, two sets S_1 and S_2 are said to be congruent if S_2 can be obtained by translating and rotating S_1. More precisely, there is a point $b \in \mathbb{R}^2$ and a matrix

$$A = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix},$$

such that if

$$T(v) = A(v) + b,$$

then

$$T(S_1) = T(S_2),$$

i.e., T is an isometry (distance preserving function). Conversely,

Proposition. *If $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ is an isometry, then ϕ is of the form*

$$\phi(p) = A(p) + b.$$

Motivated by this fact, we say S_1, $S_2 \subset \Omega$ are congruent if there is an isometry ϕ of (Ω, d_f) with

$$\phi(S_1) = (S_2).$$

Proposition. *Let ϕ be an isometry of (Ω, d_f) and σ be a curve joining two points p and q in Ω.*
(i) $l_f(\phi \circ \sigma) = l_f(\sigma)$.
(ii) Let γ be another curve making an angle θ with σ at $\sigma(t_0) = \gamma(t_0)$. Then $\phi \circ \sigma$ and $\phi \circ \gamma$ makes the same angle θ at $\phi \circ \sigma(t_0) = \phi \circ \gamma(t_0)$.

The proof is given in the next section.

5.4.1 Hyperbolic Geometry

The geometry on D defined by the function $f(p) = \frac{1}{1-\|p\|^2}$ (see Example 2) is called Hyperbolic geometry. Geodesics in D are lines passing through the origin and circles intersecting the boundary of D orthogonally.
Example 3: Consider the geodesic

$$\sigma(t) = \left(\frac{e^{2t}-1}{e^{2t}+1}, 0\right), \qquad t \in (-\infty, \infty).$$

If $a(t) = \dfrac{e^{2t}-1}{e^{2t}+1}$, then $a' = 1 - a^2$. This implies,

$$\|\sigma'(t)\|_f = 1.$$

Therefore,

$$l(\sigma|_{[0,T]}) = T.$$

So, σ has infinite length.

Note that Euclidean norm of σ' is

$$\|\sigma'(t)\| = \frac{4e^{2t}}{(e^{2t}+1)^2} \to 0, \qquad \text{as} \qquad |t| \to \infty.$$

However, because of the factor of $[1 - \|\sigma(t)\|]^{-2}$, we have

$$\|\sigma'(t)\|_f = 1.$$

It is easy to see that the parallel postulate fails in Hyperbolic geometry.

Euclid's axioms for hyperbolic geometry

1. Given two points in D, we can find a circle orthogonal to the boundary connecting these two points.
2. Any geodesic can be extended to a line.
3. Circle exists.
4. Let σ_1, σ_2 and γ_1, γ_2 be pairs of infinite geodesics intersecting orthogonally.
   Let ϕ be an isometry taking σ_1 to γ_1, i.e., $\phi \circ \sigma_1 = \gamma_1$. Then $\phi \circ \sigma_2$ is a geodesic meeting γ_1 orthogonally. Therefore, $\phi \circ \sigma_2 = \gamma_2$.
5. Failure of parallel postulate.

5.5 Riemannian Geometry

Let M be a smooth manifold with dimension n, and p be a point in it. There exists a diffeomorphism ϕ from an open neighborhood U of p to an open neighborhood $V \subset \mathbb{R}^n$ of the origin such that $\phi(p) = 0$. A curve in U is a map $\sigma : (-\epsilon, \epsilon) \to U$ such that $\sigma \circ \phi : (-\epsilon, \epsilon) \to \mathbb{R}^n$ is smooth. Consider,

$$\mathcal{C} = \{\sigma : \sigma(0) = p\}.$$

If $(\sigma_1 \circ \phi)'(0) = (\sigma_2 \circ \phi)'(0)$ then $\sigma_1 \sim \sigma_2$ for any $\sigma_1, \sigma_2 \in \mathcal{C}$. This defines an equivalence relation in $\mathcal{C}$. The set of all equivalence classes is called the tangent space at p and denoted by T_pM.

Let v be a vector in $\mathbb{R}^n$. There exists an $\epsilon > 0$ such that $tv \in V$ for all $t \in \epsilon$. Now define $\sigma(t) = \phi^{-1}(tv)$ for all $t \in \epsilon$. Hence σ is a curve passing through p and $(\sigma \circ \phi)'(0) = v$. This gives a bijection from T_pM to $\mathbb{R}^n$. Using this bijection one can define addition and scalar multiplication on T_pM to make it a vector space over $\mathbb{R}$. Let $\phi : M \to N$ be a smooth map where N is another

smooth manifold. Then $\phi_{*|p}$, the differential of ϕ at any point p is a linear transformation from T_pM to $T_{\phi(p)}N$ defined by

$$\phi_{*|p}([\sigma]) = [\phi \circ \sigma],$$

where σ is a curve passing through p. If ϕ is a diffeomorphism then $\phi_{*|p}$ is an invertible linear transformation.

The tangent bundle of M denoted by TM is a vector bundle on M consisting of all tangent spaces of M. A vector field is a smooth function $X : M \to TM$ such that $X(p) \in T_pM$. In order to define a geometry on M, we define inner products on each tangent space.

Definition: g is said to be a *Riemannian metric* on M if

(i) $g(p)$ is an inner product on T_pM for each $p \in M$.

(ii) $g(p)$ varies smoothly with p, i.e., if X and Y are two vector fields on M then $g(X, Y)$ is a smooth function on M.

A smooth manifold with a Riemannian metric is called a *Riemannian manifold*. We use g_p to denote g at $p \in M$.

Theorem 5.5.1. *Every smooth manifold admits a Riemannian metric.*

Proof. We refer to Ref. 1 for a detailed proof of the theorem. □

Let (M, g) be a Riemannian manifold and $v \in T_pM$.

$$|v| := \sqrt{g_p(v, v)}.$$

If $\gamma : [a, b] \to M$ is a curve on M then the length of γ is defined by,

$$l(\gamma) := \int_a^b |\gamma'(t)| dt.$$

Let $\sigma : [a, b] \to M$ be a curve intersecting γ at $\sigma(t_0) = \gamma(t_0)$. The angle θ between these two curve is defined by the following formula:

$$\cos(\theta) = \frac{g(\sigma'(t_0), \gamma'(t_0))}{|\sigma'(t_0)||\gamma'(t_0)|}.$$

The distance d between any two points p and q is defined by,

$$d(p, q) = \inf_{\sigma} l(\sigma).$$

If the infimum is realized at a curve γ then γ is said to be a *geodesic*. Geodesics are very important geometric quantities. It is not easy to obtain an equation of a geodesic from the above definition. There is an alternative way. We know that the geodesics of a Euclidean space are straight lines. They are solutions to the following ODE:

$$\gamma''(t) = 0.$$

To obtain an analogous analytic definition of a geodesic in a Riemannian manifold we need a notion of differentiation of vector fields on M. Let $\chi(M)$ denote the space of all vector fields on M.

Definition: A Riemannian connection is a bilinear map, $\nabla : \chi(M) \times \chi(M) \to \chi(M)$, satisfying the following properties.

(i) ∇ is $C^\infty M$-linear on first co-efficient and $\mathbb{R}$-linear on the second, i.e.,

$$\nabla_{fX} cY = fc\nabla_X Y \ \forall \ f \in C^\infty M.$$

(ii) ∇ is torsion free, i.e.,

$$\nabla_X Y - \nabla_Y X = [X, Y].$$

(iii) ∇ is metric, i.e.,

$$X.g(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla_X Z).$$

Theorem 5.5.2. *There exists a unique Riemannian connection on a Riemannian manifold.*

Proof.

$$\begin{aligned} 2g(\nabla_X Y, Z) &= X.g(Y, Z) + Y.g(Z, X) - Z.g(X, Y) \\ &\quad + g([X, Y], Z) - g([X, Z], Y) - g([Y, Z], X). \end{aligned}$$

□

Let $X(t)$ be a vector field along a curve $\gamma(t)$ which can be extended to a vector field on M. The *covariant derivative* of $X(t)$ along γ is defined by

$$\frac{D}{dt} X(t) = \nabla_{\gamma'(t)} X.$$

A curve γ is a *geodesic* if the covariant derivative of $\gamma'(t)$ along $\gamma(t)$ is zero, i.e.,

$$\frac{D}{dt}\gamma'(t) = 0, \text{ or } \gamma''(t) = 0.$$

Proposition. *If γ is a geodesic then $|\gamma'(t)|$ is constant.*

Proof.

$$\begin{aligned} \frac{d}{dt}(|\gamma'(t)|^2) &= \frac{d}{dt} g_p(\gamma'(t), \gamma'(t)) \\ &= 2g_p\left(\frac{D}{dt}\gamma'(t), \gamma'(t)\right) \\ &= 0. \end{aligned}$$

Therefore, $|\gamma'(t)|^2$ =constant. □

Theorem 5.5.3. *Let γ be a curve joining two points p and q in M. If $|\gamma'(t)|$ is constant and $l(\gamma) = d(p,q)$ then $\gamma'' = 0$.*

Proof. We refer to the first variational formula in Ref. 1. □

A diffeomorphism ϕ of M is said to be an isometry of (M, g) if for every $p \in M$

$$g_{\phi(p)}(\phi_*(v), \phi_*(w)) = g_p(v, w) \ \forall \ v, w \in T_pM.$$

Proposition. *If ϕ is an isometry of (M, g) then*

$$d(\phi(p), \phi(q)) = d(p, q).$$

Proof. Let $\sigma : [0, 1] \to M$ be a curve between p and q. Then $\phi \circ \sigma$ is a curve between $\phi(p)$ and $\phi(q)$.

$$\begin{aligned} l(\phi \circ \sigma) &= \int_0^1 |(\phi \circ \sigma)'(t)| dt \\ &= \int_0^1 g(\phi_* \sigma'(t), \phi_* \sigma'(t))^{\frac{1}{2}} dt \\ &= \int_0^1 g(\sigma'(t), \sigma'(t))^{\frac{1}{2}} dt \\ &= l(\sigma). \end{aligned}$$

This completes the proof. □

Proposition. *If ϕ is an isometry of (M, g) then*

$$\nabla_{\phi_* X} \phi_* Y = \nabla_X Y.$$

Proof. The proof follows from the equation in the proof of theorem 2. □

Proposition. *Let ϕ be an isometry of (M, g) and γ be a geodesic. Then $\phi \circ \gamma$ is also a geodesic.*

Proof.

$$\nabla_{\phi_* \gamma'} \phi_* \gamma' = \nabla_{\gamma'} \gamma'.$$

□

Example 4: Let $(M, g) = (\mathbb{R}^2, \text{std})$, and X, Y be two vector fields on $\mathbb{R}^2$. Let $X = \{X_1, X_2\}$ and $Y = \{Y_1, Y_2\}$. Here std mean the standard inner product. Tnen,

$$\nabla_X Y = X.Y$$

where

$$X.Y(p) = \left(X_1 \frac{\partial Y_1}{\partial x} + X_2 \frac{\partial Y_1}{\partial y}, X_1 \frac{\partial Y_2}{\partial x} + X_2 \frac{\partial Y_2}{\partial y} \right).$$

The geodesics are straight lines.

Example 5: Consider $S^2 = \{(x, y, z) \in \mathbb{R}^3 | x^2 + y^2 + z^2 = 1\}$. Let X and Y be two vector fields on S^2. Then

$$g(X, Y) = \langle X, Y \rangle,$$

and

$$\nabla_X Y(p) = X.Y(p) - \langle X.Y(p), p \rangle p.$$

Consider the curve $\gamma : [0, \pi] \to S^2$ given by

$$\gamma(t) = (\cos(t), \sin(t), 0),$$

$$\frac{D}{dt}\gamma'(t) = 0.$$

Hence, γ is a geodesic.

For any $A \in SO(n+1)$, $x \to Ax$ is an isometry of S^n, and, therefore, $A\gamma$ is a geodesic. Therefore, great circles are geodesics of S^n.

Example 6 : Let M be a sub-manifold of $\mathbb{R}^n$. Then $T_pM \subset T_p\mathbb{R}^n$. Define a Riemannian metric g on M by restricting the standard inner product of $\mathbb{R}^n$ to T_pM.

$$g_p(v, w) = \langle v, w \rangle, \qquad \forall v, w \in T_pM.$$

$\nabla_X Y(p)$ is the component of $X.Y(p)$ along T_pM.

Example 7 (Flat Torus) : $\{(z_1, z_2) \in \mathbb{C}^2 : |z_1| = |z_1| = 1\}$ is a sub-manifold of $\mathbb{R}^4$.

5.5.1 Riemannian curvature

Riemannian Curvature is a $(3, 1)$-tensor on M defined by

$$R(X, Y, Z) = \nabla_Y \nabla_X Z - \nabla_X \nabla_Y Z + \nabla_{[X,Y]} Z,$$

$$R(X, Y, Z, W) = g(R(X, Y, Z), W),$$

where X, Y, Z, W are vector fields on M.

Sectional Curvature : Let $v, u \in T_pM$.

$$sec(v, u) = \frac{R_p(u, v, u, v)}{g(u, u)g(v, v) - (g(u, v))^2}.$$

If u and v are orthogonal unit vectors then

$$sec_p(u, v) = R_p(u, v, u, v).$$

We say an (M, g) has constant curvature if all sectional curvatures are equal.

Example 1: $(M, g) = (\mathbb{R}^n, \text{std})$.

$$\begin{aligned} R(X, Y, Z) &= \nabla_Y \nabla_X Z - \nabla_X \nabla_Y Z + \nabla_{[X,Y]} Z \\ &= Y.(X.Z) - X.(Y.Z) + [X, Y].Z \\ &= [Y, X].Z + [X, Y].Z \\ &= 0. \end{aligned}$$

Example 2: $(M, g) = (S^n, \text{std})$. $SO(n+1)$ acts on S^n isometrically. S^n has constant sectional curvature. One can prove that it is positive.
Example 3: $(\mathbb{H}^n, \text{std})$ has constant negative curvature.

References

[1] S. Gallot, D. Hulin, and J. Lafontaine, *Riemannian Geometry* (Springer, Berlin, 2004).

[2] J. Lee, *Introduction to smooth manifolds* (Springer, Berlin, 2002).

[3] S. Kumaresan, *A course in Differential geometry and Lie groups* (Hindustan Book Agency, Pune, 2002).

6

Vector Bundles

Utsav Choudhury and Atreyee Bhattacharya

The aim in the first part of this chapter is to understand the basics of Vector bundles and K-theory. We will define vector bundles and give several examples of vector bundles that arise naturally in geometry. We will give constructions of important natural operations on vector bundles, and we will show how deformation of spaces controls the structure of vector bundles. In section 6.5 we will define K-theory, an important abelian invariant of a space. We will show that this invariant can be used to answer non-trivial questions about the geometry of a space.

The second part gives a brief overview of the Chern-Weil theory in the context of vector bundles. The Chern-Weil theory is a vast topic which has been studied from various aspects. In this short note we take the differential geometric approach. We start with smooth manifolds and affine connections and generalize the notion of connections and curvature to vector bundles with an aim to produce global invariants of vector bundles in terms of characteristic classes. We conclude with a few simple examples.

I Vector Bundles and K-theory

6.1 Introduction

Let X be a space; suppose we want to understand the space using continuous (or differentiable) maps from X to real or complex vector spaces. For instance, continuous real valued functions (resp. complex valued functions) records a lot of information about the topological nature of the space. If we take only

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_6

differentiable functions then the derivatives of the functions determine the differential structure of the space. All these are part of the process of linearization of the topological or differentiable structures. The language of vector bundles allows us to make sense of the linearization process.

We are interested in the following types of problems

1. Consider the circle S^1 as a subspace of $\mathbb{R}^2$. Can we have a continuous map from $f : S^1 \to \mathbb{R}^2$, such that $\forall x \in S^1$,
   (a) $f(x) \perp x$.
   (b) $f(x) \neq 0$.
2. Consider the unit 2-sphere S^2 as a subspace of $\mathbb{R}^3$. Can we have a continuous map from $f : S^2 \to \mathbb{R}^3$, such that $\forall x \in S^2$,
   (a) $f(x) \perp x$.
   (b) $f(x) \neq 0$.
3. In particular given any differentiable manifold X, we want to associate for each $x \in X$ an element $v_x \in T_x$ continuously. Here T_x is the tangent space at x.

We will see that the answers to these questions really depend on fundamental topological properties of the space.

6.2 Basic definitions and examples

For us F will be either $\mathbb{R}$ or $\mathbb{C}$.

Definition 6.2.1. Let B be a topological space. An n-dimensional F- vector bundle over a topological space B is a continuous map $p : E \to B$ together with a F-vector space structure on $p^{-1}(b)$ for each $b \in B$, such that B can be covered by open sets U_α for each of which there exists a homeomorphism $h_\alpha : p^{-1}(U_\alpha) \to U_\alpha \times F^n$ mapping $p^{-1}(b)$ to $\{b\} \times \mathbb{F}^n$ by linear isomorphisms for each $b \in U_\alpha$.

h_α's are called local trivializations of the vector bundle. We will state results for $\mathbb{R}$ vector bundles but the statements are also true for $\mathbb{C}$-vector bundles .

6.2.1 Examples

1. Let X be any topological space. Consider the topological space $X \times \mathbb{R}^n$. The projection $p : X \times \mathbb{R}^n \to X$ is the first projection. This is called the trivial vector bundle of rank n.
2. Let $E = [0, 1] \times \mathbb{R}/\sim$, where $(a, b) \sim (c, d)$ iff $b = -d$ and $a = 0, c = 1$ or $a = 1, c = 0$.The projection map $p : E \to S^1$ is given by $[x, y] \mapsto e^{2\pi.i.x}$, here $[x, y]$ is an equivalence class. Let $U = S^1 \setminus \{1\}$. Then $p^{-1}(U) = U \times$

$\mathbb{R}/\sim \cong U \times \mathbb{R}$. Let $V = S^1 \backslash \{-1\}$. Then $p^{-1}(V) = [0, 1/2) \cup (1/2, 1] \times \mathbb{R}/\sim$. The trivialization$\phi : p^{-1}(V) \to V \times \mathbb{R}$ is given by $\phi([x, y]) = (x - 1/2, y)$ if $x > 1/2$, else $\phi([x, y]) = (x + 1/2, y)$.

3. Let S^n be the n-dimensional sphere.

$$TS^n := \left\{(x, v) \in S^n \times \mathbb{R}^{n+1} | x \perp v\right\},$$

and consider TS^n equipped with subspace topology (subspace of $S^n \times \mathbb{R}^{n+1}$). There is a projection map $p : TS^n \to S^n$ given by $(x, v) \mapsto x)$. Note that $p^{-1}(x) \cong \mathbb{R}^n$ and $p^{-1}(x)$ is the tangent space of at x. For $x \in S^n$, let U_x be the open hemisphere containing x and bounded by $p^{-1}(x)$. Consider the function $h_x : p^{-1}(U_x) \to U_x \times p^{-1}(x) \cong U_x \times \mathbb{R}^n$, given by $h_x(y, v) = (y, \pi_x(v))$, where π_x is the orthogonal projection to $p^{-1}(x)$. Then it is clear that h_x's are homeomorphism for every $x \in S^n$.

4. Again, let S^n be the n-dimensional sphere.

$$NS^n = \left\{(x, v) \in S^n \times \mathbb{R}^{n+1} | v = tx, t \in \mathbb{R}\right\}.$$

There is a projection $p : NS^n \to S^n$ given by $(x, v) \mapsto x$. Note that $p^{-1}(x) \cong \mathbb{R}$. For $x \in S^n$, let U_x be the open hemisphere containing x and bounded by the hyperplane passing through the origin which is orthogonal to x. Let $h_x : p^{-1}(U_x) \to U_x \times p^{-1}(x) \cong U_x \times \mathbb{R}$ be the map given by $h_x(x, v) = (x, \pi_x(v))$, where $\pi_x(v)$ is the projection of v to the line $p^{-1}(x)$. Then h_x's are homeomorphism for all $x \in S^n$.

5. Let $\mathbb{R}P^n$ be the space of lines through the origin in $\mathbb{R}^{n+1}$. We can think of $\mathbb{R}P^n$ as $S^n/(x \sim -x)$. The canonical line bundle is given by $E := \left\{(l, v) \in \mathbb{R}P^n \times \mathbb{R}^{n+1} | v \in l\right\}$. The projection map $p : E \to \mathbb{R}P^n$ is given by $(l, v) \mapsto l$. Note that $p^{-1}(l) = l \cong \mathbb{R}$. For $l \in \mathbb{R}P^n$, let $U_l \subset \mathbb{R}P^n$ be the open set of all lines in $\mathbb{R}P^n$ which are not orthogonal to l. Then the functions $h_l : p^{-1}(U_l) \to U_l \times p^{-1}l \cong U_l \times \mathbb{R}$, given by $h_l(l', v) = (l', \pi_l(v))$ is a homeomorphism. Here $\pi_l(v)$ is the projection of v to the line l. This is called the canonical line bundle.

6. Again for consider the topological space

$$E^\perp := \left\{(l, v) \in \mathbb{R}P^n \times \mathbb{R}^{n+1} | v \perp l\right\}.$$

There is a projection map $p : E^\perp \to \mathbb{R}P^n$ given by $(l, v) \mapsto l$. Note that $p^{-1}(l) \cong \mathbb{R}^n$. For each $l \in \mathbb{R}P^n$ consider the same open set U_l as in the previous example. The functions $h_l : p^{-1}(U_l) \to U_l \times p^{-1}l \cong U_l \times \mathbb{R}^n$, given by $h_l(l', v) = (l', \pi_l(v))$ is a homeomorphism. Here $\pi_l(v)$ is the projection of v to the line l. Here $\pi_l(v)$ is the orthogonal projection of v to $p^{-1}(l)$.

7. Let $\mathbb{R}P^\infty = \bigcup_\infty \mathbb{R}P^n$. The union is taken under the inclusions $\mathbb{R}P^n \subset \mathbb{R}P^{n+1}$. The topology on $\mathbb{R}P^\infty$ is defined by declaring a subset U to be open in $\mathbb{R}P^\infty$ iff $U \cap \mathbb{R}P^n$ is open for all n. The inclusions $\mathbb{R}P^n \subset \mathbb{R}P^{n+1}$, induce inclusions of canonical line bundles and after taking union of these canonical line bundle we get a line bundle over $\mathbb{R}P^\infty$.

8. For n and k positive integers such that $n \leq k$, consider the following sets

$$\begin{aligned} G_n(\mathbb{R}^k) &:= \left\{V \subset \mathbb{R}^k | dim_{\mathbb{R}}(V) = n\right\}, \\ V_n(\mathbb{R}^k) &:= \left\{(v_1, v_2, \ldots, v_n) | v_i \in \mathbb{R}^k, \text{orthonormal}\right\}. \end{aligned}$$

Note that

$$V_n(\mathbb{R}^k) \subset \underbrace{S^{k-1} \times \ldots S^{k-1}}_{n}.$$

So we can give the subspace topology to $V_n(\mathbb{R}^k)$. There is a surjective map $V_n(\mathbb{R}^k) \to G_n(\mathbb{R}^k)$ given by $(v_1, \ldots v_n) \mapsto Span(v_1, \ldots, v_n)$. Therefore we can give the quotient topology to $G_n(\mathbb{R}^k)$. Note that $G_1(\mathbb{R}^k) = \mathbb{R}P^{k-1}$. Since $G_n(\mathbb{R}^k) \subset G_n(\mathbb{R}^{k+1})$, we can take the union $G_n(\mathbb{R}^\infty) := \bigcup_k G_n(\mathbb{R}^k)$. We take all the sets $U \subset G_n(\mathbb{R}^\infty)$, such that $U \cap G_n(\mathbb{R}^k)$ is open for all k, to be the open sets. This defines a topology on $G_n(\mathbb{R}^\infty)$. Let

$$E_n(\mathbb{R}^k) = \left\{(l, v) \in G_n(\mathbb{R}^k) \times \mathbb{R}^k | v \in l\right\}.$$

This is a topological space with subspace topology coming from $G_n(\mathbb{R}^k) \times \mathbb{R}^k$. There is a projection map $p : E_n(\mathbb{R}^k) \to G_n(\mathbb{R}^k)$ given by $(l, v) \mapsto l$. For each $l \in G_n(\mathbb{R}^k)$ let $\pi_l : \mathbb{R}^k \to l$ be the orthogonal projection to l. The set

$$U_l := \left\{l' \in G_n(\mathbb{R}^k) | dim(\pi_l(l')) = n\right\}$$

is open. Then $h_l : p^{-1}(U_l) \to U_l \times l \cong U_l \times \mathbb{R}^n$ given by $(l', v) \mapsto (l', \pi_l(v))$ gives a homeomorphism. Hence $p : E_n(\mathbb{R}^k) \to G_n(\mathbb{R}^k)$ is a vector bundle of rank n. Let $E_n(\mathbb{R}^\infty) = \bigcup_k E_n(\mathbb{R}^k)$, where the union is taken over the inclusions $E_n(\mathbb{R}^k) \subset E_n(\mathbb{R}^{k+1})$. Then $p : E_n(\mathbb{R}^\infty) \to G_n(\mathbb{R}^\infty)$ is a vector bundle of rank n.

9. Let M be a smooth manifold and $p \in M$. Let

$$C(M, p) = \left\{\gamma : (-\epsilon, \epsilon) \to M | \gamma(0) = p\right\}.$$

Define $T_pM := C(M, p)/\sim$ where $\gamma \sim \sigma$ if there exists a chart $h : U \to \mathbb{R}^n$ such that $p \in U$ and $(h \circ \gamma)'(0) = (h \circ \sigma)'(0)$. Let

$$TM := \left\{(p, v) \in M \times \coprod_{p \in M} T_pM | v \in T_pM\right\}.$$

There is a natural projection map $\pi : TM \to M$. Give TM the coarsest topology such that π is continuous. Let $U \subset M$ be a open subset such that $h : U \cong \mathbb{R}^n$, and let $x^i : \mathbb{R}^n \to \mathbb{R}$ be the i-th projection maps. Then $\pi^{-1}(U) = TU$ and $TU \to \mathbb{R}^n \times \mathbb{R}^n$ given by $(p, v = \sum_{i=1}^n w_i \frac{\partial}{\partial x_i}) \mapsto (p, \{w_i(p)\})$ gives a homeomorphism. The vector bundle TM is called the tangent bundle of M. It is a vector bundle of rank $dim(M)$.

6.2.2 Map of vector bundles

Definition 6.2.2. Let $p_1 : E \to M$ and $p_2 : F \to M$ be vector bundles over M. A map of vector bundles from E to F is a continuous map $f : E \to F$ such that $p_2 \circ f = p_1$, and for every $m \in M$, the induced map $f : p_1^{-1}(m) \to p_2^{-1}(m)$ is a linear transformation. A map of vector bundles is an isomorphism if it has both left and right inverses.

Lemma 6.2.3. *Let $p_1 : E \to M$ and $p_2 : F \to M$ be vector bundles over M. A map f between vector bundles E to F is an isomorphism if and only if $f : p_1^{-1}(m) \to p_2^{-1}(m)$ is an isomorphism for all $m \in M$.*

Proof. The only if part is clear. Suppose $f : p_1^{-1}(m) \to p_2^{-1}(m)$ is an isomorphism for all $m \in M$. It is clear that f is a bijection. We want to show that f^{-1} is continuous. Continuity is a local property so we can assume that $U \subset M$ is open and we have homeomorphisms $h_i : p_i^{-1}(U) \to U \times \mathbb{R}^n$ for $i \in \{1, 2\}$. Then $h_2 \circ f \circ h_1^{-1} : U \times \mathbb{R}^n \to U \times \mathbb{R}^n$ is given by $(x, v) \mapsto (x, g_x(v))$, where $g_x \in GL_n(\mathbb{R})$. The entries of the matrix g_x depend on x continuously. Hence the entries of the matrix $g_x^{-1} = \mathrm{Ad}(g_x)/\det(g_x)$ depend continuously on x. So $h_1 \circ f^{-1} \circ h_2^{-1}$ is continuous as it is given by $h_1 \circ f^{-1} \circ h_2^{-1}(x, v) = (x, g_x^{-1}(v))$. Therefore f^{-1} is continuous. □

6.2.3 Sections of a vector bundle

Definition 6.2.4. A section of a vector bundle $p : E \to B$ is a continuous map $s : B \to E$ such that $p \circ s = id_B$.

1. Since every vector space V has a distinguished element $0 \in V$ (the zero element of the vector space), every vector bundle $p : E \to B$ has a canonical section s_0, called the zero section. The map $s_0 : B \to E$ is defined by $s_0(b) = 0_{p^{-1}(b)}$ for all $b \in B$.
2. Let $p_1 : E_1 \to B$ and $p_2 : E_2 \to B$ be vector bundles, then a vector bundle isomorphism $h : E_1 \to E_2$ takes the zero section to the zero section, i.e., $f \circ s_0 : B \to E_2$ is again $s_0 : B \to E_2$. Hence h takes $E_1 \setminus s_0(B)$ to $E_2 \setminus s_0(B)$. So the Möbius bundle over S^1 is not isomorphic to the trivial bundle $S^1 \times \mathbb{R}$. Indeed, if we they were isomorphic the complement of the zero section of the Möbius bundle which is connected will be homeomorphic to the complement of the zero section of the trivial bundle which is not connected.

Lemma 6.2.5. *An n-dimensional vector bundle $p : E \to B$ is trivial if and only if it has n-sections $s_1, s_2, \ldots, s_n$ such that for all $b \in B$, $s_1(b), \ldots, s_n(b)$ are linearly independent in $p^{-1}(b)$.*

Proof. Let $(e_1, \ldots, e_n)$ be the standard basis of $\mathbb{R}^n$. If $E = B \times \mathbb{R}^n$ and $p : E \to B$ the first projection then $s_i(b) = (b, e_i)$ are continuous sections of p and $s_i(b)$'s form a basis of $p^{-1}(b)$. Conversely suppose $p : E \to B$ has n-sections $s_1, s_2, \ldots, s_n$ such that for all $b \in B$, $s_1(b), \ldots, s_n(b)$ are linearly independent

in $p^{-1}(b)$. Define $h : B \times \mathbb{R}^n \to E$ by $h(b, t_1, t_2, \ldots, t_n) = \sum t_i.s_i(b)$. It is clear that h is continuous and a map of vector bundles. Now apply 6.2.3 to show that h is an isomorphism of vector bundles.

□

Definition 6.2.6. Let $p : E \to B$ a vector bundle. A section of s of p is called non vanishing if $s(B) \cap s_0(B) = \emptyset$.

6.2.4 Examples

1. The normal bundle NS^n is isomorphic to $S^n \times \mathbb{R}$. The isomorphism is given by $(x, tx) \mapsto (x, t)$.

2. The tangent bundle of the sphere is trivial, i.e., $TS^1 \cong S^1 \times \mathbb{R}$. Indeed, we have to construct a non vanishing section $s : S^1 \to TS^1$ (lemma 6.2.5). Define s as $s(x_1, x_2) = ((x_1, x_2), (-x_2, x_1))$ for all $(x_1, x_2) \in S^1$.

3. Any section of the Möbius bundle has to intersect the zero section hence the Möbius bundle is not trivial.

4. The vector space $\mathbb{R}^4$ is also an algebra. The multiplication is given as follows: We have orthonormal basis $1, i, j, k$ such that,

$$i^2 = j^2 = k^2 = -1;$$

and

$$ij = k, jk = i, ki = j, ji = -k, kj = -i, ik = -j.$$

Writing $z_1 := (x_1, x_2, x_3, x_4), z_2 := (y_1, y_2, y_3, y_4) \in \mathbb{R}^4$ as $z_1 = x_1 + ix_2 + jx_3 + kx_4$ and $z_2 := y_1 + iy_2 + jy_3 + ky_4$, we can define the multiplication $z_1.z_2$ using the multiplication rule above for i, j, k and forcing distributivity of multiplication over addition. Check that if $z = x_1 + ix_2 + jx_3 + kx_4$ and $\bar{z} = x_1 - ix_2 - jx_3 - kx_4$ then $z.\bar{z} = |z|^2$ and hence $|z.w| = |z|.|w|$. Let $s_\beta : S^3 \to TS^3$ be given by $s_\beta(z) = (z, \beta.z)$ where $\beta \in \{i, j, k\}$ for all $z \in S^3 \subset \mathbb{R}^4$. These three sections are linearly independent in $p^{-1}(z)$ for every $z \in S^3$, where $p : TS^3 \to S^3$. Hence by lemma 6.2.5 TS^3 is a trivial bundle.

5. The canonical line bundle over $\mathbb{R}P^1$ is isomorphic to the möbius bundle over S^1. Here we identify $\mathbb{R}P^1$ with $[0, \pi]/(0 \sim \pi)$. The tautological line bundle over $\mathbb{R}P^1$ can be described as tuples (θ, x) where $\theta \in [0, \pi]$ and $x \in l$ where l is the line passing through origin that makes θ angle with x axis, then identify $(0, x)$ with $(\pi, -x)$. The möbius bundle over S^1 is $[0, \pi] \times \mathbb{R}/((0, x) \sim (\pi, -x))$.

6.3 Operations

Let $p : E \to B$ a vector bundle. So there exists an open covering $\{U_\alpha\}$ of B and $h_\alpha : p^{-1}(U_\alpha) \to U_\alpha \times \mathbb{R}^n$ homeomorphism such that the induced map

$$h_\beta \circ h_\alpha^{-1} : U_\alpha \cap U_\beta \times \mathbb{R}^n \to U_\alpha \cap U_\beta \times \mathbb{R}^n$$

is given by $(x, v) \mapsto (x, g_x(v))$ where $g_x \in GL_n(\mathbb{R})$. This gives continuous maps $\rho_{\beta\alpha} : U_\alpha \cap U_\beta \to GL_n(\mathbb{R})$ such that $\rho_{\gamma\beta} \circ \rho_{\beta\alpha} = \rho_{\gamma\alpha}$. These $\rho_{\beta\alpha}$'s are called transition function.

Conversely, given a topological space B, open covering $\{U_\alpha\}$ of B and continuous maps $\rho_{\beta\alpha} : U_\alpha \cap U_\beta \to GL_n(\mathbb{R})$ such that $\rho_{\gamma\beta} \circ \rho_{\beta\alpha} = \rho_{\gamma\alpha}$, we construct a new vector bundle E_ρ as follows : $E_\rho := \coprod(U_\alpha \times \mathbb{R}^n)/\sim$, where $(x, v) \sim (x, \rho_{\beta\alpha}(x)v)$ for $x \in U_\alpha \cap U_\beta$. So elements in E_ρ is given by equivalence classes $[x, v]$. The topology is quotient topology. There is a projection map $p : E_\rho \to X$, sending $[x, v] \to x$, since the equivalence relation does not do anything to the first factor so it is a well defined continuous map. It is easy to check that $p^{-1}(U_\alpha) \cong U_\alpha \times \mathbb{R}^n$ and hence E_ρ is a vector bundle.

Given any vector bundle $p : E \to B$ we can use the functions $\rho_{\alpha\beta}$ of E to construct E_ρ. We can construct a canonical map of vector bundle $E_\rho \to E$, given by $h := \coprod h_\alpha^{-1} : \coprod U_\alpha \times \mathbb{R}^n \to E$ and then see that h takes equivalent points to a single point. It is is easy to verify using lemma 6.2.3, that h is an isomorphism. So to give a vector bundle $E \to B$ is equivalent to give open covering $\{U_\alpha\}$ of B and continuous maps $\rho_{\beta\alpha} : U_\alpha \cap U_\beta \to GL_n(\mathbb{R})$ such that $\rho_{\gamma\beta} \circ \rho_{\beta\alpha} = \rho_{\gamma\alpha}$.

6.3.1 Direct Sum

Let $p_1 : E_1 \to B$ and $p_2 : E_2 \to B$ be vector bundles. We can choose an open cover $\{U_\alpha\}$ such that both E_1 and E_2 are trivial over each U_α such that for E_1 the transition functions are $\rho^1_{\beta\alpha}$ and for E_2 the transition functions are $\rho^2_{\beta\alpha}$. Then we take $\rho_{\beta\alpha} = \rho^1_{\beta\alpha} \oplus \rho^2_{\beta\alpha}$. Define $E_1 \oplus E_2 := E_\rho$.

Let $E' := \{(x_1, x_2) \in E_1 \times E_2 | p_1(x_1) = p_2(x_2)\}$, equip this space with subspace topology coming from $E_1 \times E_2$ and let $p : E' \to B$ be given by $p(x_1, x_2) = p_1(x_1) = p_2(x_2)$. It is easy to verify that $p : E' \to B$ is a vector bundle and there exists a vector bundle isomorphism $h : E' \to E_1 \oplus E_2$.

Example 6.3.1. 1. Direct sum of trivial bundles is again a trivial bundle.

2. A point in $TS^n \oplus NS^n$ is given by (x, v, tx) where $x \in S^n$, $v \perp x$ and $t \in \mathbb{R}$. The map $TS^n \oplus NS^n \to S^n \times \mathbb{R}^{n+1}$ given by $(x, v, tx) \mapsto (x, v + tx)$ gives an isomorphism $TS^n \oplus NS^n \cong S^n \times \mathbb{R}^{n+1}$.

3. Let $E \to \mathbb{R}P^n$ be the canonical line bundle and let $E^\perp$ be the orthogonal bundle. Then a point of $E \oplus E^\perp$ is given by (l, v, w) such that $l \in \mathbb{R}P^n, v \in l, w \perp l$. The map $E \oplus E^\perp \to \mathbb{R}P^n \times \mathbb{R}^{n+1}$ given by $(l, v, w) \mapsto (l, v + w)$ gives an isomorphism $E \oplus E^\perp \cong \mathbb{R}P^n \times \mathbb{R}^{n+1}$.

4. If $n = 1$, we have $E \cong E^{\perp}$. Indeed $(l, v) \mapsto (l, \rho(v))$, where $v \in l$ and ρ is rotation by $\pi/2$, gives the required isomorphism. So $E \oplus E \cong S^1 \times \mathbb{R}^2$, or sum of two copies of möbius bundle is trivial.

6.3.2 Inner Product

Definition 6.3.2. Let $p : E \to B$ be a vector bundle. An inner product on E is a map $<,>: E \oplus E \to \mathbb{R}$, which restricts to a positive definite symmetric bilinear form in each fiber.

A space X is paracompact if it is Hausdorff and for every open cover $\{U_\alpha\}$ of X, there exists a collection of continuous functions $\phi_\beta : X \to [0, 1]$ such that ϕ_β is zero outside U_α for some α, each $x \in X$ has a neighborhood in which only finitely many ϕ_β's are nonzero, and $\sum_\beta \phi_\beta = 1$. The collection $\{\phi_\beta\}$ is called a partition of unity subordinate to the open cover $\{U_\alpha\}$. Every compact Hausdorff space is paracompact.

Proposition 6.3.3. *Let B be a paracompact space. Then for any vector bundle $p : E \to B$, an inner product exists on E.*

Proof. Let $h_\alpha : p^{-1}(U_\alpha) \to U_\alpha \times \mathbb{R}^n$ be the local trivializations of $p : E \to B$. Then $<,>_{U_\alpha}: p^{-1}(U_\alpha) \oplus p^{-1}(U_\alpha) \to \mathbb{R}$ is given by $(x_1, x_2) \mapsto < v_1, v_2 >$. Here $(p(x_i), v_i) = h_\alpha(x_i)$ and $<,>$ is the standard norm of $\mathbb{R}^n$. For any point $(x_1, x_2) \in E \oplus E$, let $< x_1, x_2 >:= \sum_\beta \phi_\beta(p(x_1)) < x_1, x_2 >_{U_\beta}$, where $\{\phi_\beta\}$ is a partition of unity subordinate to the open cover $\{U_\alpha\}$ such that ϕ_β is zero outside U_β. □

Similarly, if $p : E \to B$ was a complex vector bundle then one can construct a Hermitian inner product.

Proposition 6.3.4. *Let B be a compact Hausdorff space. For each vector bundle $E \to B$, there exists a vector bundle $E' \to B$ such that $E \oplus E'$ is the trivial bundle.*

Ref. 1, Proposition 1.4. □

6.3.3 Tensor Product

Let $p_i : E_i \to B$ be vector bundles $i \in \{1, 2\}$. We can choose an open cover $\{U_\alpha\}$ such that both E_1 and E_2 are trivial over each U_α such that for E_1 the transition functions are $\rho^1_{\beta\alpha}$ and for E_2 the transition functions are $\rho^2_{\beta\alpha}$. Then we take $\rho^{\otimes}_{\beta\alpha} = \rho^1_{\beta\alpha} \otimes \rho^2_{\beta\alpha}$. Define $E_1 \otimes E_2 := E_{\rho^\otimes}$. Here, by $\rho^1_{\beta\alpha} \otimes \rho^2_{\beta\alpha}$ we mean for each $x \in U_\alpha \cap U_\beta$ we take the tensor product of the matrices $\rho^1_{\beta\alpha}(x)$ and $\rho^2_{\beta\alpha}(x)$.

Definition 6.3.5. Let $p : E \to B$ a vector bundle. The rank of the vector bundle is the dimension of the vector space $p^{-1}(b)$ for any $b \in B$. A rank 1 vector bundle is called a line bundle.

Lemma 6.3.6. *For a topological space B, let $Vect^1(B)$ be the set of isomorphism classes of line bundles. Then $Vect^1(B)$ is a group, where the group operation is given by $\otimes$. Moreover if B is paracompact, then every element of $Vect^1(B)$ has order 2.*

Proof. If $p_i : L_i \to B$ be line bundles for $i \in \{1, 2\}$, then $L_1 \otimes L_2$ is again a line bundle. Suppose $p : L \to B$ a line bundle then the transition functions $\rho_{\beta\alpha} : U_\alpha \cap U_\beta \to GL_1(\mathbb{R})$ takes a value in an abelian group, so the continuous functions $\rho_{\beta\alpha}^{-1}$ satisfies cocycle condition. Hence for the line bundle $E_{\rho^{-1}}$, we have $E \otimes E_{\rho^{-1}} \cong B \times \mathbb{R}$ as the transition functions of the left hand side is $\rho_{\beta\alpha} \otimes \rho_{\beta\alpha}^{-1} = 1$. If B is paracompact, then any vector bundle $E \to B$ has an inner product, so we can choose the local trivializations to be isometries in each fiber. So for a line bundle $p : L \to B$, B paracompact, we can choose the transition functions $\rho_{\beta\alpha}$ such that it takes values in the subgroup $\{1, -1\} \subset GL_1(\mathbb{R})$. Then the transition function of $L \otimes L$ takes the value 1. Hence $L \otimes L \cong B \times \mathbb{R}$. □

The last part of the proposition is not true for complex line bundles.

6.3.4 Dual

Let $p : E \to B$ be a complex vector bundle. Let $\rho_{\beta\alpha}$ be the transition functions. Let $\bar{E}$ be the vector bundle whose transition functions are complex conjugates $\rho_{\bar{\beta}\alpha}$ of $\rho_{\beta\alpha}$. So as a set $\bar{E}$ is the same as E, but if v is in any fiber of $\bar{E}$ and $\lambda \in \mathbb{C}$, $\lambda.(v) = \bar{\lambda}.v$. Here the left hand side is the action of $\mathbb{C}$ in $\bar{E}$ and the right hand side is the action of $\mathbb{C}$ on E. Equivalently $\bar{E}$ as a space is the same as E but the local trivialization $\bar{h_\alpha} : p^{-1}(U_\alpha) \to U_\alpha \times \mathbb{C}^n$ is the composition $(id \times \bar{()}) \circ h_\alpha$, where h_α is the local trivialization of E, and $\bar{()} : \mathbb{C}^n \to \mathbb{C}^n$ is the coordinate wise conjugation. Now if B is paracompact and $L \to B$ is a complex line bundle then presence of Hermitian matrices allows us to choose transition functions $\rho_{\beta\alpha} : U_\alpha \cap U_\beta \to GL_1(\mathbb{C}) = \mathbb{C} \setminus \{0\}$ to actually take values in the subgroup $S^1 \in \mathbb{C} \setminus \{0\}$. Since the transition functions of $\bar{L}$ are $\rho_{\bar{\beta}\alpha}$, hence $L \otimes \bar{L} \cong B \times \mathbb{C}$, as $\rho_{\beta\alpha}.\rho_{\bar{\beta}\alpha} = 1$.

6.3.5 Pullback

Let $p : E \to B$ be a vector bundle, with transition functions $\rho_{\beta\alpha}$. Let $f : A \to B$ a continuous map. Then f^*E, the pullback of E along f is the vector bundle whose transition functions are $\rho_{\beta\alpha} \circ f|_{f^{-1}(U_\alpha \cap U_\beta)}$. Suppose E' is the set $E' := \{(a, v) \in A \times E | f(a) = p(v)\}$. This set can be equipped with the subspace topology coming from $A \times E$. Then E' is a vector bundle over A and $E' \cong f^*E$.

Example 6.3.7.

1. Pullback of a trivial bundle is trivial bundle.
2. Let $p : E \to S^1$ be the Möbius bundle and let $f : S^1 \to S^1$ be the function given by $z \mapsto z^2$, then $f^*E \cong S^1 \times \mathbb{R}$. Indeed, there is always a map from $g : S^1 \to E$ such that $p \circ g = f$ and $g \cap s_0$ is empty. Then this gives a non zero section $g' : S^1 \to f^*E$.

We also have the following isomorphisms:

1. For $g : A \to B$, $f : B \to C$ and a vector bundle $E \to C$, we have $(f \circ g)^* E \cong g^*(f^* E)$.
2. $id^* E \cong E$.
3. $f^*(E_1 \oplus E_2) \cong f^* E_1 \oplus f^* E_2$.
4. $f^*(E_1 \otimes E_2) \cong f^*(E_1) \otimes f^*(E_2)$.

Let $f_0, f_1 : A \to B$ be continuous maps such that $\exists$ a continuous map $H : A \times [0,1] \times A \to B$, such that $F(0, A) = f_0$ and $F(1, A) = f_1$, then we say that f_0 is homotopic to f_1. Two maps from $A \to B$ being homotopic gives an equivalence relation on the set of continuous maps $C(A, B)$. We denote by $[A, B]$ the set of equivalence classes modulo this equivalence relation.

Theorem 6.3.8. *Let $p : E \to B$ a vector bundle and $f_0, f_1 : A \to B$ be two homotopic maps. Then $f_0^* E \cong f_1^* E$ if A is paracompact.*

Ref. 1, Theorem 1.6. □

Let A be a space and let $Vect^n(A)$ be the set of isomorphism classes of rank n vector bundles over A.

Corollary 6.3.9. *If $f : A \to B$ is a homotopy equivalence between paracompact spaces then $f^* : Vect^n(B) \to Vect^n(A)$ is a bijection. In particular, every vector bundle over a paracompact contractible space is trivial.*

Theorem 6.3.10. *For a paracompact topological space X, the natural map of sets $[X, G_n(\mathbb{R}^\infty)] \to Vect^n(X)$ given by $(f : X \to G_n(\mathbb{R}^\infty)) \mapsto f^*(E_n(\mathbb{R}^\infty))$ is bijective.*

Ref.1. [Theorem 1.16] □

Example 6.3.11. Let $p : TS^n \to S^n$ be the natural map. For $x \in S^n$ the fiber $p^{-1}(x)$ is an n-dimensional subspace of $\mathbb{R}^{n+1}$. Hence we have a map $f : S^n \to G_n(\mathbb{R}^\infty)$ given by $f(x) = p^{-1}(x)$. Then $f^*(E_n(\mathbb{R}^\infty)) \cong TS^n$.

6.4 Clutching function

The aim of this section is to understand the construction of vector bundles over the real spheres S^k. We will see that in this case there is a difference between complex and real vector bundles. For us F will be either $\mathbb{R}$ or $\mathbb{C}$. Let $E \to S^k$ be a F vector-bundle of rank n. The space S^k can be covered by two open discs D_+^k and D_-^k such that the intersection $D_+^k \cap D_-^k = S^{k-1} \times (-\epsilon, \epsilon)$. Now D_+^k and D_-^k are contractible hence E restricted to these open discs are isomorphic to trivial vector bundles. If we make a choice of these isomorphisms, we get local trivializations, and the transition function (it depends on the choice of the isomorphism) $S^{k-1} \times (-\epsilon, \epsilon) \to GL_n(F)$ gives a unique function (unique upto homotopy) $f : S^{k-1} \to GL_n(F)$.

Conversely, given any continuous map $f : S^{k-1} \to GL_n(F)$ (such functions are called clutching functions) we can extend it to $\tilde{f} : S^{k-1} \times (-\epsilon, \epsilon) \to GL_n(F)$ such that at each time t it is f. Then

$$E_f := \frac{D^k_+ \times F^n \coprod D^k_- \times F^n}{\sim}, \text{where } (x, v) \sim (x, \tilde{f}(x)(v)),$$

for $(x, v) \in D^k_+ \cap D^k_- \times F^n$. The evident projection map $E_f \to S^k$ gives a F-vector bundle of rank n. Suppose we have two maps $f, g : S^{k-1} \to GL_n(F)$ such that f and g are homotopic by a map $F : S^{k-1} \times I \to GL_n(F)$. Then using the same method as before we can construct a vector bundle $E_F \to S^k \times I$ such that $E_F|_{S^k \times 0} = E_f$ and $E_F|_{S^k \times 1} = E_g$ and hence $E_f \cong E_g$. This obviously gives a map $\Phi_F : [S^{k-1}, GL_n(F)] \to Vect^n_F(S^k)$. The following result is due to the fact that $GL_n(\mathbb{C})$ is path connected.

Proposition 6.4.1. *The map Φ_F is bijective for $F = \mathbb{C}$. It is not in general bijective for $F = \mathbb{R}$.*

In the case of $F = \mathbb{R}$, the sub Lie group $GL_{n,+}(\mathbb{R})$ of matrices of positive determinant is the connected component of the identity and any clutching function f whose image is in $GL_{n,+}(\mathbb{R})$ gives a vector bundle E_f which is oriented.

Example 6.4.2. 1. Every complex vector bundle over S^1 is trivial as $Vect^n_{\mathbb{C}}(S^1) = \pi_0(GL_n(\mathbb{C})) = *$.

2. In this example we are interested in vector bundles over $\mathbb{C}P^1$. Recall that $\mathbb{C}P^1$ is the quotient space of $\mathbb{C}^2 \setminus \{0\}$ under the equivalence relation $(z_0, z_1) \sim \lambda(z_0, z_1)$ for $\lambda \in \mathbb{C} \setminus \{0\}$. Writing the equivalence class of (z_0, z_1) as $[z_0, z_1]$, we get a homeomorphism $\mathbb{C}P^1 \to \mathbb{C} \cup \{\infty\} \cong S^2$ which is given by $[z_0, z_1] \mapsto z_0/z_1$ if $z_1 \neq 0$ and $[z_0, 0] \mapsto \infty$. Under this identification, the points in the disk D^2_0 inside the unit circle S^1 can be expressed uniquely in the form $[z, 1]$ with $|z| \leq 1$ and the points in the disk D^2_∞ outside S^1 can be uniquely written in the form $[1, z^{-1}]$ with $|z^{-1}| \leq 1$.

   Let $H \to \mathbb{C}P^1$ be the canonical line bundle. Recall that $H := \{(l, v) \in \mathbb{C}P^1 \times \mathbb{C}^2 | v \in l\}$. Over D^2_0 a section of the line bundle H is given by $[z, 1] \mapsto (z, 1)$ and over D^2_∞ it is given by $[1, z^{-1}] \mapsto (1, z^{-1})$. This gives trivializations of H over these disks. The common intersection of the two disks is S^1 and here we can go from one trivialization to another trivialization by multiplication by z. After identifications $D^2_0 = D^2_-$ and $D^2_\infty = D^2_+$ we get that the clutching function $f : S^1 \to GL_1(\mathbb{C}) = \mathbb{C} \setminus \{0\}$ of H is given by $f(z) = z$.

   One can also explicitly use the map

$$\mathbb{C}P^1 \to S^2 = \{(z, t) \in \mathbb{C} \times \mathbb{R} | |z|^2 + t^2 = 1\}$$

   which is given by

$$[z_0, z_1] \mapsto \frac{1}{|z_0|^2 + |z_1|^2}(2\bar{z_0}z_1, |z_0|^2 - |z_1|^2),$$

and the inverse is given by

$$(z,t) \mapsto [(1+t) : z].$$

Now we want to show that

$$H \oplus H \cong (H \otimes H) \oplus \mathbf{1}.$$

Here $\mathbf{1}$ is the trivial vector bundle of rank 1. For this it is enough to show that clutching functions of both sides are homotopic. As the clutching function of H is $f(z) = z$, therefore the clutching function of $H \oplus H$ is given by the function

$$z \in S^1 \mapsto \begin{bmatrix} z & 0 \\ 0 & z \end{bmatrix},$$

and the clutching function of $(H \otimes H) \oplus \mathbf{1}$ is given by

$$z \in S^1 \mapsto \begin{bmatrix} z^2 & 0 \\ 0 & 1 \end{bmatrix}.$$

Now, let α_t be the path in $GL_2(\mathbb{C})$ from the identity matrix to the matrix

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Then the function $F : S^1 \times I \to GL_n(\mathbb{C})$, given by

$$(z,t) \mapsto (f \oplus 1)(z).\alpha_t.(1 \oplus f)(z).\alpha_t,$$

gives the required homotopy. Here

$$(f \oplus 1)(z) = \begin{bmatrix} z & 0 \\ 0 & 1 \end{bmatrix},$$

and

$$(1 \oplus f)(z) = \begin{bmatrix} 1 & 0 \\ 0 & z \end{bmatrix}.$$

The last computation will be useful in the K-theory part.

6.5 Complex K-theory and Bott periodicity

K-theory is an invariant which captures the obstruction of a vector bundle to being trivial. In this section we will see the basic definition of K theory of complex and real vector bundles. Again let F denote either $\mathbb{R}$ or $\mathbb{C}$ and $1^n \to X$ denote the trivial F-vector bundle of rank n. Also $1^0 = X$ and 1 is the trivial bundle of rank 1.

In the collection of vector bundles on X we can define the following two equivalence relations.

Definition 6.5.1. 1. Let $E_1 \to X$ and $E_2 \to X$ be two F vector bundles on X. Set $E_1 \sim E_2$ if $E_1 \oplus 1^n \cong E_2 \oplus 1^m$ as F vector bundles for some $m, n \in \mathbb{N}$. Denote by $[E]$ the equivalence class of a vector bundle $E \to X$ under this equivalence relation.

2. We say that $E_1 \to X$ and $E_2 \to X$ are stably isomorphic (denoted by $E_1 \sim_s E_2$ if $E_1 \oplus 1^n \cong E_2 \oplus 1^n$ for some n. Denote by $[E]_s$ the equivalence class of a vector bundle $E \to X$ under this equivalence relation.

The operation $\oplus$ of vector bundles can be extended to the equivalence classes for $\sim$ or $\sim_s$ by the formula $[E] \oplus [E'] = [E \oplus E']$ (same for $\sim_s$). It is easy to verify that $\oplus$ on equivalence classes is well defined, commutative and associative. For every vector bundle $E \to X$, for X compact Hausdorff, we have a vector bundle $E' \to X$ such that $E \oplus E' \cong 1^n$ for some n. The next proposition is the formal consequence of this.

Proposition 6.5.2. *X is compact Hausdorff, then the set of $\sim$ equivalence classes of vector bundles on X forms an abelian group with respect to $\oplus$. For $F = \mathbb{R}$ we denote this group by $\widetilde{KO}(X)$ and for $F = \mathbb{C}$ we denote this group by $\widetilde{K(X)}$.*

Note that $[1^m] \oplus [E] = [E]$ and $[1^m] = [1^n] = [1^0]$ for the equivalence relation $\sim$. It is also obvious that $\sim_s$ does not give a group structure as the class $[1^0]_s = [1^n]_s$ implies $n = 0$ and $E \oplus E' \sim_s 1^0$ implies E and E' are isomorphic to 1^0. But we have cancellation for $\sim_s$. Suppose $E_1 \oplus E_2 \sim_s E_1 \oplus E_3$ and X is compact Hausdorff, then there exists E' such that $E_1 \oplus E' \cong 1^n$ for some n. Hence $E_1 \oplus E' \oplus E_2 \sim_s E_1 \oplus E' \oplus E_3$. This implies, there exists $N \geq n$ such that $1^N \oplus E_2 \cong 1^N \oplus E_3$ or $E_2 \sim_s E_3$. So the equivalence classes modulo $\sim_s$ forms a monoid ($\oplus$ being the composition of the monoid and $[1^0]_s$ being the class of the unit) such that cancellation holds.

This is similar to the case of $\mathbb{Z} \setminus \{0\}$ with multiplication of integers as monoid operation. The universal group $\mathbb{Q} \setminus \{0\}$ where we formally add the inverse of each element in $\mathbb{Z} \setminus \{0\}$ is given by the following construction

$$\mathbb{Q} \setminus \{0\} := \frac{\{(a,b) \in \mathbb{Z} \setminus \{0\} \times \mathbb{Z} \setminus \{0\}\}}{\{(a,b) \sim (c,d) | ad = bc\}}.$$

We will follow the same method to define K theory.

Definition 6.5.3. Let X be a compact Hausdorff space. Set $K(X)$ and $KO(X)$ as

$$\begin{aligned} K(X) &:= \frac{\{E - E' | E, E' \in Vect_{\mathbb{C}}(X)\}}{=}, \\ KO(X) &:= \frac{\{E - E' | E, E' \in Vect_{\mathbb{R}}(X)\}}{=}. \end{aligned}$$

Here $E_1 - E_1' = E_2 - E_2'$ iff $E_1 \oplus E_2' \sim_s E_2 \oplus E_1'$.

The relation $=$ being transitive requires cancellation theorem. Indeed, $E_1 \oplus E_2' \sim_s E_2 \oplus E_1'$ and $E_2 \oplus E_3' \sim_s E_2' \oplus E_3$, gives the equation $E_1 \oplus E_2 \oplus E_3' \sim_s E_2 \oplus E_1' \oplus E_3$, now cancelling out E_2 from both sides gives us $E_1 - E_1' = E_3 - E_3'$.

6.5.1 The abelian group structure

Note that $K(X)$ or $KO(X)$ is an abelian group where the addition is defined for the equivalence classes by the following rule

$$[E_1 - E_1'] + [E_2 - E_2'] := [E_1 \oplus E_2 - E_1' \oplus E_2'].$$

From this it is easy to verify that $[E-E] = [E'-E']$ and $[E_1 - E_1'] + [E-E] = [E_1 - E_1']$. Also inverse of $[E-E']$ is $[E'-E]$. Moreover every element $[E-E']$ is equal to $[E \oplus E'' - 1^n]$ for some n such that $E' \oplus E'' \cong 1^n$. If we have $[E-1^n] = [E'-1^m]$ then $E \oplus 1^m \sim_s E' \oplus 1^n$, hence $E \sim E'$. This way we get a natural group homomorphism $K(X) \to \widetilde{K(X)}$ (resp. $KO(X) \to \widetilde{K(X)}$) given by $[E-1^m] \mapsto [E]_\sim$. This homomorphism is clearly surjective. Kernel of this homomorphism are all those $[E-1^m]$ such that $E \sim 1^0$ or $E \sim_s 1^n$ for some n. But this implies $[E-1^m] = [1^n - 1^m]$. Therefore the kernel is isomorphic to $\mathbb{Z}$, where $[1^n - 1^m]$ is identified with $n-m$. Also there is a group homomorphism (even a ring homomorphism) $K(X) \to K(x)$ (resp. $KO(X) \to KO(x)$) for $x \in X$ given by restriction of vector bundles to x. This homomorphism gives isomorphism when restricted to the subgroup of the elements of type $[1^n - 1^m]$. Hence we get $K(X) \cong \widetilde{K(X)} \oplus \mathbb{Z}$ (resp. $KO(X) \cong \widetilde{KO(X)} \oplus \mathbb{Z}$). This isomorphism depends on the choice of the point $x \in X$. For every non zero vector bundle $E \to X$ there exists a non-zero class $[E - 1^0] \in K(X)$.

6.5.2 The ring structure and external product

There is a well defined multiplication on $K(X)$ (resp. $KO(X)$) given by the following formula.

$$[E_1 - E_1'][E_2 - E_2'] := [(E_1 \otimes E_2) \oplus (E_1' \otimes E_2') - (E_1 \otimes E_2') \oplus (E_1' \otimes E_2)].$$

It is easy to verify that the product structure is really well defined using cancellation law and we leave this verification to the reader. With this product structure $K(X)$ (resp. $KO(X)$) is a commutative unital ring with identity the class of $[1 - 1^0]$. For instance $[E - 1^0][E' - 1^0] = [E \otimes E' - 1^0]$ and $[E - 1^0][1^n - 1^0] = [E^n - 1^0]$. Here E^n means direct some of n-copies of E.

Let $f : X \to Y$ be a map, then $f^* : K(Y) \to K(X)$ is defined by the formula $f^*([E - E']) := [f^*E - f^*E']$. Upto isomorphism pullback preserves tensor product, Direct sum and trivial vector bundles. Also pullbacks of a vector bundle by two homotopic maps are isomorphic. These shows that the map f^* is a well defined ring homomorphism satisfying the following properties.

1. $(Id)^* = Id$.
2. $(f \circ g)^* = g^* \circ f^*$.
3. If f and g are homotopic then $f^* = g^*$.
4. For $x \in X$, the kernel of the ring homomorphism $K(X) \to K(x)$ is an ideal isomorphic to $\widetilde{K(X)}$ (same statement holds for $KO(X)$).

The external product $\mu : K(X) \otimes K(Y) \to K(X \times Y)$ is given by

$$\mu(a \otimes b) = pr_1^*(a).pr_2^*(b),$$

where $pr_1 : X \times Y \to X$ and $pr_2 : X \times Y \to Y$. The abelian group $K(X) \otimes K(Y)$ is a ring with multiplication given by

$$(a \otimes b).(c \otimes d) = (a.c \otimes b.d).$$

Therefore

$$\mu((a \otimes b).(c \otimes d)) = \mu(a.c \otimes b.d) = pr_1^*(a.c).pr_2^*(b.d) =$$

$$= pr_1^*(a).pr_2^*(b)pr_1^*(c).pr_2^*(d) = \mu(a \otimes b)\mu(c \otimes d).$$

This shows that μ is a ring homomorphism.

Now if we take $Y = S^2$, we get $\mu : K(X) \otimes K(S^2) \to K(X \times S^2)$. But $S^2 \cong \mathbb{C}P^1$ and the relation $(H \otimes H) \oplus 1 \cong H \oplus H$ gives the relation $(H-1)^2 = 0$ in $K(S^2)$. Here H is the canonical complex line bundle on $\mathbb{C}P^1$. So there exists a unique ring homomorphism $\alpha : \mathbb{Z}[t]/(t-1)^2 \to K(S^2)$ which maps t to the class of H in $K(S^2)$. Therefore we get by composition the following ring homomorphism

$$\mu' : K(X) \otimes \mathbb{Z}[t]/(t-1)^2 \xrightarrow{id \otimes \alpha} K(X) \otimes K(S^2) \xrightarrow{\mu} K(X \times S^2).$$

A weak formulation of Bott Periodicity is the following theorem.

Theorem 6.5.4. *The ring homomorphism $\mu' : K(X) \otimes \mathbb{Z}[t]/(t-1)^2 \to K(X \times S^2)$ is an isomorphism. Therefore $K(S^2) \cong \mathbb{Z}[t]/(t-1)^2$.*

Ref. 1. [Theorem 2.2]. □

There is a version of the previous theorem using the reduced K-theory, i.e., $\widetilde{K(X)}$. The interested reader can see [Ref. 1][Theorem 2.11] for the statement and the proof.

6.5.3 Non zero vector field on S^2

Now let us focus on the following question posed in the introduction.

Consider the unit 2-sphere S^2 as a subspace of $\mathbb{R}^3$. Can we have a continuous map from $f : S^2 \to \mathbb{R}^3$, such that $\forall x \in S^2$,

1. $f(x) \perp x$.
2. $f(x) \neq 0$.

Suppose there exists such f. Then it is a non-vanishing vector field on X. This means $f = v_1 : S^2 \to TS^2$ such that $v_1(x) \in T_xS^2$ is non zero for all $x \in S^2$. Now using the inner product on TS^2 we can construct a $v_2 : S^2 \to TS^2$ such that $v_1(x)$ is orthogonal to $v_2(x)$ for all $x \in S^2$. Hence, $v_1(x), v_2(x)$ forms

a basis of the tangent space T_xS^2 for all $x \in S^2$. Now using Gram-Schmidt process we can make $x, v_1(x), v_2(x)$ orthonormal for all $x \in S^2$. We may also assume that $e_1, v_1(e_1), v_2(e_1)$ are the standard basis e_1, e_2, e_3 of $\mathbb{R}^3$. This way we get a map $\alpha : S^2 \to SO(3)$ such that $\alpha(x)$ is the linear transformation sending the standard orthonormal basis to the orthonormal basis $x, v_1(x), v_2(x)$. Now α gives a map $\mu : S^2 \times S^2 \to S^2$ given by $(x, y) \mapsto \alpha(x)(y)$. Note that $\alpha(e_1)(y) = y$ and $\alpha(x)(e_1) = x$ by assumption. The structure $\mu : S^2 \to S^2 \times S^2$ thus obtained makes S^2 an H-space. Roughly, an H-space is a topological space which has multiplication and two sided identity but lacks associativity and inverse. For example S^1, S^3, S^7 are all H-spaces since $\mathbb{R}^2, \mathbb{R}^4, \mathbb{R}^8$ are all algebras with norms.

The map $\mu : S^2 \times S^2 \to S^2$ induces a map on K-theory $\mu^* : K(S^2) \to K(S^2 \times S^2)$. We know by Bott-periodicity $K(S^2) = \mathbb{Z}[\gamma]/(\gamma^2)$ and $K(S^2 \times S^2) = \mathbb{Z}[\alpha, \beta]/(\alpha^2, \beta^2)$. Hence we can identify μ^* with a ring homomorphism

$$\mu^* : \mathbb{Z}[\gamma]/(\gamma^2) \to \mathbb{Z}[\alpha, \beta]/(\alpha^2, \beta^2).$$

Every element of $\mathbb{Z}[\alpha, \beta]/(\alpha^2, \beta^2)$ can be written as $k + m_\alpha\alpha + m_\beta\beta + m\alpha.\beta$ for some $k, m_\alpha, m_\beta, m \in \mathbb{Z}$. Hence $\mu^*(\gamma) = k + m_\alpha\alpha + m_\beta\beta + m\alpha.\beta$. Let us show that $k = 0, m_\alpha = m_\beta = 1$.

Set $i_1 : S^2 \to S^2 \times S^2$ given by $x \mapsto (x, e_1)$ and $i_2 : S^2 \to S^2 \times S^2$ given by $x \mapsto (e_1, x)$. Since S^2 is an H-space with identity e_1, we have $\mu \circ i_j = id$ for $j \in \{1, 2\}$. This implies $(\mu \circ i_1)^*(\gamma) = \alpha$ and $(\mu \circ i_2)^*(\gamma) = \beta$. This implies $k = 0, m_\alpha = m_\beta = 1$ as $(\mu \circ i_j)^* = i_j^* \circ \mu^*$ and $(i_1)^*(\alpha) = \alpha$ and $(i_1)^*(\beta) = 0$ (similarly $(i_2)^*(\beta) = \beta$ and $(i_1)^*(\alpha) = 0$).

Now $\mu^*(\gamma) = \alpha + \beta + m\alpha.\beta$ implies $\mu^*(\gamma^2) = (\alpha + \beta + m\alpha.\beta)^2 = 2\alpha.\beta$. The element $2\alpha.\beta \neq 0 \in \mathbb{Z}[\alpha, \beta]/(\alpha^2, \beta^2)$. This gives us a contradiction as $\gamma^2 = 0$. In general using Bott Periodicity of K-theory one can prove the following stronger theorem.

Theorem 6.5.5. *There exist $n - 1$ vector fields on S^{n-1} which are linearly independent at each point (equivalently the tangent bundle is trivial) if and only if $n \in \{1, 2, 4, 8\}$.*

For a proof of this theorem, see Ref.1, Theorem 2.16.

References

[1] A. Hatcher, *Vector Bundles and K-Theory* (Cambridge University Press, New York, 2002).

II An introduction to the Chern-Weil Theory in vector bundles

6.6 Introduction

To understand the properties of a smooth manifold at an infinitesimal level, one studies the tangent space at each point on the manifold and also how the tangent space changes as the base-point moves on the manifold. One would, therefore, like to study the tangent bundle consisting of the tangent spaces at all points on the manifold. It turns out that the tangent bundle of a smooth manifold has a natural smooth structure. Generalizing the notion of the tangent bundle one constructs a vector bundle over a smooth manifold by associating a vector space of a fixed dimension (known as the *rank* of the vector bundle) at each point of the manifold in a way that given the smooth structure of the manifold, the set of all these vector spaces at all points of the manifold acquires a natural smooth structure. The simplest example of a vector bundle of rank n over a smooth manifold M is the product manifold $M \times \mathbb{R}^n$ where $n \in \mathbb{N}$. This is known as the trivial vector bundle of rank n over M. Any vector bundle of a given rank n over a smooth manifold M locally looks like a product $U \times \mathbb{R}^n$ where U is some open subset of M, but two different vector bundles of the same rank over the same manifold may differ in the process in which the local product structures are glued together to obtain a global smooth structure of that vector bundle. In order to classify the vector bundles (of a given rank) over a given manifold, one thus requires a global invariant which distinguishes two vector bundles depending on this gluing process. Characteristic classes are crucial global invariants which are used to understand how much *twisted* a vector bundle is, i.e., how far it is from being a trivial bundle.

6.7 Connection and curvature in a vector bundle

6.7.1 Connection in a smooth manifold and curvature

The idea of connection has been used in transporting some topological or geometric data related to a smooth manifold along a curve in the manifold in a nice and consistent manner. Depending on the context and what is to be transported, a connection is defined in an appropriate manner. Let's start with the simplest version of connection, known as an *affine connection* on a smooth manifold, which provides a standard technique for transporting tangent vectors from one point to another along a curve of the manifold. Thus it helps to compare the tangent spaces of two points in a manifold which are joined by a curve. An affine connection is also described in terms of a covariant derivative, which prescribes a way to take a derivative of a vector field with respect to another vector field on a manifold, i.e., to transport a vector field infinitesimally in a given direction.

Definition 6.7.1. Let M be a smooth manifold of dimension n. An affine connection on M is a $\mathbb{R}$-bilinear map

$$\begin{aligned} \nabla : \chi(M) \times \chi(M) &\to \chi(M) \\ (X, Y) &\mapsto \nabla_X Y, \quad X, Y \in \chi(M) \end{aligned}$$

satisfying the following properties

$$\begin{aligned} \nabla_{fX} Y &= f(\nabla_X Y) \text{ and} \\ \nabla_X (fY) &= X(f)Y + f(\nabla_X Y) \text{ for } \mathrm{f} \in \mathrm{C}^\infty(\mathrm{M}), \mathrm{X}, \mathrm{Y} \in \chi(\mathrm{M}) \end{aligned}$$

Then$\nabla_X Y$ is called the covariant derivative of Y with respect to X.

Considering the vector fields on a smooth manifold M as the sections of the tangent bundle TM, an affine connection on M can be re-interpreted as a way to translate the sections of the tangent bundle and to compare the fibers of TM at different points in M. One refers to an affine connection on a smooth manifold as a connection in its tangent bundle.

It can be checked that there are infinitely many connections on any smooth manifold. In particular, given any Riemannian metric g on M, there is a natural choice of an affine connection ∇^g on M which also satisfies the following additional properties

$$\begin{aligned} \nabla^g_X Y - \nabla^g_Y X &= [X, Y] \text{ and} \\ Z(g(X, Y)) &= g(\nabla^g_Z X, Y) + g(X, \nabla^g_Z Y) \text{ for } \mathrm{X}, \mathrm{Y}, \mathrm{Z} \in \chi(\mathrm{M}) \end{aligned}$$

∇^g is called the *Riemannian connection* of the metric g.

Example 6.7.2. The Riemannian connection ∇^{g_E} on the Euclidean space $(\mathbb{R}^n, g_E)$ equipped with the standard Euclidean metric g_E, is given by the directional derivative of vector fields, i.e., given vector fields X, Y on $\mathbb{R}^n$, $\nabla^{g_E}_X Y = X(Y)$. More explicitly, if $Y = (Y^1, ..., Y^n)$ with respect to the standard Euclidean (global) co-ordinates where $Y^i : \mathbb{R}^n \to \mathbb{R}$, are smooth functions for each $i = 1, ..., n$, then $\nabla^{g_E}_X Y = (X(Y^1), ..., X(Y^n))$ where at a point $p \in \mathbb{R}^n$, $X(Y^i)(p) = DY^i(p)(X(p))$ denotes the usual directional derivative of the function Y^i in the direction of the vector $X(p)$.

With this background, one can talk about translating a vector field along a curve on the manifold as follows. First of all, let us define vector fields and covariant derivative of vector fields along a curve on a manifold.

Definition 6.7.3. A vector field X along a smooth curve $c : [a, b] \to M$ is a smooth map $X : [a, b] \to TM$ such that $X(t) \in T_{c(t)}M$ for each $t \in [a, b]$.

The covariant derivative of a vector field $X(t)$ along a curve $c(t)$ is again a vector field along the same curve denoted by $\frac{DX}{dt}$, and obtained by taking the covariant derivative of X with respect to the tangent vector field $c'(t)$ of the curve c.

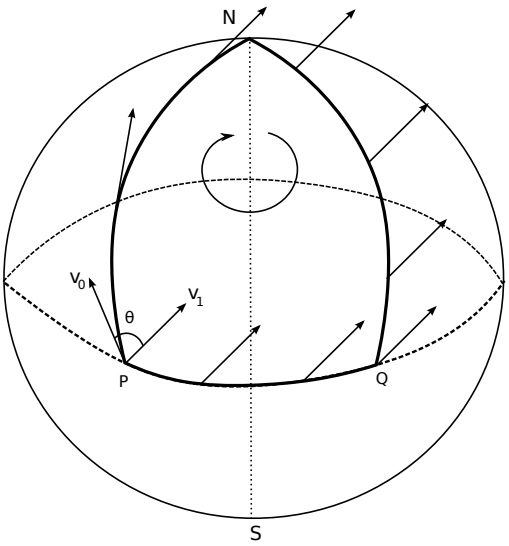


Figure 6.1: Parallel translation of the vector v_0 along the loop PNQP on S^2 using the Riemannian connection of the round metric

Definition 6.7.4. A vector field $X(t)$ along a curve $c(t)$ is said to be parallel if $\frac{DX}{dt} = 0$.

It can be checked that given a smooth curve $c : [a, b] \to M$ and any $v \in T_{c(t_0)}M$ for some $t_0 \in [a, b]$, there exists a unique parallel vector field V along c such that $V(t_0) = v$ (see Ref. 1for details). $V(t)$ is called the *parallel translate* of the vector $v = V(t_0)$ along the curve c. More precisely, we define

Definition 6.7.5. The parallel translation along a curve c from $c(t_0)$ to $c(t)$, is the linear map $P_t : T_{c(t_0)}M \to T_{c(t)}M$ which sends a vector $v \in T_{c(t_0)}M$ to the vector $V(t) \in T_{c(t)}M$ where V denotes the parallel vector field along c satisfying $V(t_0) = v$.

In the Euclidean space, parallel translation of a vector $v \in T_{c(t_0)}M$ along any curve c literally means translating that vector to $c(t)$ (i.e., it does not depend on the curve but on the points $c(t_0)$ and $c(t)$ together with the vector v) and parallel translation of a vector along a loop brings the vector back to itself. But this is not true in general. In particular, parallel translation of a vector along a loop does not usually bring the vector back to itself. Consider the example in the round sphere S^n described by Fig. 6.1, where the vector v_0 at the point P is parallel translated along the loop denoted by PNQP to the vector v_1 which makes an angle θ with the initial vector v_0. It is not difficult to check that the angle θ through which the initial vector is rotated measures the area inside the loop.

Given a Riemannian manifold (M, g) with its Riemannian connection ∇^g, one can describe the *curvature* of the manifold denoted by R, as the map

that assigns to every pair of vector fields $X, Y \in \chi(M)$ the linear operator $R(X,Y) : \chi(M) \to \chi(M)$ given by the following formula:

$$R(X,Y)Z = \nabla^g_X \nabla^g_Y Z - \nabla^g_Y \nabla^g_X Z - \nabla^g_{[X,Y]} Z, \quad X, Y, Z \in \chi(M)$$

A straight forward computation then implies the following

Lemma 6.7.6. *The curvature R of a Riemannian manifold (M, g) satisfies the following properties:*

1. $R(X,Y) = -R(Y,X)$,
2. $R(f_1X, f_2Y)(f_3Z) = f_1 f_2 f_3 R(X,Y)(Z)$, *and*
3. $R(X,Y)Z + R(Y,Z)X + R(Z,X)Y = 0$ *for* $X, Y, Z \in \chi(M)$, *and* $f_i \in C^\infty(M), i = 1, 2, 3$.

Proof. Left as an exercise. (Hint: Use the properties of Lie bracket and ∇^g.) □

Let $(U, (x_i))$ be a co-ordinate chart around a point p in the manifold M. If $X = \frac{\partial}{\partial x_i}$ and $Y = \frac{\partial}{\partial x_j}$ are coordinate vector fields on U, then $[X, Y] = 0$ and therefore the above formula reduces to

$$R(X,Y)Z = \nabla^g_X \nabla^g_Y Z - \nabla^g_Y \nabla^g_X Z$$

which says that curvature of the Euclidean space with respect to the Euclidean metric is *zero*. In other words, the curvature of a Riemannian manifold measures the extent to which the covariant derivative is non-commutative.

6.7.2 Connections in a vector bundle

Consider a vector bundle $\pi : E \to M$ over a smooth manifold M. Let $\Gamma(E)$ denote the sections of the vector bundle. One would like to find a way of transporting the sections of this vector bundle along a curve in the base manifold M in a consistent way or equivalently transporting a section infinitesimally in a given direction. Generalizing the notion of the covariant differentiation or an affine connection on the tangent bundle of a smooth manifold, one can make the following definition.

Definition 6.7.7. A connection in a vector bundle $\pi : E \to M$ denoted by ∇ is a $\mathbb{R}$-bilinear map

$$\begin{aligned} \nabla : \chi(M) \times \Gamma(E) &\to \Gamma(E) \\ (X, s) &\mapsto \quad \nabla_X s \end{aligned}$$

satisfying the following conditions:

$$\nabla_{fX} s = f(\nabla_X s) \text{ and}$$
$$\nabla_X(fs) = X(f)s + f(\nabla_X s) \text{ for f} \in \mathrm{C}^\infty\mathrm{M}, \mathrm{X} \in \chi(\mathrm{M}), \mathrm{s} \in \Gamma(\mathrm{E}).$$

We will refer to $\nabla_X s$ as the covariant derivative of s with respect to X.

Lemma 6.7.8. *Every vector bundle admits a connection.*

Proof. We first prove the result for the trivial bundle $M \times \mathbb{R}^n,\ n \in \mathbb{N}$ of rank n over a smooth manifold M :

In fact, consider the standard co-ordinates $(x_1, ..., x_n)$ of $\mathbb{R}^n$. Then $\{s_i = \frac{\partial}{\partial x_i}\}_{i=1}^n$ is a set of n-linearly independent (global) sections of $M \times \mathbb{R}^n$. We will call $s_1, ...s_n$ the (global) frame field on $M \times \mathbb{R}^n$. Set $\nabla_X s_i = 0$ for each $i = 1, .., n$ and each $X \in \chi(M)$. Any section s of the trivial bundle is of the form $s = \sum_{i=1}^n f_i s_i$ where $f_i \in C^\infty(M)$. Define

$$\nabla_X s = \sum_{i=1}^{n} X(f_i)s_i, \quad \forall X \in \chi(M).$$

This gives a well defined connection on $M \times \mathbb{R}^n$ where the covariant derivative $\nabla_X s$ is nothing but the directional derivative of s (considered as an $\mathbb{R}^n$ valued smooth map on M) in the direction of X. ∇ as defined here is called a *trivial connection* on $M \times \mathbb{R}^n$.

Next we consider any arbitrary vector bundle $\pi : E \to M$ of rank n. Let $\{U_\alpha\}_\alpha$ be a *locally finite* open covering of M (i.e., for any $p \in M$, there are only finitely many U_αs containing p) such that $\{\pi^{-1}(U_\alpha)\}$ is a trivial bundle over U_α for each α and let ∇^α denote a trivial connection in each $\{\pi^{-1}(U_\alpha)\}$. Consider a *partition of unity* $\{f_\alpha\}_\alpha$ subordinate to the covering $\{U_\alpha\}_\alpha$ i.e., is a family of non- negative smooth functions $\{f_\alpha\}$ on M such that $\text{supp}(f_\alpha) = \overline{\{p \in M : f_\alpha(p) \neq 0\}} \subset U_\alpha$ and for any point $p \in M$, $\sum_\alpha f_\alpha(p) = 1$ (since $\{U_\alpha\}_\alpha$ is locally fine, the sum is finite for any p). To check that such functions exist, the reader is referred to Ref. 4.

Now we can define

$$\nabla_X s = \sum_\alpha f_\alpha \nabla^\alpha_X s, \quad \forall X \in \chi(M), s \in \Gamma(E).$$

It follows directly that ∇ defined above is a connection in the vector bundle $\pi : E \to M$. It is also easy to see from the definition that infinitely many connections can be constructed likewise in any vector bundle.

□

Connections as differential forms

A connection ∇ in a vector bundle $\pi : E \to M$ can locally be expressed in terms of differential forms as follows. Let U be an open subset of M such that $\pi^{-1}(U) = E|_U$ is trivial and let $\{s_i\}_{i=1}^n$ be linearly independent sections giving the frame field on $E|_U$. Given any vector field X defined on U, one has

$$\nabla_X s_j = \sum_{i=1}^{n} \omega_j^i(X)s_i$$

where $\omega_j^i(X)$ is a smooth function on U for each $i = 1, .., n$. From Definition 6.7.7, it follows that $\omega_j^i(fX) = f\omega_j^i(X)$ for each $i, j = 1, 2, ..n$ and for any $f \in C^\infty(U)$. Consequently, each ω_j^i defines a differential 1-form on U. Let $\omega = (\omega_j^i)$. Then ω is a $n \times n$ matrix whose components are differential 1-forms on U. Such a matrix is called a $M(n, \mathbb{R})$ valued differential 1-form. From now on, we will refer to ω as the *connection form* of ∇ on U. Clearly, the connection ∇ is completely determined by ω on U.

6.7.3 Curvature in a vector bundle

Given a connection ∇ in an arbitrary vector bundle $\pi : E \to M$ of rank n as above, one can talk about its curvature. The definition generalizes the notion of the curvature of a Riemannian connection in the tangent bundle of a Riemannian manifold. For any $X \in \chi(M)$, the corresponding covariant derivation $\nabla_X : \Gamma(E) \to \Gamma(E)$ is a linear operator which in case of the trivial bundle $M \times \mathbb{R}^n$, is precisely the action of the fixed vector field X on smooth $\mathbb{R}^n$ valued functions on M. Thus when ∇ denotes a trivial connection, it follows from the construction of ∇ and from the definition of Lie bracket that

$$\nabla_X \nabla_Y - \nabla_Y \nabla_X - \nabla_{[X,Y]} = 0,$$

i.e.,

$$\nabla_X(\nabla_Y s) - \nabla_Y(\nabla_X s) - \nabla_{[X,Y]} s = 0 \quad \forall \quad X, Y \in \chi(M); s \in \Gamma(M \times \mathbb{R}^n).$$

This formula is not true for an arbitrary vector bundle E with an arbitrary connection which leads to defining the curvature of a vector bundle with respect to a given connection as follows.

Definition 6.7.9. Given a connection ∇ on a smooth vector bundle $\pi : E \to M$ over a smooth manifold M, the curvature corresponding to the connection ∇ denoted by R, is the map that assigns to every pair of vector fields $X, Y \in \chi(M)$ the linear operator $R(X, Y) : \Gamma(E) \to \Gamma(E)$ defined by

$$R(X, Y)(s) = \frac{1}{2}\{\nabla_X \nabla_Y s - \nabla_Y \nabla_X s - \nabla_{[X,Y]} s\}, \quad s \in \Gamma(E).$$

Thus curvature of a trivial vector bundle with respect to a trivial connection is zero. The curvature of an arbitrary vector bundle, thus, in a way measures its deviation from a trivial bundle with a trivial connection.

Lemma 6.7.10. *The curvature R of a vector bundle $\pi : E \to M$ with a connection ∇ satisfies the following properties:*

1. $R(X, Y) = -R(Y, X)$,
2. $R(f_1 X, f_2 Y)(f_3 s) = f_1 f_2 f_3 R(X, Y)(s)$ *for any* $X, Y \in \chi(M)$, *and* $f_i \in C^\infty(M), i = 1, 2, 3$; *and* $s \in \Gamma(E)$.

Proof. Left as an exercise. (Hint: Use the properties of Lie bracket and of the connection ∇.) □

Curvature as a differential form

Just like a connection ∇ in a vector bundle $\pi : E \to M$, the corresponding curvature R can also be expressed locally in terms of differential forms. As in Section 6.7.2, consider an open subset U of M such that $\pi^{-1}(U) = E|_U$ is trivial and let $\{s_i\}_{i=1}^n$ be linearly independent sections of $E|_U$. Let $\omega = (\omega_j^i)$ be the connection form of ∇ on U as defined in Section 6.7.2. Given any two vector fields X, Y on U, one can write

$$R(X,Y)s_j = \sum_{i=1}^{n} \Omega_j^i(X,Y)s_i \tag{6.1}$$

where $\Omega_j^i(X,Y)$ is a smooth function on U for each $i,j = 1,..,n$. Using Lemma 6.7.10, one can see that $\Omega_j^i(X,Y) = -\Omega_j^i(Y,X)$ and $\Omega_j^i(f_1X, f_2Y) = f_1f_2\Omega_j^i(X,Y)$ for each $i,j = 1,2,..n$ and for any $f_1, f_2 \in c^\infty(U)$ i.e., in other words, each Ω_j^i defines a differential 2-form on U. Writing $\Omega = (\Omega_j^i)$, one obtains a $n \times n$ matrix Ω whose components Ω_j^i are differential 2-forms on U. Thus Ω defines a $M(n,\mathbb{R})$ valued differential 2-form on U and we will call it the *curvature form* of ∇ on U.

It is but natural to expect a relation between the connection and the curvature form of ∇ on U as ∇ and R are closely related. The relation is described below.

Theorem 6.7.11. *Given a vector bundle $\pi : E \to M$ and a connection ∇, the connection form ω and the curvature form Ω are related as follows*

$$\Omega = \omega \wedge \omega + d\omega$$

where $\omega \wedge \omega = (\omega \wedge \omega)_j^i$ and $d\omega = (d\omega)_j^i$, $i,j = 1,..n$; are $n \times n$ matrices whose components are differential two forms defined on U given by $(\omega \wedge \omega)_j^i = \sum_{l=1}^n \omega_l^i \wedge \omega_j^l$ and $(d\omega)_j^i = d(\omega_j^i)$ respectively. Writing the above relation component-wise, for $i,j = 1,..n$; one obtains

$$\Omega_j^i = \sum_{l=1}^{n} \omega_l^i \wedge \omega_j^l + d(\omega_j^i) \tag{6.2}$$

Proof. Combining definitions of curvature $R(X,Y)$ (for vector fields X,Y) on an open set U and the connection form as above,

$$\begin{aligned}
2R(X,Y)s_j &= \nabla_X(\nabla_Y s_j) - \nabla_Y(\nabla_X s_j) - \nabla_{[X,Y]}s_j \\
&= \nabla_X\left(\sum_{i=1}^{n} \omega_j^i(Y)s_i\right) - \nabla_Y\left(\sum_{i=1}^{n} \omega_j^i(X)s_i\right) - \sum_{i=1}^{n} \omega_j^i([X,Y])s_i \\
&= \sum_{i=1}^{n} X(\omega_j^i(Y))s_i + \sum_{i,l=1}^{n} \omega_j^l(Y)\omega_l^i(X)s_i \\
&- \sum_{i=1}^{n} Y(\omega_j^i(X))s_i - \sum_{i,l=1}^{n} \omega_j^l(X)\omega_l^i(Y)s_i - \sum_{i=1}^{n} \omega_j^i([X,Y])s_i
\end{aligned} \tag{6.3}$$

Now using the following formulae (check them!!)

$$2d\omega_j^i(X,Y) = X(\omega_j^i(Y)) - Y(\omega_j^i(X)) - \omega_j^i([X,Y]),$$

$$2\omega_l^i \wedge \omega_j^l(X,Y) = \omega_l^i(X)\omega_j^l(Y) - \omega_l^i(Y)\omega_j^l(X);$$

in (6.3) and writing the expressions in terms of matrices, the required result follows. □

Corollary 6.7.12. *As above let $\pi : E \to M$ be a vector bundle of rank n with a connection and its curvature form ω and Ω respectively defined on an open set U in M. Then the exterior derivative of Ω (also defined on U) is given by*

$$d\Omega = \Omega \wedge \omega - \omega \wedge \Omega$$

i.e., writing component-wise for each $i, j = 1, ..n;$,

$$d(\Omega_j^i) = \sum_{l=1}^{n} \{\Omega_l^i \wedge \omega_j^l - \omega_l^i \wedge \Omega_j^l\}.$$

Proof. Using Theorem 6.7.11, one has for each $i, j = 1, ..n;$

$$\Omega_j^i = \sum_{l=1}^{n} \omega_l^i \wedge \omega_j^l + d(\omega_j^i).$$

Taking exterior derivative on both sides, and using the fact that $d^2 = 0$, one obtains;

$$d(\Omega_j^i) = \sum_{l=1}^{n} \{d(\omega_l^i) \wedge \omega_j^l - \omega_l^i \wedge d(\omega_j^l)\}$$

Thus in terms of matrices one obtains,

$$\begin{aligned} d\Omega &= d\omega \wedge \omega - \omega \wedge d\omega \\ &= (\Omega - \omega \wedge \omega) \wedge \omega - \omega \wedge (\Omega - \omega \wedge \omega) \\ &= \Omega \wedge \omega - \omega \wedge \Omega. \end{aligned}$$

□

6.7.4 Behaviour of the connection and curvature forms under change of (local) trivializations

So far the connection form ω and the curvature form Ω have been studied in a single open set U of M where the restriction $E|_U = \pi^{-1}(U)$ of the vector bundle $\pi : E \to M$ is trivial. One would like to understand how the local expressions of connection and curvature form change if one moves from one local trivialization to another around the same point in M. these transformation formulae are described below.

As before, consider a vector bundle $\pi : E \to M$ of rank n with a connection ∇ Also consider two open subsets U_α, U_β in M such that $U_\alpha \cap U_\beta \neq \emptyset$ with trivializations

$$\phi_\alpha : \pi^{-1}(U_\alpha) \to U_\alpha \times \mathbb{R}^n,$$
$$\phi_\beta : \pi^{-1}(U_\beta) \to U_\beta \times \mathbb{R}^n,$$

and the transition function $h_{\alpha\beta} : U_\alpha \cap U_\beta \to GL(n, \mathbb{R})$. Let $\omega_\alpha, \Omega_\alpha$ and $\omega_\beta, \Omega_\beta$ denote the connection and curvature forms on U_α, U_β respectively. Then the following transformation formulae hold on $U_\alpha \cap U_\beta$.

Proposition 6.7.13. *With $\omega_\alpha, \Omega_\alpha$; $\omega_\beta, \Omega_\beta$ and U_α, U_β as above, one has the transformation formulae*

$$\begin{aligned} \text{(a)} \quad \omega_\beta &= h_{\alpha\beta}^{-1}\omega_\alpha h_{\alpha\beta} + h_{\alpha\beta}^{-1} d(h_{\alpha\beta}), \\ \text{(b)} \quad \Omega_\beta &= h_{\alpha\beta}^{-1}\Omega_\alpha h_{\alpha\beta}. \end{aligned}$$

Proof. (a) Start with the frame fields $s_1, ..s_n$ and $\tilde{s}_1, ..\tilde{s}_n$ on U_α and U_β corresponding to the trivializations ϕ_α and ϕ_β respectively. For $p \in U_\alpha \cap U_\beta$, one can write for each $j = 1, ..n$;

$$\tilde{s}_j = \sum_{i=1}^{n} (h_{\alpha\beta})_j^i(p) s_i \tag{6.4}$$

which follows from the properties of local trivializations and corresponding transition functions where $h_{\alpha\beta} = ((h_{\alpha\beta})_j^i)$. For $X \in \chi(M)$, consider the covariant derivation operator $\nabla_X : \Gamma(E) \to \Gamma(E)$. Recall that $\omega_\alpha = ((\omega_\alpha)_j^i)$ and $\omega_\beta = ((\omega_\beta)_j^i)$ for $i, j = 1, ..n$. Applying ∇_X to (6.4) and using the properties of connection form, one obtains

$$\sum_{l=1}^{n} (\omega_\beta)_j^l(X)\tilde{s}_l = \sum_{i=1}^{n} d((h_{\alpha\beta})_j^i)(X)s_i + \sum_{i,l=1}^{n} (h_{\alpha\beta})_j^i(\omega_\alpha)_i^l(X)s_l \tag{6.5}$$

Combining (6.4) and (6.5), and equating the i-th components from both sides, one further obtains

$$\sum_{l=1}^{n} (\omega_\beta)_j^l(X)(h_{\alpha\beta})_l^i = d((h_{\alpha\beta})_j^i)(X) + \sum_{l=1}^{n} (h_{\alpha\beta})_j^l(\omega_\alpha)_l^i(X) \tag{6.6}$$

As $X \in \chi(M)$, is arbitrary and the above equation holds for each $i, j = 1, ...n$; one can write in matrix form

$$h_{\alpha\beta}\omega_\beta = \omega_\alpha h_{\alpha\beta} + d(h_{\alpha\beta}). \tag{6.7}$$

which implies (a).

(b) From Theorem 6.7.11 it follows that on U_β, one has

$$\Omega_\beta = \omega_\beta \wedge \omega_\beta + d(\omega_\beta)$$

i.e., writing componentwise as in (6.2),

$$(\Omega_\beta)^i_j = \sum_{l=1}^{n} (\omega_\beta)^i_l \wedge (\omega_\beta)^l_j + (d(\omega_\beta))^i_j.$$

For simplicity of notations, denote $h_{\alpha\beta}$ by h. Using (6.7), one makes the following observations

$$\begin{aligned} \omega_\beta \wedge \omega_\beta &= (h^{-1}\omega_\alpha \wedge \omega_\alpha h + h^{-1}dh) \wedge (h^{-1}\omega_\alpha h + h^{-1}dh) \\ &= h^{-1}\omega_\alpha h + h^{-1}\omega_\alpha \wedge dh + (h^{-1}dhh^{-1}) \wedge \omega_\alpha h + (h^{-1}dhh^{-1}) \wedge dh \end{aligned}$$

and

$$d\omega_\beta = (dh^{-1} \wedge \omega_\alpha)h + (h^{-1}d\omega_\alpha)h - h^{-1}\omega_\alpha \wedge dh.$$

Combining the above two equations and using the fact that $h^{-1}h = Id$ (and therefore, taking exterior derivative, $h^{-1}dh + hdh^{-1} = 0$), one finally obtains

$$\Omega_\beta = h^{-1}\omega_\alpha h + h^{-1}d\omega_\alpha h = h^{-1}\Omega_\alpha h \tag{6.8}$$

□

6.7.5 Connections and curvature in a complex vector bundle

One can talk about connections and corresponding curvature of a complex vector bundle over a smooth manifold in a manner similar to the case of real vector bundles discussed earlier, although one needs to make suitable modifications. Consider a complex vector bundle $\pi : E \to M$ of complex rank n over a smooth manifold M. Here the set $\Gamma(E)$ of all sections of the vector bundle forms a module over the ring of all complex valued smooth functions $C^\infty(M;\mathbb{C})$ on M. A connection can be defined as follows.

Definition 6.7.14. A connection ∇ in a complex vector bundle $\pi : E \to M$ is a complex bilinear map

$$\begin{aligned} \nabla : \chi(M) \times \Gamma(E) &\to \Gamma(E) \\ (X, s) &\mapsto \quad \nabla_X s \end{aligned}$$

satisfying the following conditions:

$$\nabla_{fX} s = f(\nabla_X s) \text{ and}$$

$$\nabla_X(fs) = X(f)s + f(\nabla_X s) \text{ for f} \in \mathrm{C}^\infty(\mathrm{M},\mathbb{C}), \mathrm{X} \in \chi(\mathrm{M}), \mathrm{s} \in \Gamma(\mathrm{E}).$$

One can define the connection form of ∇ locally as follows. First consider the complex linear combinations of differential l-forms on M and call them the complex differential l-forms on M. Let us denote the set of all complex differential l-forms on M by $\mathcal{A}^l(M,\mathbb{C})$. One can define the exterior derivative $d : \mathcal{A}^l(M,\mathbb{C}) \to \mathcal{A}^{l+1}(M,\mathbb{C})$ by extending the ordinary exterior derivative complex linearly. Then one obtains the co-chain complex $\{\mathcal{A}^l(M,\mathbb{C}), d\}$ and

the corresponding complex de Rham cohomology which can be denoted by $H_{DR}(M, \mathbb{C})$ and one can check that

$$H^*_{DR}(M, \mathbb{C}) \equiv H^*(M, \mathbb{C})$$

where $H^*(M, \mathbb{C})$ denotes the singular cohomology of M over $\mathbb{C}$. On an open set $U \subset M$ with $E|U = \pi^{-1}(U)$ trivial, consider a frame field $s_1, ...s_n \in \Gamma(E|U)$. Then as in the case of a real vector bundle, one can write

$$\nabla_X s_j = \sum_{i=1}^{n} \omega_j^i(X) s_i$$

where each ω_j^i defines a complex differential 1-form on U and $\omega = (\omega_j^i)$ is the $n \times n$ matrix known as the connection form of ∇ on U. In a similar fashion, one can define the curvature form $\Omega = (\omega_j^i)$ whose components (ω_j^i) are complex differential two forms defined U. Thus in this case, ω and Ω respectively define $M(n, \mathbb{C})$ valued 1-form and 2-form on U. It also follows directly from the construction of ω and Ω that Theorem 6.7.11, Corollary 6.7.12 and both the transformation formulae given in Proposition 6.7.13 continue to hold in the complex vector bundle set up as well. The only difference is that the transition functions $h_{\alpha\beta}$ (as described in Proposition 6.7.13) in this case will have values in $GL(n, \mathbb{C})$.

6.8 Characteristic classes

The next step would be to construct some global invariants for a vector bundle that respect natural bundle maps and work as tools to distinguish two given vector bundles of the same rank over a manifold. (Recall that given two vector bundles $\pi_i : E_i \to M_i$, $i = 1, 2$; a bundle map between these two bundles is described by a pair (f, F) where $f : M_1 \to M_2$ and $F : E_1 \to E_2$ are smooth maps such that the following diagram commutes

$$\begin{array}{ccc} E_1 & \xrightarrow{F} & E_2 \\ \downarrow{\scriptstyle \pi_1} & & \downarrow{\scriptstyle \pi_2} \\ M_1 & \xrightarrow{f} & M_2 \end{array}$$

and for each $x \in M_1$, the restricted map $F : E_1|_x \to E_2|_{f(x)}$ which preserves the fibers of the bundles, is linear.) The first thing to observe is that the curvature of a vector bundle endowed with some connection is described in terms of locally defined differential forms on the base manifold giving local information about the vector bundle and its deviation from the trivial bundle of the same rank. Using these local information and properties of a curvature form, one can try to construct global differential forms on the base manifold. If one can further prove this form to be closed, then it corresponds to a de Rham cohomology

class of the manifold which is a topological invariant. With this strategy one proceeds as follows. First one combines all these local differential forms into a global one using an algebraic entity known as invariant polynomials which is described below.

6.8.1 Invariant polynomials

Definition 6.8.1. Let $M(n, \mathbb{K})$ denote the space of all $n \times n$ matrices and $GL(n, \mathbb{K})$ the space of all invertible matrices of the same order over $\mathbb{K}$ where $\mathbb{K}$ denotes the field of real or complex numbers. A polynomial function $P : M(n, \mathbb{K}) \to \mathbb{K}$ (by a polynomial function $P : M(n, \mathbb{K}) \to \mathbb{K}$, one means a polynomial in the entries of matrices) is said to be an invariant polynomial if

$$P(X) = P(A^{-1}XA) \ \ \forall \ X \in M(n, \mathbb{K}), A \in GL(n, \mathbb{K}).$$

Two of the most elementary examples of invariant polynomials are the determinant and trace of an $n \times n$ matrix. These are invariant polynomials of degree n and 1 respectively. It can be checked that the set all invariant polynomials as defined above is a commutative algebra denoted by I_n with respect to the standard addition and multiplication of polynomials.

Properties of an invariant polynomial

A few properties of invariant polynomials are given here. For a detailed proof of these properties see Ref. 4.

(1) P is an invariant polynomial if and only if

$$P(XY) = P(YX) \ \ \forall \ X, Y \in M(n, \mathbb{K}).$$

(2) An invariant polynomial P can be uniquely expressed as a real or complex polynomial of elementary symmetric functions $\sigma_1, ..., \sigma_n$ where the elementary symmetric functions are as follows. Given $X \in M(n, \mathbb{K})$, $\sigma_i(X)$ denotes the i-th elementary symmetric function of the eigenvalues of X given by the equation

$$det(I + tX) = 1 + t\sigma_1(X) + t^2\sigma_2(X) + ... + t^n\sigma_n(X).$$

Then it follows from the above equation that for each $i = 1, .., n; \sigma_i : M(n, \mathbb{K}) \to \mathbb{K}$ is an invariant polynomial.

Constructing global differential forms on a smooth manifold using (local) curvature forms

In this section we will construct global differential forms on a smooth manifold from (local) curvature forms of a given vector bundle over the manifold. Naturally, one needs the transformation formulae for the curvature forms while moving from one local trivialization of the vector bundle to another around the same point on the manifold and then an invariant polynomial to define a well defined global differential form.

Consider a (real or complex) vector bundle $\pi : E \to M$ of (real or complex) rank n over a smooth manifold M with a connection ∇ and curvature R. Then there is an open covering $\{U_\alpha\}$ of M such that for each α, $E|_{U_\alpha}$ is trivial and the curvature R is given by the curvature form $\Omega_\alpha = ((\Omega_\alpha)^i_j)$, an $n \times n$ matrix whose components are differential two forms $(\Omega_\alpha)^i_j$ defined on U_α. Moreover, the curvature forms $\{\Omega_\alpha\}$ and $\{\Omega_\beta\}$ defined respectively on U_α and U_β are related to each other on $U_\alpha \cap U_\beta$ by

$$\Omega_\beta = h^{-1}_{\alpha\beta}\Omega_\beta h_{\alpha\beta}$$

where $h_{\alpha\beta} : U_\alpha \cap U_\beta \to GL(n, \mathbb{K})$, $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, denotes the transition function as before.

Combining the locally defined curvature forms together one defines $M(n, \mathbb{K})$ valued global differential forms on M (i.e., $n \times n$ matrices whose components are differential forms on M) in the following manner. Observe that the forms Ω_α and Ω_β are similar to each other in $U_\alpha \cap U_\beta$. Applying any invariant polynomial P on the $n \times n$ matrices Ω_α one obtains a globally well defined differential form as in this case one has $P(\Omega_\alpha) = P(\Omega_\beta)$ whenever $U_\alpha \cap U_\beta \neq \emptyset$. Let us denote this global differential form on M by $P(\Omega)$ for any invariant polynomial P. In fact, one makes a stronger conclusion.

Proposition 6.8.2. *Given an invariant polynomial $P \in I_n$ of degree l, the global differential form $P(\Omega)$ as defined above is a closed $2l$-form on M.*

In the case of a complex vector bundle $\pi : E \to M$ of complex rank n, $P(\Omega)$ as above (P is an invariant polynomial of degree l acting on $n \times n$ complex matrices) defines a closed complex differential $2l$-form on M.

For a proof of the above proposition, see Ref. 4.

For an invariant polynomial P of degree l acting on $n \times n$ real or complex matrices, consider the de Rham cohomology class $[P(\Omega)] \in H^{2l}_{DR}(M, \mathbb{K})$, $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, of the closed (real or complex) differential form $P(\Omega)$ of degree $2l$ on a smooth manifold M. One observes that

Proposition 6.8.3. *The de Rham cohomology class $[P(\Omega)] \in H^{2l}_{DR}(M, \mathbb{K})$, $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, for each l is independent of the choice of the connection ∇ on the (real or complex) vector bundle $\pi : E \to M$.*

For a proof of the above proposition, see Ref. 4.

The above proposition says that not only the differential forms are independent of local trivializations of the vector bundle but are also independent of which connection is chosen. Thus given a vector bundle, a connection or its curvature play only an auxiliary role to define some de Rham cohomology classes of the manifold which are global invariants. Therefore one can denote $P(\Omega)$ by $P(E)$ and call it the characteristic class of E corresponding to the invariant polynomial P.

Furthermore, one has the following result which in particular says that two isomorphic vector bundles have the same characteristic classes. The proof of this result is not discussed in this note. See Ref. 4 for proof.

Proposition 6.8.4. *The characteristic class is natural with respect to a bundle map, i.e., if $f : N \to M$ is a smooth map between manifolds and $\pi : E \to M$ is a vector bundle (real or complex) over the smooth manifold M, then for the pull back bundle $f^*(E)$ over N one has*

$$P(f^*(E)) = f^*(P(E)) \in H^{2k}(N, \mathbb{C}), \quad \textit{for any invariant polynomial } P,$$

where the pull back bundle $f^*(E)$ over N is defined fiber-wise by

$$f^*(E)|_x = E|_{f(x)},$$

for each $x \in N$.

6.8.2 Chern classes and Pontrjagin classes

In this section we will talk about two important characteristic classes. The former, called the *Chern classes*, are defined for complex vector bundles; while the latter, called *Pontrjagin classes*, are defined for real vector bundles.

Chern classes

Definition 6.8.5. For a complex vector bundle $\pi : E \to M$ of complex rank n over a smooth manifold M, the Chern class of degree l denoted by $c_l(E)$ is the characteristic class corresponding to the degree l invariant polynomial

$$\left(\frac{-1}{2\pi i}\right)^l \sigma_l \in I_n(\mathbb{C}),$$

where σ_l denotes the elementary symmetric function of degree l. As a cohomology class in terms of the local curvature form Ω of the vector bundle, one can write,

$$\left[det\left(Id - \left(\frac{-1}{2\pi i}\right)\Omega\right)\right] = 1 + c_1(E) + c_2(E) + ... + c_n(E) \in H^*_{DR}(M, \mathbb{C}).$$

The expression on the right hand side is called the total Chern class of the vector bundle denoted by $c(E)$.

The first observation that one makes is that

Proposition 6.8.6. *For each l, $1 \leq l \leq n$, the Chern class $c_l(E)$ defines a real cohomology class i.e., $c_l(E) \in H^{2l}_{DR}(M, \mathbb{R})$.*

Again we have skipped the proof here which can be found in Ref. 4.

In fact, one can say more. The normalizing constant term attached to the symmetric function in the definition of the Chern class of degree l ensures that this represents a de Rham cohomology class with integral coefficients i.e., for each l, $1 \leq l \leq n$, $c_l(E) \in H^{2l}_{DR}(M, \mathbb{Z})$ for a complex vector bundle E of rank n over M.

Pontrjagin classes

Definition 6.8.7. For a real vector bundle $\pi : E \to M$ of rank n over a smooth manifold M, the Pontrjagin class of degree l denoted by $p_l(E) \in H^{4l}_{DR}(M, \mathbb{R})$ is the characteristic class corresponding to the degree $2l$ invariant polynomial

$$\left(\frac{1}{2\pi}\right)^{2l} \sigma_{2l} \in I_n(\mathbb{R}),$$

where σ_{2l} denotes the symmetric function of degree $2l$. In terms of the local curvature form Ω of the vector bundle one can write,

$$\left[\det\left(Id - \left(\frac{1}{2\pi}\right)\Omega\right)\right] = 1 + p_1(E) + p_2(E) + ... + p_{[\frac{n}{2}]}(E) \in H^*_{DR}(M, \mathbb{R}).$$

The expression on the right hand side is called the total Pontrjagin class of the vector bundle denoted by $p(E)$.

As for the Chern classes, here also the normalizing constant attached to the symmetric function in the definition of the Pontrjagin class of degree l ensures that this represents a de Rham cohomology class with integral coefficients i.e., for each l, $1 \leq l \leq [n/2]$, $p_l(E) \in H^{4l}_{DR}(M, \mathbb{Z})$ for a real vector bundle E of rank n over M.

Remark 6.8.8. The above definition of the Pontrjagin class of degree l is quite similar to that of the Chern class of degree $2l$ for a complex vector bundle. The difference is that, in the real case, i.e., while defining Pontrjagin classes, in view of the following fact (for a proof of the fact see Ref. 4), one only considers symmetric functions of even degrees.

Proposition 6.8.9. *Given an invariant polynomial P of odd degree, the characteristic class $P(E) = 0$ where $\pi : E \to M$ is areal vector bundle.*

This is not true when E is a complex vector bundle.

Relating Chern and Pontrjagin classes

Given a real vector bundle $\pi : E \to M$ of rank n over a smooth manifold M, one has its Pontrjagin classes $p_l(E) \in H^*_{DR}(M, \mathbb{Z})$. Consider the complexification $E \otimes \mathbb{C}$ of E which is a complex vector bundle of complex rank n over M. Then one can talk about the Chern classes $c_l(E) \in H^*_{DR}(M, \mathbb{Z})$. These two characteristic classes are closely related to each other as given by the following proposition.

Proposition 6.8.10. *Let $\pi : E \to M$ be a real vector bundle of rank n over a smooth manifold M, and $E \otimes \mathbb{C}$ its complexification. Then with notations described above, one has*

$$p_l(E) = (-1)^l c_{2l}(E \otimes \mathbb{C}) \in H^{4l}_{DR}(M, \mathbb{Z}).$$

For a proof of the above proposition the reader is referred to Ref. 4. The above formula can also be used as the definition of the Pontrjagin classes of a real vector bundle E (in terms of the Chern classes of its complexification $E \otimes \mathbb{C}$).

Both the Chern and Pontrjagin classes behave well with respect to the sum of two vector bundles as described below. For a verification of the following result the reader is referred to Ref. 4.

Proposition 6.8.11. *(1) If E and F are complex vector bundles over a smooth manifold M, then*

$$c_l(E \oplus F) = \sum_{j=0}^{l} c_j(E) c_{l-j}(F),$$

and hence the total Chern class is given by

$$c(E \oplus F) = c(E)c(F).$$

(2) If E and F are real vector bundles over a smooth manifold M, then

$$p_l(E \oplus F) = \sum_{j=0}^{l} p_j(E) p_{l-j}(F),$$

and hence the total Pontrjagin class is given by

$$p(E \oplus F) = p(E)p(F).$$

6.8.3 Examples

From the previous discussions, it follows that all the Chern and Pontrjagin classes of trivial bundles (real and complex respectively) over any smooth manifold are zero. In fact, the curvature of a trivial connection on a trivial vector bundle vanishes identically. One would like to know about the Chern (or Pontrjagin) classes of some non-trivial vector bundles. Here we discuss the simplest example of the tangent bundle of sphere and its characteristic classes.

(a) The tangent bundle of the Riemann sphere

Consider the Riemann sphere (which can be thought as the complex projective space $\mathbb{C}P^1$) with a holomorphic local co-ordinate chart $(U, (z))$. Let $T\mathbb{C}P^1$ denote the complex tangent bundle of $\mathbb{C}P^1$. A complex tangent vector at a point in U is of the form $\lambda \frac{\partial}{\partial z}$, $\lambda \in \mathbb{C}$. One can show that $c_1(T\mathbb{C}P^1) \neq 0$. (Since this is a real vector bundle of rank two, there is no question of the existence of its Pontrjagin classes.)

Consider the Kähler metric g which with respect to the holomorphic co-ordinate (z), is of the form

$$g = \frac{dz d\bar{z}}{(1 + |z|^2)^2}.$$

Consequently, in this co-ordinate chart U, the curvature 2-form Ω looks like

$$\Omega = \frac{2dz \wedge d\bar{z}}{(1+|z|^2)^2},$$

and the first Chern class is given by

$$c_1(T\mathbb{C}P^1) = \left[\frac{i}{2\pi}\mathrm{tr}(\Omega)\right].$$

Computing its integral over $\mathbb{C}P^1$, one obtains

$$\int c_1 dz \wedge d\bar{z} = \frac{i}{\pi}\int \frac{dz \wedge d\bar{z}}{(1+|z|^2)^2} = 2.$$

Thus it represents a non-zero cohomology class, i.e., $T\mathbb{C}P^1$ is not a trivial bundle.

(b) The tangent bundle of the sphere of dimension bigger than two

The tangent bundle of sphere S^n, $n \geq 3$; is real vector bundle of rank n. Although, in general, it is not a trivial bundle, but it turns out that all its Pontrjagin classes vanish. This can be seen as follows.

Consider the trivial bundle $T\mathbb{R}^{n+1} = \mathbb{R}^{n+1} \times \mathbb{R}^{n+1}$, i.e., the tangent bundle of $\mathbb{R}^{n+1}$. Then its restriction $E = T\mathbb{R}^{n+1}|_{S^n}$ is also a trivial real vector bundle over S^n of rank $n+1$ and all its Pontrjagin classes

$$p_j(E) = 0, \text{for} 1 \leq j \leq [\frac{n+1}{2}],$$

vanish identically. Define the sub-bundle L of E by $L = (TS^n)^{\perp}$ in E. Then L is real line bundle over S^n such that

$$E = TS^n \oplus L.$$

Then using the formula in the second part of Proposition 6.8.11, it follows that all the Pontrjagin classes $p_j(TS^n) = 0$ for j, $1 \leq j \leq [\frac{n}{2}]$; also vanish identically.

References

[1] S. Gallot, D. Hulin and J. Lafontaine, *Riemannian Geometry*, Third edition, (Springer-Verlag, Berlin, 2004).

[2] J. W. Milnor and J. D. Stasheff, *Characteristic Classes*, (Princeton University Press, 1976).

[3] R. O. Wells Jr. , *Differential Geometry on Complex Manifolds*, (Prentice Hall, 1973; Springer, 1979)

[4] S. Morita, *Geometry of Differential Forms* (American Mathematical Society, Providence, RI, 2001).

[5] S. Morita, *Geometry of Characteristic Classes* (American Mathematical Society, Providence, RI, 2001).

7

Special Topics: A Crash Course on Knots

Mahan Mj

We give a quick introduction to Knot Theory following standard sources in the subject. With the help of examples, we illustrate the ideas of knot invariants, knot groups, Wirtinger presentation, torsion, Seifert surfaces, Skein relations, Alexander and Jones polynomials, and linking number. A few examples of knots and links are given in the appendix.

7.1 Introduction: Equivalence between Knots

We shall be mainly using Dale Rolfsen's *Knots and Links* [1] as the primary source below.

Definition 7.1.1. An embedded copy of the circle S^1 in Euclidean 3-space $\mathbb{R}^3$ or the 3-sphere S^3 is called a **knot**.

The union of finitely many copies of S^1 in Euclidean 3-space $\mathbb{R}^3$ or the 3-sphere S^3 is called a **link**.[1]

Knot theory mainly attempts to answer the question:

Question 7.1.2. *Given two knots K_1, K_2 in $\mathbb{R}^3$ are they equivalent?*

To answer this question, we need to come up with appropriate notions of equivalence. Usually two equivalent notions are used:

[1] See Appendix for diagrams of a few simple knots and links.

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_7

Definition 7.1.3. K_1 and K_2 are said to be **equivalent** if there is an orientation-preserving homeomorphism $h : \mathbb{R}^3 \to \mathbb{R}^3$ such that $h(K_1) = K_2$.

K_1, K_2 are said to be **isotopic** if there is a one-parameter family h_t of embeddings ($t \in [0, 1]$) of S^1 in $\mathbb{R}^3$ such that $h_0(S^1) = K_1$ and $h_1(S^1) = K_2$.

An isotopy h_t is called an ambient isotopy if h_t can be extended to a one-parameter family of diffeomorphisms of $\mathbb{R}^3$.

The **Isotopy Extension Theorem** shows that two knots are equivalent if and only if they are isotopic.

A weaker equivalence between knots is obtained by demanding only that their complements in $\mathbb{R}^3$ (equipped with an orientation) are homeomorphic via an orientation-preserving homeomorphism.

In a famous paper *Knots are determined by their complements. J. Amer. Math. Soc. 2 (1989), no. 2, 371-415*, by Cameron Gordon and John Luecke, the authors showed that if the complements of two tame knots are homeomorphic via an orientation-preserving homeomorphism, then the knots are equivalent. This is referred to cryptically as:

Theorem 7.1.1. *Knots are determined by their complements.*

7.2 Knot Invariants

It therefore suffices to study **algebraic invariants** of the knot complement $\mathbb{R}^3 \backslash K$. Here, by an algebraic invariant of a space X, we mean a natural way of associating to X an algebraic gadget, e.g. a group or a module, or a polynomial $A(X)$, such that if X and Y are homeomorphic, then $A(X)$ and $A(Y)$ are isomorphic.

Examples include:

1. The fundamental group $\pi_1(X)$,
2. More generally, higher homotopy groups $\pi_n(X)$,
3. Homology groups $H_n(X)$,
4. (The dual) Cohomology groups $H^n(X)$.

These invariants have been discussed in various chapters of this book.[2]

7.3 The Knot Group

In general, the conceptually simplest invariant is the knot group $\pi_1(\mathbb{R}^3 \backslash K)$. This is quite a sensitive invariant and distinguishes between most distinct knots. It is also easy to compute given a planar projection of the knot as we show below. The problem with the invariant is that if the same knot has two different planar projections, then they give two different presentations of the same group; and in general it is hard to decide if two presentations give isomorphic groups or not (this is the so-called **Isomorphism Problem**).

[2] We refer, in particular, to Chaps. 1, 3, 4, and 5 for more details on these topics.

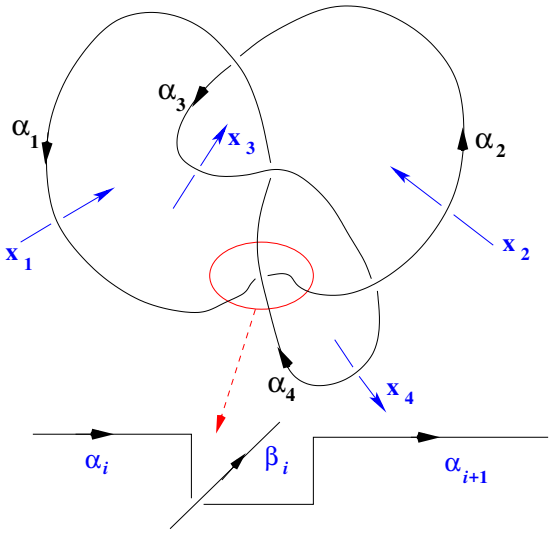


Figure 7.1: Wirtinger presentation of a knot. The strands between two successive crossings are denoted by α_i. In this figure, actually for a figure-8 knot, there are four strands, $\alpha_1, ...\alpha_4$. In the lower figure, β_i represents any other strand.

7.3.1 Wirtinger presentation

Instead of giving a general description of the algorithm to compute the Wirtinger presentation, we illustrate it by means of the example in Fig. 7.1.

For the example of Fig. 7.1, the generators of the group are $x_1, \cdots, x_4$, which are constructed as described below. In the general situation, we consider an oriented loop starting from a point $*$ very far away coming down to the plane on which the knot is projected, going below a strand and then going back to $*$. For convenience we only give an orientation on the piece of this loop that crosses the knot below a strand of the knot. In Fig. 7.1 x_1 crosses the strand α_1 from below. Proceeding anticlockwise, we come to the encircled region, where there is a knot crossing. Then x_2 crosses the strand α_2 from below. Note that the loop indicated by x_1 cannot be homotoped to the loop indicated by x_2 as the strand indicated by α_4 comes in the way. Thus all maximal strands not disconnected by another strand crossing from the top gives rise to a generator of the knot.

Relations: Now we compute the relations:

At the encircled crossing in the diagram, there are three of the x's. x_4 crosses from right to left both below and above the crossing, x_1 crosses from bottom to top on the left of the crossing and x_2 crosses from bottom to top on the right of the crossing. This gives rise to the relation

$$x_4 x_1 = x_2 x_4.$$

Note that this could also be written as

$$x_4 x_1 x_4^{-1} = x_2.$$

Similarly for the other crossings, we get

1. $x_2 x_3 = x_4 x_2$
2. $x_1 x_3 = x_2 x_1$
3. $x_4 x_3 = x_3 x_1$

This gives us a full presentation of the complement of the knot described above, which is also called the figure 8 knot (see Fig. 7.5).

It turns out that (any) one of the relations can always be dropped as it is a consequence of the remaining ones.

7.3.2 The first homology

The first homology $H_1(X)$ is the abelianization of the fundamental group, i.e., it is the group obtained by declaring that all the generators commute.

Theorem 7.3.1. *$H_1(\mathbb{R}^3 \backslash K) = \mathbb{Z}$ for any knot K.*

Again we illustrate this in the figure 8 knot complement case. We need to abelianize the 4 relations above. These give, respectively,

1. $x_1 = x_2$
2. $x_3 = x_4$
3. $x_3 = x_2$
4. $x_4 = x_1$

Thus we have that

$$H_1(\mathbb{R}^3 \backslash K) = \langle x_1, x_2, x_3, x_4 : x_1 = x_2 = x_3 = x_4 \rangle = \mathbb{Z}.$$

7.4 Torsion

Since all knots have the same first homology, $H_1(\mathbb{R}^3 \backslash K)$ is of no use as a knot invariant. However, it can be used to extract finer invariants by passing to finite index subgroups of $\pi_1(\mathbb{R}^3 \backslash K)$.

Let $X = S^3 \backslash K$ and let X_j be the $j-$fold cyclic cover of X. The torsion-part of $H_1(X_j)$ is called the $j-$th torsion invariant of K.

This can be defined purely algebraically as follows.
$\pi_1(S^3 \backslash K) \to H_1(S^3 \backslash K) = \mathbb{Z}$ be the abelianization map. Compose this with the map $\mathbb{Z} \to \mathbb{Z}/k$. Let N_k be the kernel of this map. Then the abelianization of N_k is $H_1(X_k)$ and its torsion part is the $k-$th torsion invariant of K. We refer the reader to Dale Rolfsen's *Knots and Links*, Chapter 6, pgs. 145-150 where explicit computations of torsion invariants of the trefoil knot (Fig. 7.5) are made [1]. In particular, the second torsion invariants of the trefoil and the figure eight knots are $\mathbb{Z}/3$ and $\mathbb{Z}/5$ respectively.

7.5 Seifert surfaces

A Seifert surface for a knot or link K is a connected bicollared compact surface Σ with $\partial\Sigma = K$. Any oriented knot or link K has an oriented Seifert surface bounding it. A Seifert surface for K is constructed as follows. Take a planar

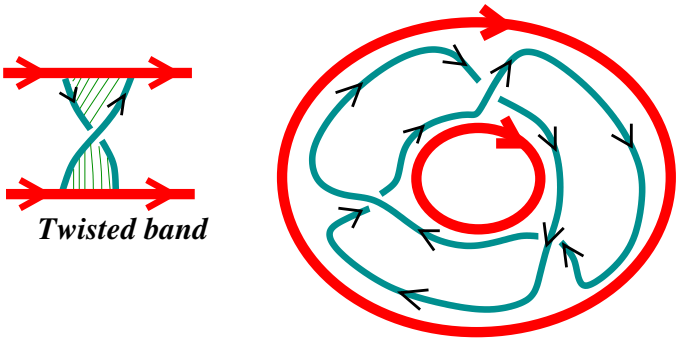


Figure 7.2: Seifert surface for the trefoil (see Fig. 7.5). The thick circles bound oriented disks. These are glued together using twisted bands as on the left.

projection of K. Near each crossing point, delete the over- and undercrossings and replace them by 'short-cut' arcs preserving orientation. This gives rise to a disjoint collection of oriented simple closed curves. They bound disks, which may be pushed slightly off each other if necessary to make them disjoint. Finally we connect these disks together at the original crossings using half-twisted strips. The result is a surface with boundary K.

7.6 Alexander Polynomial

To compute torsion invariants we used finite cyclic covers. The Alexander polynomial is computed using an infinite cyclic cover corresponding to the map $\pi_1(S^3\backslash K) \to H_1(S^3\backslash K) = \mathbb{Z}$. Let X denote the knot complement and S a Seifert surface. Cut X open along S and attach infinitely many copies end to end to obtain the infinite cyclic cover Y. Let t be the generators of the deck transformation group (isomorphic to $H_1(S^3\backslash K) = \mathbb{Z}$). Thus $H_1(Y)$ can be regarded as a $\mathbb{Z}[t, t^{-1}]-$module called the **Alexander module**. The presentation matrix for the Alexander module is called the **Alexander matrix**. When the number of generators, k, is less than or equal to the number of relations, s, then the ideal generated by all $k \times k$ minors of the Alexander matrix is called the **Alexander ideal**. When the Alexander ideal is principal, its generator is called an **Alexander polynomial** of the knot. A detailed computation of the Alexander polynomial of the trefoil ($= t^2 - t + 1$) may be found in Dale Rolfsen's *Knots and Links*, Pages 163-165. [1] For a list of Alexander polynomials for knots, see Table 1 of Ref. [6].

7.7 Skein Relations

Skein relations are of the form

$$F(L_0, L_+, L_-) = 0,$$

where L_0, L_+, L_- are the three possible crossings at a point. The three knots that differ at one point, because of these three crossing patterns, are denoted by K_0, K_+, K_-.

An example is given by the following figure (from Mina Aganagic's article [2]).

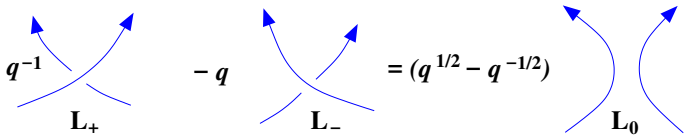


Figure 7.3: Skein relation for the Jones polynomial

Finding an F which produces polynomials independent of the planar projection used in a recursion is not easy. Jones uncovered an underlying structure of skein relations when he discovered planar algebras. A skein relation can be thought of as defining the kernel of a quotient map from the planar algebra of tangles. Such a map gives rise to a knot polynomial if all closed diagrams are taken to some (polynomial) multiple of the image of the unknot.

7.7.1 Alexander polynomial

We have given a geometric description of the Alexander polynomial above. Conway discovered the following skein relation that computes the Alexander polynomial.

$$A_{K_+} - A_{K_-} = (q^{\frac{1}{2}} - q^{-\frac{1}{2}})A_{K_0}.$$

7.7.2 Jones polynomial

The Jones polynomial was discovered by Vaughan Jones in 1984. It is a Laurent polynomial in $q^{\frac{1}{2}}$. The figure given above actually gives us the way to compute the Jones polynomial by furnishing the skein relation the Jones polynomial satisfies:

$$q^{-1}J_{K_+} - qJ_{K_-} = (q^{\frac{1}{2}} - q^{-\frac{1}{2}})J_{K_0},$$

and its value for the unknot. An example of computation of J(trefoil) with the choice of J(unknot) $= 1$, can be found in Chap. 14 of this book.

7.8 Linking Number

The linking number is an invariant of a link having two components, K_1 and K_2. In a sense this was the oldest knot or link invariant. It was discovered by Gauss. Choose a planar projection of the link onto a plane and define the linking number to be half the number of crossings counted with sign (using the right hand thumb rule after orienting the link, or using the $\pm$ convention of Fig. 7.3), i.e.,

$$m(K_1, K_2) = \frac{1}{2} \sum_{\text{crossings}(K_1, K_2)} \text{sign(crossing)}.$$

See the following figure (from Ref. [2])

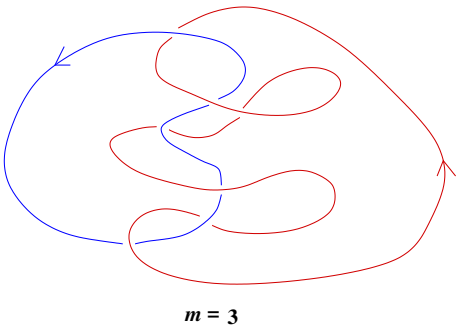


Figure 7.4: A planar projection of a link.

Gauss's discovery of the linking number came from his study of electrostatics, and he gave the following formula (see Ref. [2,3] for instance) describing the same topological invariant:

$$m(K_1, K_2) = \frac{1}{2\pi} \oint_{K_1} \oint_{K_2} \frac{\mathbf{x}_1 - \mathbf{x}_2}{|\mathbf{x}_1 - \mathbf{x}_2|^3} \cdot (d\mathbf{x}_1 \times d\mathbf{x}_2),$$

where $\mathbf{x}_1, \mathbf{x}_2$ are the position vectors on loops K_1, K_2 respectively. Some use of this Gauss formula can be found in Chapter 9.

For a list of Alexander polynomials and linking numbers for a class of links, see Table II of Ref. [6].

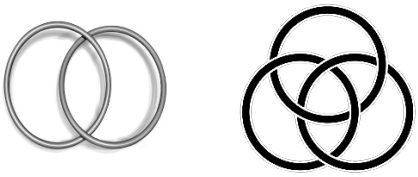
Appendix: Examples of knots and links

A few common examples of knots are given in Fig. 7.5.

unknot trefoil figure-8

Figure 7.5: Examples of knots [8].

Two common links are shown in Fig. 7.6.



Hopf link Borromean rings

Figure 7.6: Examples of links [9].

References

[1] D. Rolfsen, *Knots and links*, Mathematics Lecture Series, No. 7. (Publish or Perish, Inc., Berkeley, Calif., 1976).

[2] M. Aganagic, String Theory and Math: Why This Marriage Can Last, Mathematics and Dualities of Quantum Physics, Bulletin of the American Mathematical Society, **53**, 93 (2016).

[3] J. Baez and J. P. Muniain, Gauge fields, knots and gravity. Series on Knots and Everything, 4. (World Scientific Publishing Co., River Edge, NJ, 1994).

[4] J. H. Conway, An enumeration of knots and links, and some of their algebraic properties. 1970 Computational Problems in Abstract Algebra (Proc. Conf., Oxford, 1967) pp. 329-358 (Pergamon, Oxford)

[5] V. F. R. Jones, A polynomial invariant for knots via von Neumann algebras. Bulletin of the American Mathematical Society, (N.S.) **12**, 103 (1985).

[6] M. D. Frank-Kamenetskii, A. V. Vologodskiii, "Topological aspects of the physics of polymers: The theory and its biophysical applications", Sov. Phys-Uspekhi **24**, 679 (1981).

[7] Louis H Kauffman, *Knots and Physics*, 3rd Edition, Series on Knots and Everything: Volume 1, (World Scientific, Singapore, 2001).

[8] https://en.wikipedia.org/wiki/Trefoil_knot;
https://commons.wikimedia.org/wiki/File:Figure8knot-01.png.

[9] https://commons.wikimedia.org/wiki/File:Hopf_link.png;
https://commons.wikimedia.org/wiki/Category:Borromean_rings

8

Special Topics: A Short Course on Group Theory

Bobby Ezhuthachan

These lectures provide a brief introduction to group theory, largely focussing on finite groups. After giving the basic definition of groups, we start with a discussion on abelian groups followed by a discussion on non-abelian groups in the next section. Along the way, we define normal subgroups and conjugacy classes and discuss the commutator subgroup and abelianization. In the final sections, we discuss the examples of the Quaternionic group, as well as two examples of continuous groups- the rotation group, in particular its connection with the group of special unitary matrices in two dimensions as well as the conformal group.

8.1 Groups and Physics

Symmetry plays an important role in Physics. In classical theory, Noether's theorem relates symmetries of the action to conservation laws. So for instance, if the action, or Hamiltonian in the phase space formulation, is invariant under rotations, then the system described by the action, has total angular momentum conserved in time, just as invariance of the action under translation in space implies the conservation of total momentum in time. The set of such symmetry transformations which leaves something invariant (the action in this case), forms what is mathematically called a "group". The action of these symmetry transformations on a physical system are described by matrices- called group representations. Taking products of such matrices, corresponds to doing successive symmetry transformations of the system.

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_8

The ideas of symmetry transformations and groups are even more powerful in quantum theory. For instance in quantum mechanics, if a system has some symmetry, then the corresponding group-matrices commute with the Hamiltonian. This fact along with a well known theorem called the Schur's lemma, can help in solving the problem of finding the energy levels of a system in some cases.

Group theory helps in classifying the various particles found in nature, since the various fundamental particles of nature, as described by the standard model of particle physics, correspond to different representations of the corresponding symmetry group.

It is sometimes the case that the ground state is invariant under a smaller set of symmetries than the Hamiltonian. This is known as ' spontaneous symmetry breaking' and plays a crucial role in very important and diverse physical phenomenon, like superconductivity and the 'Higgs mechanism' which gives mass to various fundamental particles in nature. Most phases of matter can be classified by the amount of symmetry that they break. As an example, in crystals, the underlying invariance of the Hamiltonian under continuous translations is broken to a set of discrete translations in the crystalline phase.

So, in short, the study of Group theory is well motivated in Physics. Before discussing some properties and examples of groups, we begin with the basic definition of groups in the next section.

8.2 Basic definition

A Group G is a set of elements $(g_1, g_2, ...)$, equipped with a composition rule, that basically tells us how composing any two elements of this set gives rise to a third $g_i \circ g_j = g_k$. This set along with the composition rule, has to satisfy the following conditions.

- Associativity: $(g_i \circ g_j) \circ g_k = g_i \circ (g_j \circ g_k)$
- Existence of an identity element denoted by e such that $e \circ g_i = g_i \ \forall \ i$.
- Existence of an inverse element for every element g_i for which we use the notation by g_i^{-1} such that $g_i^{-1} \circ g_i = e$

As a corollary of these properties, it follows that e and g_i^{-1} are unique and that $g \circ e = g$ and $g \circ g^{-1} = e$.

If the elements of the set G can be labelled by an integer, then the group is called *discrete*, while if its labelled by continuous numbers, then the group is called *continuous*. If a discrete group has a finite number of elements, then the number of elements is called the *order* of the group and is usually denoted by the symbol $|G|$.

In general, the composition rule for the group elements is not commutative. However if in special cases, the composition rule is commutative for all elements, then such a group is called an *abelian* group, else its called *non-abelian*.

The condition that two elements g_i, g_j commute, can be recast as $g_i^{-1} \circ g_j^{-1} \circ g_i \circ g_j = e$. An element of a group $g \in G$ is called a commutator if it can be expressed as $g = a^{-1} \circ b^{-1} \circ a \circ b \equiv [a, b]$. Where a, b are any two elements of G. Stated in terms of the commutator, a group is *abelian* iff the only commutator element of the group is the identity.

Some simple examples of Groups:

- The set of integers under addition ($\mathbb{Z}$).
- The set of continuous rotations of a rigid body
- Permutation of n objects
- The set of unitary $n \times n$ Matrices.

In the next section, we will first discuss the *abelian* case, and then in later sections discuss *non-abelian* groups.

8.3 Abelian Groups

For the *abelian* group, as is standard, we will borrow the notation from $\mathbb{Z}$. We therefore denote the composition rule by the addition symbol ($+$), the identity element by 0 and the inverse by $-g$. So that $g_i + g_j = g_k$ and $g + (-g) \equiv g - g = 0$.

If a subset of elements of a group, itself forms a group under the same composition law, then its called a subgroup. For example, the group $\mathbb{Z}$ has a subgroup which is simply obtained by multiplying each element by any specific integer say N, that is: $N\mathbb{Z}$ is a subgroup of $\mathbb{Z}$. Given an abelian group G and some subgroup H, we can form a set of *equivalence classes*, called *coset*, where the elements of each class consists of all such elements, (g, x) which are related as: $\{g, x \subset G;$ such that $g = x + h;$ where $h \subset H\}$. It is easy to check that the above relation is an equivalence relation (which we denote as: $(g \sim x)$), because the relation is reflexive, transitive and associative.

- (a) reflexivity: $g \sim g$, because $g = g + 0$
- (b) transitivity: $g \sim x \Rightarrow x \sim g$, because $g = x + h \Rightarrow x = g - h$
- (c) associativity: if $g \sim x$ and $x \sim y \Rightarrow g \sim y$, because if $g = x + h_1$ and $x = y + h_2$, then $g = y + h_1 + h_2$

Each such coset class, is denoted by $[g]$, where g is any representative element of that class. For abelian groups the coset , which is the set of all such distinct equivalence classes, also has a group structure and hence is also called the *Quotient Group*. The notation used for such a Quotient group formed out of G and H is G/H. The identity element of this group denoted by [0] is simply the set of all elements of the type $(0 + h)$ which is just the full subgroup H.

In particular, for $G = \mathbb{Z}$ and $H = N\mathbb{Z}$, the elements of G/H are the equivalence classes $([0], [1], ...[N-1])$. For example with $N = 2$, we have the elements [0] and [1]. Where $[0] = (0, 2, -2, 4, -4, 6, -6, ...)$ and $[1] =$

$(1, -1, 3, -3, 5, -5, 7, -7, ...)$. Its easy to see that, $[0] + [1] = [1]$, $[1] + [1] \equiv 2[1] = [0]$. This Group is therefore in one - one correspondence with the group of two elements $(0, 1)$ under the composition law *addition modulo two*, $(a+b = c \ mod \ 2)$. This group is denoted as $\mathbb{Z}_2$. Similarly for general N, $N[1] = 0$ and the quotient group $\mathbb{Z}/N\mathbb{Z} \cong \mathbb{Z}_N$, which is the group of N integers-$(0, 1, ...N-1)$ under the composition rule *addition modulo* N, $(a + b = c \ mod \ N)$

In a general abelian group, the element obtained by adding the same element many times must be again an element of the group. If the group is finite, then it must be the case that for some positive integer N, $g+g+g...(Ntimes) \equiv Ng = 0$. Then, N is called the *order of the element* g. The set of all elements of finite order, in an abelian group forms a sub group, called the *torsion* subgroup and denoted by $t(G)$.

Problem 8.3.1: Show that $t(G)$ is a subgroup.

If all elements of a group is generated by a set of elements $(g_1, g_2, ...g_r)$, their inverse and the identity, the group is called a *finitely generated abelian group* $(FGAG)$. That is, $g = \sum_{i=1}^{r} N_i g_i; \ \forall \ g \in G$. Here, $N_i \in \mathbb{Z}$. If these generators are linearly independent, that is: $0 = \sum_{i=1}^{r} N_i g_i; \ \Rightarrow N_i = 0 \ \forall i$, then the $FGAG$ is called *free*. A free $FGAG$ with r independent generators is said to have rank r. A group generated by one element is called *cyclic*. The group $\mathbb{Z}$ is a cyclic group of order infinity, while $\mathbb{Z}_N$ is a cyclic group of finite order N.

Definition 8.3.1 (**Group Homomorphisms**)**.** Given two groups (not necessarily abelian) G and H, if there exists a map $f(G)$, which maps elements of G into elements of H, then such a map is called a *Homomorphism*.

1. $f(g) \in H; \ \forall g \in G$
2. $f(g_1 \circ g_2) = f(g_1) \bullet f(g_2); \ \forall g_i \in G$

Here the $\circ$ and the $\bullet$ denote the composition laws in G and H respectively.

The subset of elements (x) of the group G, which maps to the identity in H is called $ker \ f$. $(x \in kerf; \ f(x) = e)$. While $Im \ f$ is the subset of elements y in H which have been mapped from some element of G. ie: $y = f(g)$.

Problem 8.3.2: Show that both $ker \ f$ and $Im \ f$ are subgroups of G and H respectively.

It is clear that the map from G to $Im \ f$ is not in general an *isomorphism*. That is, the map need not be *one-one onto*. This is so because, more than one element in G, can map to same element in H. $f(g_1) = f(g_2)$. This means that $f(g_1) \bullet f^{-1}(g_2) = e \Rightarrow f(g_1 \circ g_2^{-1}) = e \Rightarrow g_1 \circ g_2^{-1} \in kerf$.

Returning to the case of abelian groups, this means that g_1 - $g_2 = h$, $h \in kerf$. This implies that any two elements which map to the same element in H belong to the same equivalence class, and therefore is a single element of the *Quotient* group $G/kerf$. This *Quotient* group has now elements in one

to one correspondence with the elements of $Im\ f$, so that the two groups are isomorphic. ie: $G/kerf \cong Imf$

A few applications of this result are given below.

1. Consider the homomorphism between the group $\mathbb{Z}$ and $\mathbb{Z}_2$. $f(2n) = 0$ and $f(2n+1) = 1$. The kernel of this map $f(x) = 0$ is simply all even integers: ie: $kerf = 2\mathbb{Z}$, so that we recover: $\mathbb{Z}/2\mathbb{Z} \cong \mathbb{Z}_2$.

2. Consider any group H which is a free $FGAG$ of rank r generated by the set $(g_1, ..., g_r)$. This can be thought of as the image of the homomorphism from the group $G = \bigoplus_{i=1}^{r} \mathbb{Z}$. The map being simply: $(N_1, ...N_r) \in G;\ f(g) = \sum_{i=1}^{r} N_i g_i \in H$. Since the group H which is the image of f is a free $FGAG$, the $kerf = \{0\}$. Then it follows that any free $FGAG$ of rank r, $H \cong \bigoplus_{i=1}^{r} \mathbb{Z}$.

3. By a similar argument, any cyclic group G with generator g of finite order N, $(x \in G;\ x = ng;\ Ng = 0)$ is isomorphic to $\mathbb{Z}_N$. To get this result, we simply start from the group $\mathbb{Z}$ and then take $f(n) = ng;\ n \in \mathbb{Z}$. Since $Ng = 0$, it follows that $kerf = N\mathbb{Z}$. Therefore, any cyclic group $G \cong \mathbb{Z}/N\mathbb{Z} \cong \mathbb{Z}_N$.

4. More generally, any $FGAG$ with r generators $(g_1, ...g_r)$ not necessarily free, is isomorphic to $\mathbb{Z} \bigoplus \mathbb{Z} \bigoplus \mathbb{Z}...(m\ times) \bigoplus \mathbb{Z}_{k_1} \bigoplus \mathbb{Z}_{k_2} ... \bigoplus \mathbb{Z}_{k_n}$, where the set of integers $(m, k_1, ...k_n)$ are fixed for a given $FGAG$, and $m + n = r$. m is known as the *rank* of the $FGAG$.

We sketch a proof of the last statement below.

Problem 8.3.3: For a $free\ FGAG$ of $rank = r$ with generators $(g_1, ..., g_r)$, the subset of elements generated by $(k_1 g_{i_1}, ...k_p g_{i_p})$ is always a subgroup. Where $(g_{i_i} ....g_{i_p})$ are any p generators from the set of r generators of the $free\ FGAG$ and $(k_1, ..., k_p)$ are all integers.

In fact, it turns out that *all* subgroups of a $FGAG$ can be generated this way. This means, that any subgroup H of a free $FGAG$ is isomorphic to $k_1\mathbb{Z} \bigoplus k_2\mathbb{Z} \bigoplus ... \bigoplus k_p\mathbb{Z}$.

We can now construct a homomorphism, as before, from the group $G = \bigoplus_{i=1}^{r} \mathbb{Z}$ to $H = FGAG$. The map being, again as before, $f(G) = f(N_1, ...N_r) = \sum_{i=1}^{r} N_i g_i$. Then the kernel of this map, being a subgroup of G,

- $kerf \cong k_1\mathbb{Z} \bigoplus k_2\mathbb{Z} \bigoplus ... \bigoplus k_p\mathbb{Z}$, for some set of positive integers $(p, k_1, ...k_p)$.

Then the desired result follows.

- $H = Imf \cong G/kerf = \mathbb{Z} \bigoplus \mathbb{Z} \bigoplus \mathbb{Z}...(m\ times) \bigoplus \mathbb{Z}_{k_1} \bigoplus \mathbb{Z}_{k_2} ... \bigoplus \mathbb{Z}_{k_p}$, with $m + p = r$

8.4 Nonabelian cases: Conjugacy class, cosets

In this section, we will start discussing the *non-abelian* case. We will introduce some definitions, some of which we have already seen in the abelian case, and some which are interesting and non-trivial only when the group is non-abelian. For the non abelian groups, we will use the notation gh for denoting the composition $g \circ h$.

Definition 8.4.1 (*conjugacy classes*)**.** All the elements of a group can be placed in various conjugacy classes. Two elements $(x,\ y)$ of a group G are said to be conjugate to each other, if there exists some element g of the group G, such that $gxg^{-1} = y$. The conjugation relation is an equivalence relation.

Problem 8.4.1: Show that conjugation is an equivalence relation

All elements which are so related belong to the same conjugacy class. It is trivial to see that for abelian groups, there are no non-trivial conjugacy classes. Every element is conjugate only to itself.

- For the group of rotation matrices in three dimensions, all rotations by the same angle but about different axis fall into the same conjugacy class.
- For the group of two dimensional special unitary matrices, all matrices having the same trace are conjugate to each other.

Definition 8.4.2 (*coset*)**.** Given any element (g) in G, and a subgroup H, one can form a set denoted as $[g]$, which has elements $(g, gh_1, gh_2, ..)$; $\forall h_i \in H$. This subset of elements is denoted as $gH \equiv [g]$. A coset is a set whose elements are all such distinct classes $[g_1], [g_2]$.... Its usually denoted as G/H.

Problem 8.4.2: Show that if there are two such classes $[g_1]$ and $[g_2]$, then either they share all elements or none.

The number of elements in each such class $[g]$ is equal to $|H|$- the order of H. Therefore $|G|/|H|$ = number of distinct coset classes. So it follows that for any subgroup H of G, $|H|$ factorizes $|G|$.

Cosets have been introduced in the context of abelian groups. Unlike in the abelian case however, cosets do not form a group in general. This is because the coset classes do not satisfy the composition rule for groups. ie: All elements of $[g_1][g_2] \neq [g_1 g_2]$. This is so because due to the non-abelian nature of the group, $g_1 h_1 g_2 h_2 \neq g_1 g_2 h_3$ for any $g_i \in G$ and $h_i \in H$. Another way of expressing this is $g_1 H g_2 H \neq g_1 g_2 H$, A coset would have been a Group, iff $(g_i H g_i^{-1} = H \ \forall\ i)$. This brings us to the definition of a *normal subgroup*.

Definition 8.4.3 (*Normal Subgroup* and *Quoient Groups*)**.** A subgroup which satisfies the property that $gHg^{-1} = H$; $\forall g \in G$ is called a normal subgroup.

This means that the conjugate of any element of H is also in H. If we construct a coset out of a normal subgroup, then the coset so formed is a Group, called the Quotient Group. For the abelian case, all subgroups are trivially normal.

Problem 8.4.3: Show that $kerf$, where f is a homomorphism between two groups G and H is a normal subgroup of G

8.5 Commutator subgroup and abelianization

As an example of a normal subgroup, we consider the *commutator* sub group. Given a group G, the commutator subgroup $[G, G]$ is the subgroup which is generated by all the commutator elements $[a, b]$. So, the elements of the subgroup $[G, G]$ are $(x \in [G, G];\ x = [a_1, b_1]^{n_1}[a_2, b_2]^{n_2}...[a_i, b_i]^{n_i}...)$. This automatically implies that the product of two commutator elements is also a commutator element. All such elements clearly form a group. In fact it forms a Normal subgroup. This follows from the following two properties.

- $[a, b]^{-1} = b^{-1}a^{-1}ba \equiv [b, a]$
- $g[a, b]g^{-1} = [a_c, b_c];\quad a_c = gag^{-1}$ and $b_c = gbg^{-1}$

This further means that

1. $\forall\ x \in [G, G],\ x^{-1} \in [G, G]$
2. $\forall\ x \in [G, G]\ and\ \forall g \in G,\ gxg^{-1} \in [G, G]$

- The first property along with the fact that the identity is also a commutator element implies that $[G, G]$ is a subgroup.
- From the second property, it follows that $[G, G]$ is a normal subgroup.

The Quotient group $G/[G, G]$ is always *abelian*. This means that $g_1[G, G]g_2[G, G] = g_2[G, G]g_1[G, G]$. As a check of this statement, do the following problem.

Problem 8.5.1: Show $g_1[a_1, b_1]g_2[a_2, b_2] = g_2[a_1^c, b_1^c]g_1[g_1, g_2][a_2, b_2]$ where $a^c = g_2^{-1}g_1ag_1^{-1}g_2$ and similarly for b^c.

This process of constructing an abelian group from a non-abelian group is called *abelianization*.

8.6 Examples of Groups

In this and the following section, we will consider various examples of groups.

8.6.1 The Quaternionic Group

The Quaternionic group is the set of elements denoted as $(1, -1, \mathbf{i}, \mathbf{j}, \mathbf{k}, -\mathbf{i}, -\mathbf{j}, -\mathbf{k})$, with the composition rule: $\mathbf{ii} = \mathbf{jj} = \mathbf{kk} = -1$ and $\mathbf{ij} = -\mathbf{ji} = \mathbf{k},\ \mathbf{jk} = -\mathbf{kj} = \mathbf{i},\ \mathbf{ki} = -\mathbf{ik} = \mathbf{j}$. $(1,\ -1)$ commute with all elements. In particular, 1 is the identity element and -1 has order two. (**-i**, **-j**, **-k**) are the inverse elements of (**i**, **j**, **k**) respectively

One representation of these elements are as 2×2 σ matrices and the identity matrix.

Problem 8.6.1: Check that the following representations of the elements of the Quaternionic group, indeed satisfy the group composition rules.
$\mathbf{i} \equiv i\sigma_1$, $\mathbf{j} \equiv i\sigma_2$, $\mathbf{k} \equiv -i\sigma_3$, $\pm 1 \equiv \pm \mathbb{I}_{2\times 2}$

Problem 8.6.2: Write down all the conjugacy classes of the group.

Problem 8.6.3: Write down all the subgroups of the Quaternionic group. Check that the order of these subgroups are indeed factors of the order of the Quaternionic group.

Problem 8.6.4: Check which of these subgroups are normal subgroups.

Problem 8.6.5: Using the definition of the commutator subgroup show that for the Quaternionic group, the commutator subgroup is $\{1, -1\}$.

Problem 8.6.6: *Abelianization of the Quotient Group*: Since the commutator subgroup is a Normal subgroup, the coset formed out of it $G/[G,G]$ is a Quotient group. We will now check that it is abelian and identify which abelian group it is.

- Find the cosets gH where H is the commutator subgroup. Show that they are $(1,-1)$, $(\mathbf{i},-\mathbf{i})$, $(\mathbf{j},-\mathbf{j})$, $(\mathbf{k},-\mathbf{k})$
- Denoting each of these classes as $[1],[i],[j],[k]$ respectively, show that they form an abelian group. Show explicitly that the composition rule of the Quotient group is $[i][i] = [j][j] = [k][k] = [1]$, $[i][j] = [j][i] = [k]$, $[j][k] = [k][j] = [i]$

  and $[j][i] = [i][j] = [k]$
- Using the notation for abelian groups, where we use $+$ for composition, 0 for identity, so that $[1] = 0$, write the above rules as: $2[i] = 2[j] = 2[k] = 0$ and $[i] + [j] = [k]$, $[j] + [k] = [i]$, $[i] + [k] = [j]$.
- Simplify the above and show that not all are independent relations. Show that the independent relations are:
- $2[i] = 2[j] = 0$ and $[i] + [j] = [k]$.
- Hence show that the Quotient group obtained by abelianization of the Quaternionic group is isomorphic to $\mathbb{Z}_2 \oplus \mathbb{Z}_2$

8.6.2 Rotations of a rigid body

Now we will consider examples of continuous groups. The first example we consider is that of continuous rotations of a rigid body. These are characterized by 3×3 marices which satisfy

$$RR^\tau = R^\tau R = I_{3\times 3}, \text{ with } \det[\mathrm{R}] = +1 \tag{8.1}$$

Here R^τ is the transpose matrix. Its easy to check, the set of all such matrices form a group. This is the group $SO(3)$. Here R is a three dimensional real

matrix. The above equation reduces the number of independent elements of the matrix to just three. Under rotations, the length of a vector **V**remains invariant. ie $\mathbf{RV} \cdot \mathbf{RV} = \mathbf{V} \cdot \mathbf{V}$. Let W be an eigenvector of R with eigenvalue λ. Physically we expect that under rotation the vector remains real, so that we take λ to be real. Then,

$$\mathbf{RW} \cdot \mathbf{RW} = \lambda^2 \mathbf{W} \cdot \mathbf{W} = \mathbf{W} \cdot \mathbf{W}$$

Also the determinant condition means that $\lambda_1\lambda_2\lambda_3 = 1$. Taken together, this means that there exists atleast one eigenvector with eigenvalue $= +1$. Therefore this eigenvector does not change under rotations. This eigen-direction is called the *axis of rotation*.

Problem 8.6.7: Choose a basis, where the $z - axis$ is the *axis of rotation*, and the x and y axes are any two orthonormal directions. In this basis show, using equation(1), that R can be written as follows:

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & \sin(\theta) \\ 0 & -\sin(\theta) & \cos(\theta) \end{bmatrix}$$

The angle θ is called the *angle of rotation*. So that any Rotation matrix is fully characterized by an *angle and an axis of rotation*. It follows from this form of R that $cos(\theta) = \frac{\mathrm{tr}R-1}{2}$. So the angle of rotation is given in terms of the trace of R.

So any rotation is characterized by $(\hat{\mathbf{n}},\ \theta)$. We can now parametrize this space as follows.

- Let $\hat{\mathbf{n}}$ be any direction on the sphere. This means that we distinguish between the axis of rotations $\hat{\mathbf{n}}$ and -$\hat{\mathbf{n}}$.
- Then we measure θ in counter-clockwise direction of $\hat{\mathbf{n}}$. This way, we can restrict the range of θ to values $0 \leq \theta < \pi$. This is because, in this way of parametrizing, rotation by angle $\pi + \theta$ in counter clockwise direction to $\hat{\mathbf{n}}$, is same as counter clockwise rotation by $\pi - \theta$ around -$\hat{\mathbf{n}}$.
- In particular, rotation by π around $\hat{\mathbf{n}}$ is same as rotation by π around -$\hat{\mathbf{n}}$.
- We can now geometrically represent the space of all $SO(3)$ matrices as points inside a ball of size π, where the radial direction is the $\hat{\mathbf{n}}$, and the length along this direction is θ and because of the identification mentioned in the previous bullet, the surface of this ball has its diametrically opposite points identified. Therefore the configuration space of $SO(3)$ Matrices is in one to one correspondence with the points of a $Ball/\mathbb{Z}_2$, where the $\mathbb{Z}_2$ action identifies the opposite points on the surface of the ball.

There is another way in which we can represent these $SO(3)$ matrices. Since these represent continuous rotations, we can build a finite rotation as a

series of several smaller rotations, in the limit where each rotation is taken to be infinitesimal and the number of such rotations is taken to be infinity. ie:

$$R = \Pi_{i=1}^{N} R(\epsilon), \;\; lt : (N \to \infty, \; \epsilon \to 0, \;\; N\epsilon = 1). \tag{8.2}$$

Here $R(\epsilon)$ is a infinitesimal rotation. An infinitesimal rotation can be parametrized as

$$R(\epsilon) = \mathbb{I}_{3\times 3} + \epsilon \mathbf{T} \tag{8.3}$$

Using the equation $RR^{\tau} = \mathbb{I}_{3\times 3}$, we can show that $\mathbf{T}$ is an antisymmetric matrix. Any antisymmetric matrix in three dimensions may be parametrized in terms of three *generators* $\{T_i\}$, $\mathbf{T} = \sum_{i=1}^{3} t_i T_i$, where t_i are independent parameters, and

$$T_1 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad T_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} \quad T_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$$

These generators satisfy the following algebra: $[T_i, T_j] = \epsilon_{ijk} T_k$, where$[A, B] = AB - BA$ and ϵ_{ijk} is the completely antisymmetric tensor with entries ± 1. $\epsilon_{123} = 1$. We can now build a finite rotation, as a product of such infinitesimal rotations.

$$R(t_1, t_2, t_3) = \lim_{N\to\infty} \left(\mathbb{I}_{3\times 3} + \frac{1}{N} \sum_{i=1}^{3} t_i T_i \right)^N = \exp\left(\sum_{i=1}^{3} t_i T_i \right) \tag{8.4}$$

It is more conventional to express

$$R(t_i) = \exp\left(i \sum_{i=1}^{3} \theta_i L_i \right) = \exp(i\theta \hat{\mathbf{n}} \cdot \mathbf{L}) \tag{8.5}$$

Where $L_i = iT_i$, $\theta_i = -t_i$, $(\theta = \sqrt{t_1^2 + t_2^2 + t_3^2})$, $\theta \; \hat{\mathbf{n}} = (\theta_1, \theta_2, \theta_3)$ and $\mathbf{L} = (L_1, L_2, L_3)$.

The L_i's are Hermitian matrices and satisfy the following algebra: $[L_i, L_j] = i\epsilon_{ijk} L_k$. This algebra is identical to that satisfied by the two dimensional σ matrices. As we will see in the next section, the σ matrices are the generators of the $SU(2)$ matrices in two dimensions. This isomorphism between the two algebras is because of the fact that the $SU(2)$ and $SO(3)$ groups are related by a homomorphism, which we will discuss in the next section.

8.7 $SU(2)$ and $SO(3)$

In this final section, we will discuss the group of two dimensional $SU(2)$ matrices and write down a homomorphism from this group to the group of $SO(3)$ matrices.

8.7.1 $SU(2)$ matrices

$SU(2)$ matrices in two dimensions are complex matrices with determinant $+1$ and satisfying

$$UU^\dagger = \mathbb{I} \tag{8.6}$$

Its easy to show that any such matrix can be parametrized as

$$U = \begin{pmatrix} \alpha & \beta \\ -\beta^* & \alpha^* \end{pmatrix} = \begin{pmatrix} a_0 + ia_3 & a_2 + ia_1 \\ -a_2 + ia_1 & a_0 - ia_3 \end{pmatrix} = a_0\mathbb{I} + i\mathbf{a}\cdot\boldsymbol{\sigma}$$

where, $\alpha = a_0 + ia_3$, $\beta = a_2 + ia_1$ are complex numbers and $(a_0^2 + \mathbf{a}\cdot\mathbf{a} = 1)$ $\mathbf{a} = (a_1, a_2, a_3)$ and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$

Problem 8.7.1: Show it.

Therefore it follows that the configuration space of all $SU(2)$ matrices are in one-one correspondence with the points on a S^3.

Parametrizing $a_0 = \cos(\theta)$ and $|\mathbf{a}| = \sin(\theta)$, we can write: $U = e^{i\theta\hat{\mathbf{n}}\cdot\boldsymbol{\sigma}}$, where $\hat{\mathbf{n}} = \frac{\mathbf{a}}{|\mathbf{a}|}$. So the σ matrices are the generators of the $SU(2)$ group in the same sense that $\mathbf{L}_i$'s were of the $SO(3)$ group. Also, the $SU(2)$ and $SO(3)$ generators satisfy the same algebra. We will now see that the two groups are homomorphic.

Homomorphism from $SU(2)$ to $SO(3)$

We start with the observation that for every 3 dimensional vector one can write a corresponding two dimensional traceless, hermitian matrix as follows:

$$\mathbf{V} = \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix} \Rightarrow \upsilon = \begin{pmatrix} V_3 & V_- \\ V_+ & -V_3 \end{pmatrix} = \mathbf{V}\cdot\boldsymbol{\sigma}, \quad V_\pm = V_1 \pm iV_2$$

The length of the vector is related to the determinant of the matrix as: $|V|^2 = -\det(\upsilon)$.

After rotation vector $\mathbf{V}$ goes over to vector $\mathbf{V}' = \mathbf{RV}$ with the same length. $|V'| = |V|$. This means that the corresponding two dimensional matrices, υ and $\upsilon' = \mathbf{V}'\cdot\boldsymbol{\sigma}$ must have the same determinant. Since both these matrices are Hermitian, have the same trace and same determinant, it must be that they are related by a unitary transformation as given below:

$$\upsilon' = U_R \upsilon U_R^\dagger \tag{8.7}$$

In fact we can take the U_R to be $SU(2)$ matrices, because the $det(U)$ and $det(U^\dagger)$ cancel each other in the RHS of the above equation. So, we can conclude that the rotation of a vector $\mathbf{V}$ in three dimensions, corresponds, in two dimensions, to acting on the matrix υ by a $SU(2)$ matrix and its inverse from the left and right respectively.

We can use this equation to find an explicit map from $SU(2)$ to the R matrices. We first note that:

$$v' = \sum_i V_i' \sigma_i = \sum_i R_{ij} V_j \sigma_i = (a_0 \mathbb{I} + i\mathbf{a}\cdot\boldsymbol{\sigma})(\mathbf{V}\cdot\boldsymbol{\sigma})(a_0\mathbb{I} - i\mathbf{a}\cdot\boldsymbol{\sigma}), \tag{8.8}$$

where R_{ij}'s are the components of the rotation matrix.

One can then simplify the LHS further by using the following identities:

$$(\mathbf{V}\cdot\boldsymbol{\sigma})(\mathbf{W}\cdot\boldsymbol{\sigma}) = (\mathbf{V}\cdot\mathbf{W})\mathbb{I} + i(\mathbf{V}\times\mathbf{W})\cdot\boldsymbol{\sigma}, \tag{8.9a}$$

$$\mathbf{V}' = a_0^2\mathbf{V} + \mathbf{a}(\mathbf{V}\cdot\mathbf{a}) + (\mathbf{V}\times\mathbf{a})\times\mathbf{a} + 2a_0(\mathbf{V}\times\mathbf{a}). \tag{8.9b}$$

We then get

$$R_{ij}V_j = \Big((a_0^2 - \mathbf{a}\cdot\mathbf{a})\delta_{ij} + 2a_ia_j + 2a_0\epsilon_{ijk}a_k\Big)V_j. \tag{8.10}$$

Then comparing the coefficient of V_j on both sides we get the desired explicit form of R_{ij}

$$R = \begin{pmatrix} a_0^2 + a_1^2 - a_2^2 - a_3^2 & 2(a_1a_2 + a_0a_3) & 2(a_1a_3 - a_0a_2) \\ 2(a_1a_2 - a_0a_3) & a_0^2 + a_2^2 - a_1^2 - a_3^2 & 2(a_2a_3 + a_0a_1) \\ 2(a_1a_3 + a_0a_2) & 2(a_2a_3 - a_0a_1) & a_0^2 + a_3^2 - a_1^2 - a_2^2 \end{pmatrix},$$

where as before, $a_0^2 + \mathbf{a}\cdot\mathbf{a} = 1$. $R(a_0, \mathbf{a})$ is therefore the desired map which takes an element of $SU(2)$ labelled by $(a_0, \mathbf{a})$ to an element of $SU(3)$. We can easily show that the kernel of this map is $a_0^2 = 1,\ \mathbf{a} = \mathbf{0}$. ie: $R(\pm 1, \mathbf{0}) = \mathbb{I}$. Therefore the kernel is simply $\pm\mathbb{I}_{2\times 2}$. This is the subgroup $\mathbb{Z}_2$. $\mathbb{Z}_2$ is called the *centrer* of $SU(2)$. In general, the *centre* of a group is the subset of all elements which commutes with every other element. For $SU(N)$ group the centre is $\mathbb{Z}_N$. So we get the relation that $SU(2)/\mathbb{Z}_2 \cong SO(3)$. Thus there is two-one relation between elements of the $SU(2)$ and $SO(3)$.

We can further identify the angle and axis of rotation with the φ and $\hat{\mathbf{n}}$ appearing in the parameterization of the $SU(2)$ matrix $U_R = e^{i\varphi\hat{\mathbf{n}}\cdot\boldsymbol{\sigma}}$ as follows: $\theta = 2\varphi$ and angle of rotation is same as $\hat{\mathbf{n}}$. This is left as an exercise below. This relation between the angles θ and φ is related to the existence of half integral spin particles in quantum mechanics— a direct consequence of the $SU(2)$ representation of rotations.

Problem 8.7.2: Prove the identity given by Eq. (8.9a).

Problem 8.7.3: Using the identity of Eq. (8.9a), show Eq. (8.9b).

Problem 8.7.4: By choosing $\mathbf{V} \parallel \mathbf{a}$, show that $\mathbf{V}' = \mathbf{V}$, and thus prove that $\hat{\mathbf{n}}$ is indeed the direction of the axis of rotation.

Problem 8.7.5: By choosing $\mathbf{V} \perp \mathbf{a}$, and using the fact that the angle of rotation $\cos(\theta) = \mathbf{V}'\cdot\mathbf{V}/|V|^2$, show that $\theta = 2\varphi$. So that U_R can be parametrized directly in terms of the axis and angle of rotation as $U_R = e^{i(\theta/2)\hat{\mathbf{n}}\cdot\boldsymbol{\sigma}}$.

8.8 Conformal transformations

Conformal transformations are transformations which preserve the angle between curves. The angle between two curves at a point of intersection p is the angle that the tangents to the two curves make at p. It is given by

$$\cos(\theta) = \frac{d\mathbf{s}_1.d\mathbf{s}_2}{|d\mathbf{s}_1||d\mathbf{s}_2|},$$

where $d\mathbf{s}_1$ and $d\mathbf{s}_2$ are infinitesimal vectors along the tangent directions at the point p. It is clear, from the expression of the angle that it is invariant under:

- global translations ($x \to x + a$),
- global rotations(or Lorentz transformations, depending on whether the space is Euclidean or Minkowskian)($x_i \to R_{ij}x_j$)
- global scaling of coordinates ($x \to \lambda x$). In this case, Both $d\mathbf{s}_1.d\mathbf{s}_2$ and $|d\mathbf{s}_1||d\mathbf{s}_2|$ scale by a factor of λ^2, but these factors cancel in the denominator and numerator terms.
- In fact, the expression would be invariant if the numerator and denominator both scaled locally by the same function $f(x)$, which could then cancel among each other. This would mean demanding that $ds_1.ds_2 \to f(x)ds_1.ds_2$ and $|ds_1||ds_2| \to f(x)|ds_1||ds_2|$. We can now ask, what kind of coordinate transformations $x \to x'(x)$ can achieve such a local scaling.

To find such coordinate transformations, it is useful to consider the invariant distance $dS^2 = g_{ij}dx^i dx^j$. The distance dS^2 is defined to be invariant under any general coordinate transformations, which means that under any coordinate transformations, when

$$dx^i \to \frac{\partial x'^i}{\partial x^j}dx^j, \quad \text{the metric} \quad g_{ij} \to \frac{\partial x^k}{\partial x'^i}\frac{\partial x^l}{\partial x'^j}g_{kl},$$

so that the statement that $ds_1.ds_2 \to f(x)|_p ds_1.ds_2$ is the same as demanding that the metric $g_{ij} \to \frac{1}{f(x)}g_{ij}$.

- So a conformal transformation is a coordinate transformation $x'(x)$ under which $g_{ij}(x) \to g'_{ij}(x') = h(x)g_{ij}(x)$ where $h(x)$ is any local function of the coordinates.
- It is easy to see that these transformations form a group. Under two such transformations, $x \to x'(x) \to x''(x'(x))$, the metric will transform as

$$g_{ij}(x) \to g'_{ij}(x'(x)) = h'(x)g_{ij}(x),$$

$$g'_{ij}(x') \to g''_{ij}(x'') = h''(x')g'_{ij}(x') = h''(x'(x))h'(x)g_{ij}(x) = \hat{h}(x)g_{ij}(x),$$

where $\hat{h}(x) = h''(x'(x))h'(x)$. The inverse coordinate transformation, assuming it exists, is the inverse element while the identity is $x'(x) = x$.

We study these transformations by first looking at the infinitesimal forms of coordinate transformations. So, we look at $x'(x) = x + \epsilon(x)$ and $h(x) = 1 + g(x)$. Under these transformations, the metric change is as follows:

$$\frac{\partial x^a}{\partial x'^i}\frac{\partial x^b}{\partial x'^j} g_{ab}(x) = h(x) g_{ij}(x), \tag{8.11}$$

which for infinitesimal transformations becomes

$$(\delta^a_i - \partial_i \epsilon^a)(\delta^b_j - \partial_j \epsilon^b) g_{ab} = (1 + g(x)) g_{ij}(x) \tag{8.12}$$

We will be working mostly on flat space, so that we can take $g_{ij} = \delta_{ij}$. Then the above equation simplifies to:

$$\partial_i \epsilon_j + \partial_j \epsilon_i = -g(x)\delta_{ij} \tag{8.13}$$

We will first analyze these equations in two dimensions, where it is simple. In two dimensions, i, j take values $1, 2$. Then the above two equations are:

$$\partial_1 \epsilon_1 = \partial_2 \epsilon_2, \quad \text{and} \quad \partial_1 \epsilon_2 = -\partial_2 \epsilon_1. \tag{8.14}$$

These are just the Cauchy-Riemann conditions, and it means that in two dimensions, under a conformal transformations, $(z = x + iy) \to f(z)$ and similarly for $\bar{z} = x - iy \to \bar{f}(\bar{z})$. So any holomorphic transformation is a conformal transformation in two dimensions.

We now try to construct the generators of these conformal transformations. To find the generators, let us first recall how it is done for the case of translations and rotations.

- The generator of translations is simply the derivative operator $\frac{\partial}{\partial x^i}$. Any function $f(x + a) = e^{a\partial_x} f(x)$. For a infinitesimal,

  $$f(x + a) = f(x) + a\partial_x f(x) = (1 + a\partial_x) f(x).$$

  In particular for $f(x) = x$, $x' = x + a = (1 + a\partial_x)x$.

- Similarly for the case of two dimensional infinitesimal rotations $x' = x + y\epsilon$ and $y' = y - \epsilon x$, we know that the generator is simply $(x\partial_y - y\partial_x)$, as for any function

  $$f(x', y') = f(x + \epsilon y, y - x\epsilon) = f(x, y) + \epsilon(y\partial_x - x\partial_y) f(x, y).$$

  In particular, for $f(x) = x$ we have $x' = (1 + \epsilon(y\partial_x - x\partial_y))x$ and similarly $y' = (1 + \epsilon(y\partial_x - x\partial_y))y$.

So, now we can use the same method to find the generators for the conformal transformations in two dimensions. We first write

$$z' = z + \epsilon(z) = \left[1 + \left(\sum_{n=-\infty}^{\infty} \epsilon_n z^{n+1}\right)\partial_z\right] z, \tag{8.15}$$

where in the RHS, we have done a Laurent expansion of $\epsilon(z)$ and ϵ_n are its infinitesimal coefficients. So the generators are $L_n = z^{n+1}\partial_z$. We have similar generators from the anti holomorphic transformations on $\bar{z}$, viz., $\tilde{L}_n = \bar{z}^{n+1}\partial_{\bar{z}}$. These generators, satisfy the following algebra:

$$[L_n, L_m] = (m-n)L_{m+n}, \quad [\tilde{L}_n, \tilde{L}_m] = (m-n)\tilde{L}_{m+n} \tag{8.16}$$

These are the generators of the conformal transformations in two dimensions. In $d > 2$ however, the analysis of Eq. (8.12) will lead to a finite set of symmetries. The analog of these in two dimensions are the ones generated by the following subset, which form a subalgebra— the set generated by (L_{-1}, L_0, L_{+1}) and $(\tilde{L}_{-1}, \tilde{L}_0, \tilde{L}_{+1})$. We will try to see what finite transformations these correspond to.

1. $L_{-1} + \tilde{L}_{-1} = \partial_z + \partial_{\bar{z}} = \partial_x$,
2. $i(L_{-1} - \tilde{L}_{-1}) = \partial_y$,
3. $L_0 + \tilde{L}_0 = z\partial_z + \bar{z}\partial_{\bar{z}} = x\partial_x + y\partial_y = r\partial_r = \partial_{\ln(r)}$, (r is the usual radial coordinate)
4. $L_0 - \bar{L}_0 = x\partial_y - y\partial_x$,
5. $L_1 = z^2\partial_z = -\partial_{z^{-1}}$.

From the above, we can conclude that:

1. $L_{-1} \pm \tilde{L}_{-1}$ generate translations,
2. $L_0 + \tilde{L}_0 = \partial_{\ln(r)}$, from analogy with translations, generates translations along $\ln r$. But translations along $\ln r$ corresponds to constant scaling of r and therefore of x, y. So this is the generator for constant scaling.
3. $L_0 - \tilde{L}_0$ is the generator of rotation.
4. $L_1 = -\partial_{z^{-1}}$, again generates translations in $\frac{1}{z}$. So it takes

$$\frac{1}{z} \to \frac{1}{z} + a \Rightarrow z \to \frac{z}{1+az},$$

where a is complex. Written in terms of $\mathbf{X} = (x, y)$ the transformation is

$$\mathbf{X}' = \frac{\mathbf{X} + \mathbf{a}(|X|^2)}{1 + 2\mathbf{a}\cdot\mathbf{X} + |a|^2|X|^2}, \quad \mathbf{a} = (a_1, a_2). \tag{8.17}$$

This is known as the *Special Conformal transformation(SCT)*

5. Taken together the translations, rotations, scaling and special conformal transformations together form a group — the global conformal group. For any $d \geq 2$ this group is isomorphic to $SO(d+1, 1)$, which is the Lorentz group in $d+2$ dimensions. For $d > 2$, this is the full conformal group, while in $d = 2$ this group of transformations is a part of the infinite dimensional transformations, that we discussed.

References

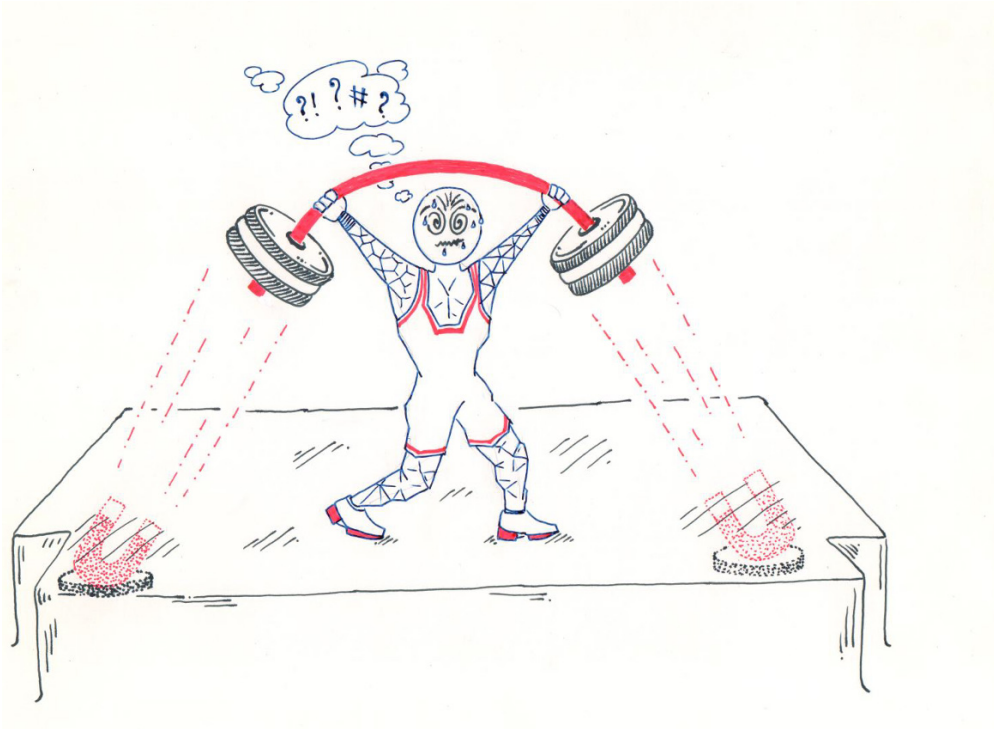
[1] For the discussion on abelian groups, I have mostly followed the book by M. Nakahara, *Geometry, Topology and Physics* (Taylor & Francis, Boca Raton, FL, USA, 2003). *Chapter-3*, *section- 3.1*

[2] The discussion on $SO(3)$ and the homomorphism with $SU(2)$ is discussed in the classical mechanics textbook by J. V. José and E. J. Saletan, *Classical Dynamics: A contemporary approach*, (Cambridge U. Press, 1998). *chapter-8*, *section-8.2.1, 8.4*.

[3] The discussion on conformal transformations is available in many places. One place is the text book by P. Francesco, P. Mathieu and D. Sénéchal, *Conformal Field Theory* (Springer, New York, 2012).

Part II

Physics Problems



9

Use of Topology in physical problems

Somendra M. Bhattacharjee

Some of the basic concepts of topology are explored through known physics problems. This helps us in two ways, one, in motivating the definitions and the concepts, and two, in showing that topological analysis leads to a clearer understanding of the problem. The problems discussed are taken from classical mechanics, quantum mechanics, statistical mechanics, solid state physics, and biology (DNA), to emphasize some unity in diverse areas of physics.

It is the real Euclidean space, $\mathbb{R}^d$, with which we are most familiar. Intuitions can therefore be sharpened by appealing to the relevant features of this known space, and by using these as simplest examples to illustrate the abstract topological concepts. This is what is done in this chapter.

9.1 The Not-so-simple Pendulum

An ideal pendulum is our first example. It is not necessarily a simple harmonic oscillator (SHO), though the small amplitude motion can be well approximated by a linear oscillator. This difference is important for dynamics, and a topological analysis brings that out.

The planar motion of a pendulum in the earth's gravitational field is described by a generalized coordinate q where q is the angle θ as in Fig. 9.1.

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_9

9.1.1 Mechanics

The equation of motion (with all constants set to 1) can be written as a second order equation or two first order equations involving the momentum p as

$$\ddot{q} + \sin q = 0, \text{ or } \begin{cases} \dot{q} &= p, \\ \dot{p} &= -\sin q, \end{cases} \tag{9.1}$$

where a dot represents a time derivative, and the conserved energy as

$$E = \frac{1}{2}\,\dot{q}^2 + (1 - \cos q). \tag{9.2}$$

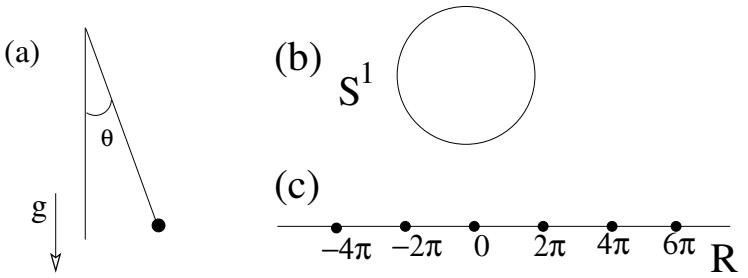


Figure 9.1: Planar pendulum: (a) A bob of mass $m(=1)$ is suspended by a massless rigid rod of length $L(=1)$ in a uniform gravitational field with $g(=1)$ as the acceleration due to gravity. The generalized coordinate q is angle θ. The configuration space for θ is (b) a circle S^1 or, (c) the real axis with equivalent points $x = x + 2n\pi, n \in \mathbb{Z}$. Here $\mathbb{Z}$ represents the set of all integers, positive, negative, 0. We may choose $q = x$ to be the length of the arc along the circular contour. A simple harmonic oscillator corresponds to $q = x \in R$ without any equivalent point.

Let us make a list of some of the relevant results known from mechanics.

1. The potential energy has minima at $q = 2n\pi$, and maxima at $q = (2n + 1)\pi$, $n = 0, \pm 1, ...$, i.e., $n \in \mathbb{Z}$. These represent the *stable* and the *unstable* equilibrium points.

2. The stable motion for small energies, $E < 2$, are oscillations around the minimum energy point $q = 0$. Let's call these *type-O* motion.

3. For larger energies, $E > 2$, the motion consists of rotations in the vertical plane, clockwise or anticlockwise. These are *type-R* motion.

4. There is a very special critical one that separates the above two types, viz., the case with $E = 2$, when E is equal to the potential energy at the topmost position ($q = \pi$). Let's call it *type-C*.

The strangeness of the critical one is its infinite time period[1]. Since most of the time is spent near the top, it looks like an inverted pendulum at the unstable equilibrium point.

5. For all problems of classical mechanics or statistical mechanics, there are two spaces to deal with, the *configuration space* for the set of values taken by the degrees of freedom, and the *phase space*, where the configuration space is augmented by the set of values of the momenta. *What sort of "spaces" are these*?

6. That there are three different classes of orbits cannot be overemphasized. The equation of motion is time reversible under $t \to -t, q \to q, p \to -p$. This time reversibility is respected by the type-O motion, but not by type-R because a right circular motion would go over to a left one. For type-R, the symmetry is explicitly broken by the initial conditions, which, however, do not play any crucial role for type-O.

7. As a coupled first order equations, Eq. (9.1), has fixed points at $q = n\pi, p = 0$, which are *centres* for *even* n but *saddle points* for *odd* n.

9.1.2 Topological analysis: Teaser

A topological analysis of the motion would be based on possible continuous deformations of one solution or trajectory to the other, without involving any explicit solution of Eq. 9.1.

Why deform? This is tantamount to asking whether there is any qualitative change in motion, as opposed to a detailed quantitative one, for a small change in energy or in the initial conditions. A small change in the amplitude of vibration due to a small change in energy is like a continuous deformation of the trajectory. In topology, the rule of deformation is to bend or stretch in whatever way we want except that neither distinct points be identified (no gluing) nor any tearing be done.[2] Such transformations are called continuous transformations.[3]

If the energy is changed continuously from E_0 to E_1 by defining a continuous function $E(\tau)$, say $E(\tau) = E_0 + (E_1 - E_0)\tau$ with $\tau \in [0, 1]$ do the trajectories in phase space get deformed continuously?[4] The answer is not necessarily yes. This is where the global properties of the phase space or the configuration space come into play. The continuous deformations then help us both in characterizing the phase space and in classifying the trajectories. We show that the three

[1] This can be seen by integrating Eq. 9.2 for the time taken to go from $q = 0$ to $q = \pi$ as $\int_0^\pi \sqrt{\sec(q/2)}\ dq \to \infty$ (from $q \to \pi$).

[2] Why these restrictions? We shall see in Sec. 9.2.1 that "gluing" is an equivalent relation that changes the space. Similarly, tearing changes the space by redefining the neighbourhoods at the point of cut.

[3] See Sec. 9.6 for a discussion on continuous functions

[4] This is a virtual change and not a real time-dependent change of energy. At every τ, the pendulum executes the motion for that energy

types of motion, O, R and C belong to three different classes of curves in the appropriate phase space.

The topological analysis is done by identifying (i) the configuration space and the phase space, (ii) the possible trajectories on these spaces, and for Hamiltonian systems, (iii) the constant energy "surface" (or manifolds) for possible real motions. We do these qualitatively first and then discuss some of the features in more detail.

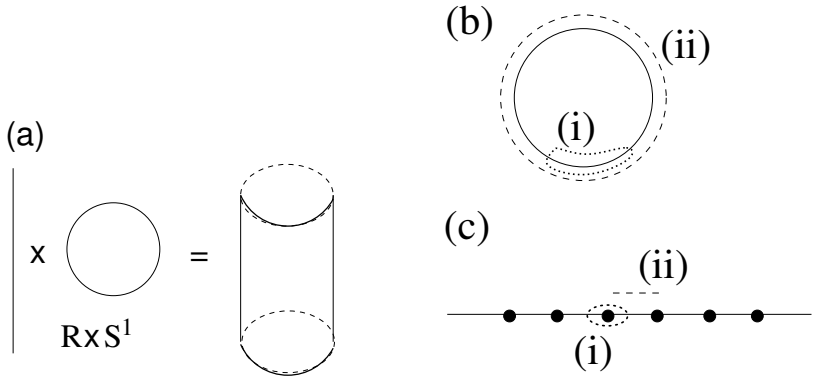


Figure 9.2: (a)The direct product phase space is a cylinder. (b) Type-1 and type-2 trajectories are marked (i) and (ii) in (b) and (c). The trajectory going completely around the circle in (b) maps to an open line in (c). Consequently any closed loop on $\mathbb{R}$ can be shrunk to a point. The real line is the *universal cover* of S^1.

Configuration space and phase space:

The first step is to construct the configuration and the phase spaces. The values taken by the degrees of freedom defines a set. One then defines a topology on it by defining the open sets, thereby generating the topological space to be called the configuration space. The phase space is obtained by adding the momenta variables to the configuration space.

For a pendulum, with q an angle, the configuration space is a circle S^1. It is also possible to represent the configuration space as the real line $\mathbb{R}$ with an identification of all points $x = x + 2n\pi$, $n \in \mathbb{Z}$ as in Fig. 9.1. The momentum, being real, trivially belongs to $\mathbb{R}$, the real line. The phase space is therefore $S^1 \times \mathbb{R}$. The direct product phase space is the surface of a *cylinder* (see Fig. 9.2), where the radius of the cylinder is not important.

Note that the phase space can also be viewed as the extended space $\mathbb{R} \times \mathbb{R}$ with proper interpretation of one $\mathbb{R}$.

Trajectories:

Since the generalized coordinates and the conjugate momenta change continuously with time, the motion of the bob generates a curve in the phase space. This curve is called a trajectory. The second step is to construct all the possible trajectories.

As Eq. (9.1) is time reversible, any piece of trajectory in the upper half plane of the (q, p) phase space with arrow to the right (indicating the direction of motion), has a mirror image in the lower half $(p \to -p)$ with arrow towards left. As a corollary, this time-reversed pair meets to form *a closed orbit*, if and only if there is a point with $p = 0$. In addition, uniqueness of solution for a given initial condition forbids crossing of distinct trajectories.

Let us now combine all the above features. We find that the possible trajectories in the extended $\mathbb{R}^2$ space are either closed loops around $q = 2n\pi$ or open curves. See Fig. 9.3. There is the transitional one that connects the saddle points at $q = (2n + 1)\pi$ at $p = 0$. On the cylindrical phase space, all are closed orbits; one type (for $E < 2$) enclosing the stable fixed point $(0, 0)$, and another type encircling the cylinder (with $E > 2$). The latter can be grouped into two inequivalent classes, namely the time reversed partners (see item 6 in Sec. 9.1.1). It is obvious that one type cannot be deformed into the other one if we follow the rules of deformation on the cylinder. The special one is $E = 2$, a conjoined twin connected at one point. It requires a pinching (i.e., identification) of two points on the curve for $E < 2$. A tearing is required as E exceeds 2 by any amount no matter how small. Moral: The topology of the phase space naturally separates the three types of orbits. The cylindrical topology forbids transformation of one to the other.

Another way to see the change is to use energy $E(p, q)$ as a parameter or replace p by E. As E depends quadratically on p, the cylindrical phase space becomes a U-tube. To be noted that the horizontal $p = 0$ circle on the cylinder has now become the vertical circle in the middle of the U as the minimum energy is zero for $\theta = 0$ but 2 for $\theta = \pi$. The motion is then given by the intersection of the U-tube with an E =const plane. See Fig. 9.3. The three classes of closed loops are now easy to see. The corresponding orbits in the configuration space are shown in Fig. 9.2.

The peculiarity of the critical case is revealed by the response of the pendulum to a vanishingly small random perturbation at say the turning or the top point. There will be no drastic change for the $E > 2$ or the $E < 2$ cases. But for the figure eight case, when $E = 2$, the motion would consist of any combination of clockwise (C) and anticlockwise (A) orbits like CCAAAACACCC... . That is to say, all infinitely long two letter words are possible trajectories, and any two words differing in at least one letter are distinct.

Problem 9.1.1: Suppose acceleration due to gravity $g = 0$. The phase space is still a cylinder but the motion is different. Discuss how the topological arguments change, by focusing on the change in the U-tube for energy as $g \to 0$.

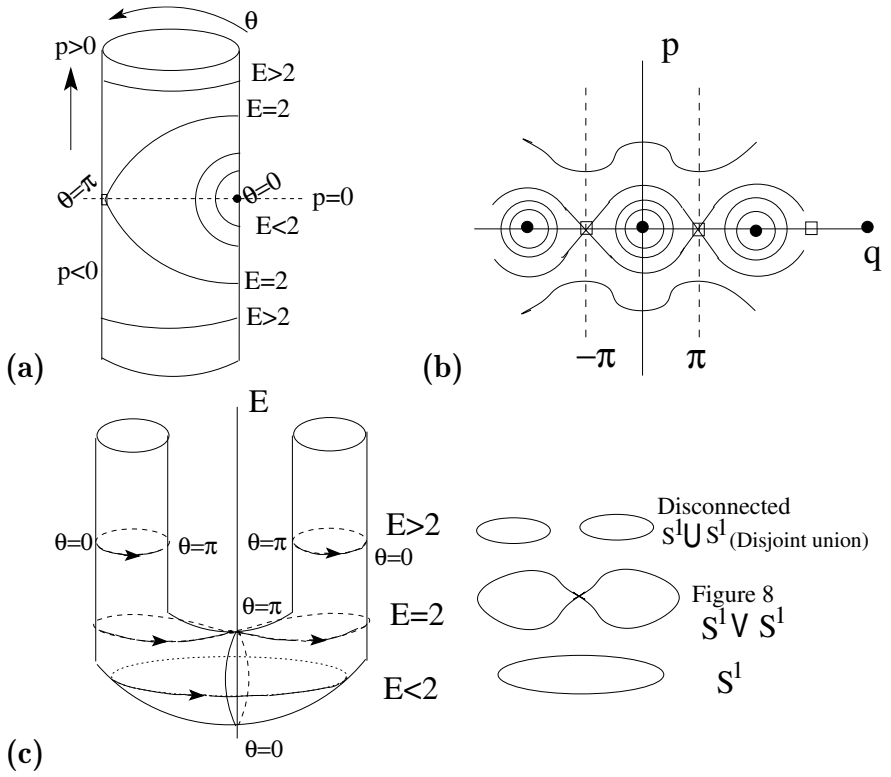


Figure 9.3: Trajectories on (a) the cylindrical phase space, and (b) the extended space $\mathbb{R}^2$. The closed orbits around $\theta = 0$ in (a) are equivalent to the closed orbits around the stable points $q = 2n\pi$ denoted by filled dots in (b) for $E < 2$. In contrast, the closed orbits in (a) encircling the cylinder for $E > 2$ correspond to the open ones in (b). The critical trajectory connects the unstable points $q = (2n + 1)\pi$ represented by unfilled squares in (b) but it just encircles the cylinder in (a) with a point of contact. The U-tube space when E is used as an axis is shown in (c). The actual trajectories are the E =const plane intersections of the U-tube. On the right, three different types of intersections: closed for $E < 2$ (topologically equivalent to a circle S^1), figure 8 for $E = 2$ (two circles with one common point, called the wedge sum $S^1 \vee S^1$), and two disconnected closed pieces for $E > 2$ (disjoint union of two circles, $S^1 \cup S^1$ but with $S^1 \cap S^1 = \emptyset$).

9.2 Topological analysis: details

We now discuss how topology is used in the description. Let us remember that a topology on a set X of points require a set of subsets, τ, to be called open sets, such that (i) $\emptyset$ (Null set) and full set X are open, i.e., $\emptyset, X \in \tau$, (ii) any finite or

infinite union of open sets is also open, (iii) any *finite* intersection of open sets, i.e. members of τ, belongs to τ. Under these conditions, the set of subsets, τ, is called a topology on X, while (X, τ) is said to constitute a topological space. See Ref. [4].

A useful procedure to define a topology on a set is to embed it in a known space. Then use the intersections of the open sets of the known space with the set in hand to define the open sets in it. For example, S^1 can be drawn in a two dimensional space and the open sets on this curve can be defined as the intersection of the curve with the rectangles (or Disks). The open sets on the circle then form the basis for the topology on S^1. The topology thus defined is called the *Inherited topology or the subspace topology.* Since we shall mostly be working with these inherited topologies, we shall not be explicit about it anymore, unless something else is meant. Such embeddings are useful in most physics problems but there are many cases for which no embeddings are possible.

Right now we rely on our intuition of spaces.

9.2.1 Configuration Space

Our familiarity with the real d-dimensional Euclidean space R^d allows us the luxury of thinking of other spaces in terms of R^d. With that, it might be possible to make topological identifications of spaces. There should be a statutory warning that proving the topological equivalence of spaces in general could be a notoriously difficult problem.

S^1 as the configuration space

Our intuition of angle leads us naturally to S^1. If we think of the set of values $q \in [0, 2\pi]$, we get S^1 only if we identify 0 and 2π. This can be seen by gluing the two ends of a piece of a string (i.e. implement the "*periodic boundary condition*"). A simple but concrete way of seeing this is to note that a continuous map takes $q \to z = e^{iq}$ with z defining the unit circle in the complex plane.

$\mathbb{R}$ as the configuration space: equivalence relation, quotient space

A slightly different identification is required for the extended real line used in Fig. 9.1. This involves an *equivalence relation* that any $x \in \mathbb{R}$ is equivalent to all points $x + na$ for $n \in Z$. It is like a translational symmetry in one dimension. Any point on $\mathbb{R}$ can then be brought into the interval $[0, a]$ or $[-a/2, a/2]$ by addition or subtraction of suitable multiples of a. E.g., $-a/2 = a/2 - a$ sets $-a/2$ as equivalent to $a/2$. This finite closed interval with end point identification is S^1. If the equivalence is denoted by the symbol $\sim$, i.e., $x \sim x + na$, then under this equivalence condition, the relevant space is different from $\mathbb{R}$. It is denoted by $\mathbb{R}/\sim$, and is called the *quotient space* for the equivalence relation $\sim$. The space obtained via an equivalence as $\mathbb{R}/\sim$ is topologically equivalent to S^1. A

general feature we see here is the possibility of construction of new spaces from a given space by defining an equivalence relation on it.

More examples For the closed interval $I = [0, 1]$, we may define the periodic boundary condition as an end point equivalence relation $0 \sim 1$. Then $I/\sim$ is homeomorphic to S^1, where *homeomorphism* is synonymous to "topologically identical". To be more systematic, one defines (i) a direct map (i.e., a function) $f : I \to S^1$ as $f(x) = e^{i2\pi x}$, and successive maps (ii) a map $f_1 : I \to I/\sim$, and (iii) an inverse map $f_2 : I/\sim \to S^1$, so that $f_2(f_1(x)) = f(x)$ for all $x \in I$.

Note that, we *do not* get S^1 from $[0, 1]$ if the end points are *not identified*. That they are different can be seen by removing one point from each one of the two sets. In the closed interval case we get two disconnected pieces whereas for S^1 we get an open interval. If we take an open interval[5] like $q \in (0, 1)$, then it is actually equivalent to the whole real line as one may verify by the map $x \to X = \tan[\pi(x - \frac{1}{2})]$ with $X \in (-\infty, +\infty)$.

Problem 9.2.1: The change in the topological space by an equivalence relation has important consequences in physics too. Take the case of $[0, 1]$ and S^1 under the equivalence condition. For single particle quantum mechanics, the first case corresponds to the boundary condition where the wave function $\psi(x) = 0$ at $x = 0, 1$ while the second one to periodic boundary condition with $\psi(0) = \psi(1)$.

Take the conventional momentum operator $-i\hbar d/dx$ with the eigen value equation $-i\hbar\frac{d\psi}{dx} = p\psi$. Show that there is no valid solution (i.e., p is not real) for the $[0, 1]$ case while p is real for S^1.

Pendulum vs harmonic oscillator:

The importance of the topology of the configuration space can be understood by comparing the S^1 case with the space for the linearized simple pendulum. For the latter, the configuration space is just a small part of the circle (small angles), which can be extended to the whole real line as for a linear harmonic oscillator. For this space $\mathbb{R}$, there is only one stable fixed point at $x = 0$, and the phase space has only one kind of orbit, namely the closed orbit of libration type. All the richness of the full pendulum at various energies come from the nontrivial topology of the configuration space.

9.2.2 Phase space

Now that we know the configuration space, we may go over to the phase space. The momentum part is easy - it is a real number for the pendulum, $p \in (-\infty, +\infty)$, i.e., $p \in \mathbb{R}$. For n degrees of freedom there are n momenta, and so the space for the momenta is $\mathbb{R} \times \mathbb{R} ... \times \mathbb{R} = \mathbb{R}^n$. Since the momentum

[5]The standard convention is to use parenthesis $(,)$ to denote openness. Here the boundary points a, b are not in the set (a, b).

and the position are independent variables, we have a product space of $\mathbb{R}^n \times C$ where C is the configuration space, as seen for the planar pendulum.

The motion of the pendulum is restricted to the constant energy subspace of the phase space as shown in Fig. 9.3. Beyond visualization, the differences in the nature of the spaces show up through their topological invariants, e.g., by the fundamental group.

Topological invariants — homotopy groups:

One way of exploring a topological space X is by mapping known spaces like circles, spheres, etc in X. The case with circles tells us how many classes of nonequivalent closed loops that start and end at a point x_0 can exist in X. Any point in X can be chosen as the base point x_0. Two loops that can be deformed into one another are called homotopic [5]. All such homotopic loops can be clubbed together as a single class, with any one of them as a representative one. One may also define a product of two loops C_1 and C_2 by going along C_1 and then from x_0 along C_2. There will be classes of loops that are homotopic to C, and therefore the multiplication is for the classes. Representing a class by $[\gamma]$ for all loops homotopic to γ, the multiplication rule can be written as $[C] = [C_2]\ [C_1]$. An inverse of a loop C can be defined as the loop traversed in the opposite direction, so that $C\ C^{-1}$ is homotopic to a point (i.e., a trivial loop). More formally, all these imply that the closed loops rooted at x_0 form a group under the operation of loop multiplications. This group is called the *fundamental group* of X, $\pi_1(X, x_0)$. For a connected space, i.e., if any two points can be connected by a path in X, the special point, the base point can be chosen arbitrarily. We shall therefore drop x_0 from the notation.

The fundamental group of the space for $E < 2$ is $\pi_1(\text{“}E < 2\text{”}) = \mathbb{Z}$ where the base point has been dropped. This part of the space is like the outer surface of a bowl. In contrast, the space for $E > 2$ is disjoint – two tubes, and the loops will depend on whether the base point, x_0, is in one or the other circle. Although $\pi_1(\text{“}E > 2\text{”}, x_0) = \mathbb{Z}$, but x_0 in one circle cannot be connected by a path to the point on the other. Such a disjoint set is characterized by the zeroth homotopy group $\pi_0(\text{“}E > 2\text{”}) = \{-1, 1\} \equiv \mathbb{Z}_2$ with two elements signifying two components. The critical surface is again two circles but with one common point forming figure 8. Such a union of spaces with one common point is called a wedge sum, indicated by a $\vee$. This fundamental group is now a nonabelian group $\pi_1(\text{“}E = 2\text{”}) = \mathbb{Z} \star \mathbb{Z}$ a free group of two elements.

The difference in the fundamental group tells us that the spaces are not identical, i.e., not homeomorphic.

Problem 9.2.2: Show that a square with periodic boundary conditions is equivalent to $S^1 \times S^1$. Take a piece of paper and glue the sides parallelly. This is a torus which is associated with two holes. Define the appropriate equivalence relation ($\sim$), and convince yourself that the compact notation is $[0,1] \times [0,1]/\sim\ = S^1 \times S^1$. This “=” means “homeomorphic” or loosely speaking “topologically equivalent”.

We have already seen two spaces formed by two circles in the case of a pendulum. Compared to the disconnected space with $\pi_0 = \{-1, 1\}$, a torus has $\pi_0 = 0$, just as figure 8 which we obtained from two circles with one point equivalence. Both (torus and figure 8) are connected spaces. That a torus is not topologically figure 8 is established by $\pi_1(S^1 \times S^1) = \mathbb{Z} \times \mathbb{Z}$ while $\pi_1(\text{figure } 8) = \mathbb{Z} \star \mathbb{Z}$.

Problem 9.2.3: Discuss the connection between the fundamental groups of the constant energy spaces mentioned above and the real trajectories of the pendulum.

Problem 9.2.4: The phase space for a simple harmonic oscillator is $\mathbb{R}^2$ with $\pi_1(\mathbb{R}^2) = 0$. Discuss the motion with respect to the corresponding E-x surface.

Problem 9.2.5: The energy of a free one-dimensional quantum particle is given by $E_k = ak^2$ with the wavevector $k \in \mathbb{R}$. The state with the wavefunction described by k may called left-moving or right moving for $k < 0$ or $k > 0$. For a particle on a lattice (lattice spacing=1), the translational symmetry makes two k values equivalent if they differ by a reciprocal lattice vector. In other words the k-space is like Fig. 9.1c. The equivalence relation makes the relevant space S^1 as in Fig. 9.1b. A linear representation of S^1 is the interval $[-\pi, \pi]$ with the identification of the two end points. With this identification, a right moving particle at $k = \pi$ becomes a left moving particle at $k = -\pi$ as defined earlier, but one should keep in mind the presence of reciprocal lattice vectors . Draw E_k vs k in this 1st Brillouin zone.

Problem 9.2.6: (a) Argue that the configuration space of the pendulum in full space (spherical pendulum) is S^2 (surface of a three dimensional sphere). (b) Discuss the possible types of motions using topological arguments. (c) In spherical coordinates S^2 can be described by (θ, ϕ) where $\theta \in I = [0, \pi]$, and $\phi \in [0, 2\pi]$. Why is S^2 not a product space $I \times S^1$?

Problem 9.2.7: If all the boundary points of a square are made equivalent, then it is topologically equivalent to a surface of a sphere S^2. Take a piece of cloth or paper and use a string to bring all the boundary points together, as one does to make a bag. Or take a square and an isolated point. Connect all the points on the boundary to that point.

Problem 9.2.8: Bloch's theorem, in solid state physics, is generally proved for a lattice with periodic boundary conditions, i.e., on a torus (an n-torus for an n-dimensional crystal. E.g., a torus is obtained by identifying opposite edges of a square. Note that if all points on the boundary of a square are identified (spherical boundary condition) one gets S^2. Is Bloch's theorem valid for the spherical boundary condition? Are the reciprocal vectors defined for the spherical boundary condition?

Problem 9.2.9: Bulk and edge states: In the tight binding model, a quantum particle hops on a square lattice. Find the energy eigen states under the following situations. Pay attention to bulk and edge states. (i) A particle on $S^1 \times S^1$. In this case there are only bulk states. (ii) With spherical boundary condition, i.e., on S^2. There are no edges. But are the bulk states same as in (i)? (iii) Klein

Bottle. No edges but different from (i) and (ii). (iv) Periodic boundary condition in one direction, i.e., on a finite cylinder. There are now two edges. (v) Antiperiodic boundary condition, i.e., a Möbius strip. There is now one single edge.

Problem 9.2.10: Argue that the configuration space for the planar motion of a double pendulum is $S^1 \times S^1$. If we consider the full three dimensional space, then the configuration space is $S^2 \times S^2$.

Problem 9.2.11: A classical Hamiltonian system with n degrees of freedom is *integrable* if there exists n conserved quantities or "first integrals". In such a case, the motion is confined on an n-torus $S^1 \times ... \times S^1$. Here the product space indicates that the motions can be handled independently. This is easy to see in the action angle variables where the n angles constitute the n-tori. Convince yourself about this for the pendulum case and for the well-known Kepler problem. This result is useful in the context of the important KAM theorem.

Problem 9.2.12: Kapitza Pendulum: The equation of motion of a pendulum under a periodic vertical drive is $\ddot{\theta} + (g + a(t)) \sin\theta = 0$ where $a(t) = a(t + \tau)$ is the periodic vertical modulation of the point of suspension. The inverted pendulum at $\theta = \pi$ is stable if the amplitude of the drive exceeds some critical value. This is called a Kapitza pendulum. By eliminating high frequency components, the Kapitza pendulum can be described by an effective potential, $V_{\rm eff} = -g\cos\theta - g_2 \cos 2\theta$. Discuss the motion of an inverted pendulum under a periodic vertical drive or under $V_{\rm eff}$ vis-a-vis Fig. 9.3.

Problem 9.2.13: Show that a plane with a hole is equivalent (homeomorphic) to a cylinder. With the hole as the origin, use polar coordinates r, ϕ so that $r = 0$ is excluded (=hole). Now do a mapping $r \to X = \ln r$ so $X \in (-\infty, +\infty)$, i.e., $X \in \mathbb{R}$ and ϕ defines S^1. Therefore $\mathbb{R}^2 - \{0\}$ (also written as $\mathbb{R}^2 \backslash \{0\}$) is a cylinder.

Solve the free particle quantum mechanics problem in $\mathbb{R}^2 - \{0\}$ in r, ϕ coordinates. What are the boundary conditions? Do the same on the cylinder by transforming the Schrödinger equation to X, ϕ variables. The main point of this exercise is to see the importance of one missing point that changes the topology of the space.

Problem 9.2.14: Show that a sphere with a hole $S^2 \backslash$N (N=north pole) is equivalent to a plane. The formal proof is by stereographic projection. A sphere with two holes (north and south poles) is like a cylinder, which in turn is a plane with a hole. What about a sphere with three missing points?

Problem 9.2.15: What is the advantage of going from the cylinder to the extended real plane as in Fig. 9.3? The real line or plane has the special feature that any closed loop can be shrunk to a point. Such a space is called simply-connected (as opposed to multiply-connected as in the previous problem). A practical usefulness may be seen by considering a damped pendulum described by $\ddot{\theta} + \gamma\dot{\theta} + \sin\theta = 0$, where γ is the friction coefficient. Now energy is not conserved, $dE/dt < 0$, so that the pendulum ultimately for $t \to \infty$ comes to rest at the stable fixed point $\theta = 0$. Draw the possible trajectories (phase portrait) of this damped pendulum for

different values of γ and starting energy ($E > 2, E < 2$) both on the cylinder and on the extended space.

9.3 Topological spaces

Is the combination of two real variables q, p equivalent to a two dimensional Euclidean plane? The question arises because even if we take q, and p as the two directions of the xy plane, still we may not be in a position to define a distance between two points (q_1, p_1) and (q_2, p_2). The second point is that for a physical system described by two variables, the state space may locally be like a plane (two dimensional) but different global connectivities may imply important qualitative differences. E.g., for a torus and a sphere, a small neighbourhood of a point may be described by the tangent plane at that point and would look similar, but globally they are different. Let us concentrate on the first issue now.

The absence of a metric (or distance) is a generic problem we face whenever we want to draw a graph of two different parameters. Take, for example, a plot of pressure P and volume V for a verification of Boyle's law. The plot reassuringly gives us a branch of a hyperbola, which is defined as the locus of a point such that the difference in the distance from two fixed points remain constant. But it would be ridiculous to define an Euclidean distance between (P_1, V_1) and (P_2, V_2). Still, we know, graph plotting does work marvelously.

The identification is done in steps through topology. First an appropriate topological space is defined which can be identified with the similar topological space in $\mathbb{R}^2$. Then, use the equivalence of this topological space and a metric or distance based $\mathbb{R}^2$.

Let's start with the real line. To define a topology we need a list or a definition of open sets. Let's define all sets of the type $(a, b), (b > a)$ and their unions as open sets. The null set $\emptyset$ and the full set are also members of the set of open sets. That these subsets form a topology on $\mathbb{R}$ is easy to check. The set of subsets with the union and intersection rules then defines a topological space. For the real line we used only the "greater than" or "less than" relation, without defining any distance or metric.

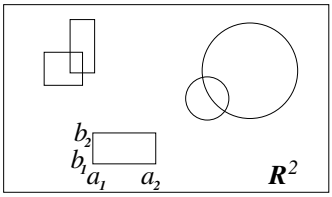


Figure 9.4: Basis for R^2 as a topological space: open rectangles or disks.

This topology can be extended to $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ by defining the sets of open rectangles $(a_1, a_2) \times (b_1, b_2)$. See Fig. 9.4. By doing this we defined a topology in $\mathbb{R}^2$ without using any distance. The next step is to define a metric, the usual

Euclidean distance in $\mathbb{R}^2$ with which open disks $D = (x, y|(x-a)^2+(y-b)^2 < \epsilon)$ can be defined around a point (a, b). It is known that the topology defined by the open rectangles and their unions is the same as the one defined by the disks.

By this procedure, with the help of boxes, the (q, p) phase-space can be taken as a topological space equivalent to $\mathbb{R}^2$. This equivalence allows one to see all the geometric features of $\mathbb{R}^2$ in the graphs we plot or in the phase space, without explicitly defining the distance.

An important feature of the topology of the phase space is that it is *connected and simply connected*, i.e., one may go from any point to any other point, and any two paths connecting two points can be deformed into each other.[6] A connected phase space is nice because that is a sufficient condition for the applicability of equilibrium statistical mechanics (generally called *ergodicity* — that one can go from any state to any other). However, a phase space may as well be in disconnected pieces in the sense that two parts may be separated by infinite energy barriers. Such spaces might be relevant for phase transitions where the phase space may get fragmented into pieces ("broken ergodicity" or ordered systems).

Problem 9.3.1: In Fig. 9.4, an infinite number of open boxes are used as "basis" sets to define the topology of $\mathbb{R}^2$. As a vector space, we need only two unit vectors $\mathbf{i}, \mathbf{j}$ where the number 2 of $\mathbb{R}^2$ determines the number of basis vectors. Where is this "2" when defined as the topological space? Argue that this dimensionality comes from the number of spaces required to construct the boxes.

Problem 9.3.2: We defined the topology for S^1 by embedding it in $\mathbb{R}^2$ (subspace topology). Is it possible to define a topology on S^1 without any embedding?

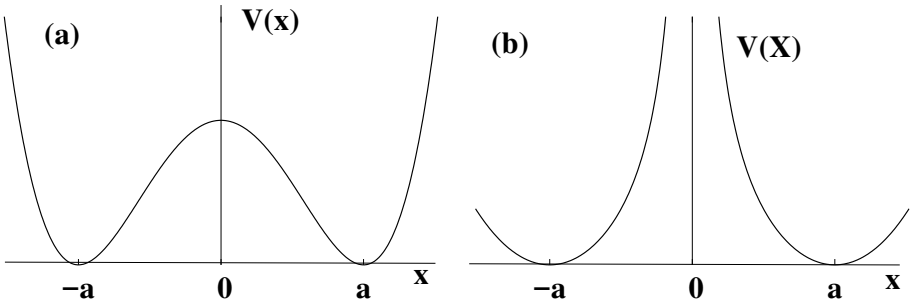


Figure 9.5: (a) A double well potential. (b) A double well potential but with an infinite barrier inbetween. The barrier cannot be crossed.

[6] A connected space has zeroth homotopy group $\pi_0 = 0$. A simply-connected space means $\pi_1 = 0$.

Problem 9.3.3: Consider the one-dimensional motion of a particle in a double well $V(x) = \frac{1}{2}K(x^2 - a^2)^2$. See Fig.9.5(a). Discuss the nature of the configuration space and of the phase space. Locate the fixed points and draw the phase space portrait.

Problem 9.3.4: Suppose the barrier of the double well potential is infinitely high (Fig. 9.5b.) Argue that the configuration space consists of disconnected pieces. Draw the possible phase portraits.

9.4 More examples of topological spaces

Let us now consider a few other well-known spaces used in condensed matter physics. Once the spaces are identified, their topological classifications in terms of fundamental groups and higher homotopy groups help us in identifying the defects that may occur, the type of particles that may be seen and so on. Here we just construct the spaces in a few cases.

A class of condensed matter systems involve ordered states. like crystals, magnets, liquid crystals, etc. These states or phases have a symmetry described by a group H which is a subgroup of the expected full symmetry G. For example, the Hamiltonian of N interacting particles is expected to be invariant under full translational and rotational symmetry, G, but a crystal, described by the same Hamiltonian has only a discrete set of space group symmetries. Such phenomena of ordering is known as symmetry breaking. The ordered state is then described by a parameter that reflects this subgroup structure of the state. The allowed values of the order parameter constitutes the topological space for the ordered state and this space is called the order parameter space. Of all the ordered states, ferromagnets and liquid crystals are easy to describe. We discuss these spaces below.

9.4.1 Magnets

A ferromagnet is described by a magnetization vector $\mathbf{M}$. In the paramagnetic phase $\mathbf{M} = 0$ but a ferromagnet by definition has $\mathbf{M} \neq 0$. For simplicity (e.g. at a particular temperature or zero temperature) we take M =const , only its direction may change.

Magnets can be of different types depending on the nature of the vector. If $\mathbf{M}$ takes only two directions up or down, then it is to be called an Ising magnet. If $\mathbf{M}$ is a two dimensional vector, it is an xy magnet, and for a three dimensional vector it is an Heisenberg magnet. In the ferromagnetic phase the origin ($M = 0$) is not allowed and so any vector space description will be of limited use. What is then required is a topological description of the allowed values of the magnetization. Since ferromagnetism is a form of ordering of the microscopic magnetic vectors, we call the space an *order parameter space* $\mathcal{O}$.

It is now straightforward to see that the order parameter spaces $\mathcal{O}$ are of the following kinds:

(i) Ising: $\mathcal{O} = \mathbb{Z}_2$, (0,1) i.e., up or down
(ii) xy: $\mathcal{O} = S^1$ (circle), i.e., the angle of orientation, θ.
(iii) Heisenberg: $\mathcal{O} = S^2$ (surface of a 3-dimensional sphere), i.e., angle of orientation, i.e., θ, ϕ.
(iv) n-vector model: there are situations where the space could be $S^n, n > 2$.

If we take a macroscopic d-dimensional magnet, then at each point of the sample ($\mathbf{x} \in \mathbb{R}^d$) we define a magnetization vector $\mathbf{M}(\mathbf{x})$ or a mapping $\mathbf{M} : \mathbb{R}^d \to \mathcal{O}$. That such a mapping can be nontrivial has important implications. Instead of a fullfledged analysis of the mapping, it helps to see how loops and spheres in real space map to the orderparameter space. E.g., when we move along a closed loop in real space, the order parameter describes a closed loop in $\mathcal{O}$. The nature of these closed loops in $\mathcal{O}$ is given by the fundamental group $\pi_1(\mathcal{O})$. A nontrivial π_1 indicates there are loops that cannot be shrunk to a point. This, in turn, means that if a loop in real space is shrunk, there will be problems with continuous deformation of the spins; there has to be a singularity where the orientation of the spin cannot be defined. These are called topological defects. In $d = 2$, loops will enclose point defects while in $d = 3$, loops will enclose line defects, with the elements of $\pi_1(\mathcal{O})$ as the "charges" of these defects.

We just quote here the results [5] that $\pi_1(S^1) = \mathbb{Z}$, and $\pi_1(S^n) = 0, n > 1$. These mean that only for the xy-magnet there will be point defects in two dimensions and line defects in three dimensions. In particular, Heisenberg magnets will have no point (line) defects in two (three) dimensions. Any Heisenberg spin configuration in real space can be changed to any other by local rearrangements of spins. In contrast, for a 2-dimensional xy magnet, a configuration with a point defect of say charge=1 *cannot be converted by local rearrangements of the spins* to a defectless configuration. The question of continuity of a mapping (i.e., a function) using topology is discussed separately in Sec. 9.6.

Problem 9.4.1: There seems to be an obsession for S^n, but that's not for no reason. Prove that S^n is the only compact simply-connected[7] "surface" in n-dimensions ($n \geq 2$).[8] (Poincaré's conjecture.)

Problem 9.4.2: Berezinskii-Kosterlitz-Thouless transition: it is known that the cost to create a unit charge defect in the 2-d xy model is $\varepsilon \ln L$ for, say, a square lattice of size $L \times L$. Since the defect can be anywhere on the lattice, argue that the free energy of a single defect at temperature T is $f(T) = \varepsilon \ln L - cT \ln L$, where c is some constant. Take the defect free state as of zero free energy. Show that free defects may form spontaneously if $T > T_{BKT} = \varepsilon/c$. This phase transition is called the Berezinskii-Kosterlitz-Thouless (BKT transition).

[7] A simply-connected space is one where any loop can be contracted to a point. This means its fundamental group is trivial, $\pi_1 = 0$.

[8] Any compact, simply connected n-dimensional "surface" is equivalent to S^n. Remember that S^n is the surface of a sphere in $(n+1)$-dimensional space, $\sum_{i=1}^{n+1} x_i^2 = 1$.

9.4.2 Liquid crystals

Nematics: $\mathbb{R}P^2$

Lest we created the impression that the world is just S^n's, we look at a different ordered system, namely liquid crystals. A nematic liquid crystal consists of rod like molecules where the centres of the rods are randomly distributed as in a liquid but the rods have a preferred orientation $\mathbf{N}$. This looks like a magnet but it isn't so because a rod does not have a direction, i.e., it is like a headless arrow. A flipping of a rod won't change anything in contrast to $\mathbf{M} \to -\mathbf{M}$. As a direction in 3-dimensions, the order parameter space $\mathcal{O}_{\text{nematic}}$ should have been S^2 but not exactly. Two points on a sphere which are diametrically opposite represent the same state, and therefore there is an equivalence relation on the sphere that *antipodal points are equivalent*, $\mathbf{N} \to -\mathbf{N}$. This is not just the hemisphere but a hemisphere with the diametrically points identified on the equator. This is called the real projective plane $S^2/\mathbb{Z}_2 = \mathbb{R}P^2$. In general, $S^n/\mathbb{Z}_2 = \mathbb{R}P^n$.

As an ordered system, we would like to know how the headless arrows can be arranged in space. This requires the behaviour of the map $\mathbf{N} : \mathbb{R}^d \to \mathbb{R}P^2$.

Biaxial nematics

Instead of rod like molecules, one may consider rectangular parallelepiped with 2-fold rotational symmetry corresponding to the 2π rotations around the three principal axis. Such a liquid crystal is called a biaxial nematics. The order parameter space is the sphere S^2 with the equivalence relation, $\sim$, under the three rotations. This "$\sim$" is not just the identification of the antipodal points but, in addition, the equivalence of four sets of points (corners of the box) on the surface of the sphere. The generic notation $\mathcal{O} = S^2/\sim$ is too cryptic to have any use. This is where the symmetry operations as a group is useful.

The sphere is actually a representation of the rotational symmetry. If n_1, n_2 are any two allowed values of the order parameter, they are related by the three dimensional rotation group $G = SO(3)$. By keeping any one value fixed, say n_1, all others can be generated by the application of the group elements of G. However, the special symmetry of the biaxial nematics, a subgroup of four elements, $H = D_2$, keeps the order parameter invariant, i.e., if $h \in H$, then $n_1 = hn_1$. Then, there is some $g \in G$, for which $n_2 = gn_1 = ghn_1$, so that n_2 is generated by all group elements of the type gh with $h \in H$ and $g \in G$ but not in H. What we get is the coset of H in G, G/H. So, instead of the generic representation as $S^2/\sim$, we may use groups to represent the order parameter space as a coset space, $\mathcal{O} = SO(3)/D_2$. The similarity of notations (quotient space in topology and coset space in group theory) is not accidental but is because of the similarity of the underlying concepts. It is now straightforward to generalize to any other point group symmetry. It will be the corresponding coset space.

Problem 9.4.3: Instead of $SO(3)$, one may consider $SU(2)$. Under this mapping, show that D_2 goes to a eight member nonabelian group, Q, the group of quaternions. Therefore, $\mathcal{O} = SU(2)/Q$.

Problem 9.4.4: Show that the order parameter spaces for magnets can be written in terms of groups as the following coset space:

1. the xy case: $\mathcal{O} = SO(2)$, or $\mathcal{O} = U(1)$. Note that the coset space is a group in this case.
2. the Heisenberg case: $\mathcal{O} = SO(3)/SO(2)$, or $\mathcal{O} = SU(2)/U(1)$.

What is $\mathbb{R}P^n$?

A real projective space is obtained by identifying the points which differ by a scale factor. If any point $\mathbf{x} \in \mathbb{R}^{n+1}$ is equivalent to $\lambda\mathbf{x}$ for any real $\lambda \neq 0$, then under this equivalence relation $(\mathbb{R}^{n+1}\backslash\{0\})/\sim = \mathbb{R}P^n$. Geometrically, all points on a straight line through the origin are taken as equivalent. The space then consists of unit vectors $\mathbf{n}$ with $\mathbf{n}$ equivalent to $-\mathbf{n}$.

Problem 9.4.5: What is the configuration space of a rigid diatomic molecule in 3 dimensions?
Ans: $\mathbb{R}^3$ for the centre of mass and S^2 for orientation of the molecule. In case the two atoms are identical then it is $\mathbb{R}^3 \times \mathbb{R}P^2$.

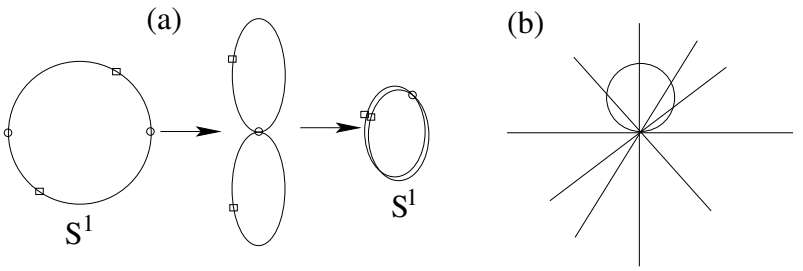


Figure 9.6: (a) Twist a circle which brings two antipodal points together. Then fold the two circles so that again antipodal points are on top of eachother. (b) The space of all lines through origin. Any point on a line, except the origin, are equivalent to all others on the same line. This space is S^1.

$\mathbb{R}\boldsymbol{P^1}$**:** Take a circle and identify the diametrically opposite points. See Fig. 9.6a. This is easy to do with a rubber band. The folding process shows that $S^1/\mathbb{Z}_2 = \mathbb{R}P^1 = S^1$. Another way of looking at $\mathbb{R}P^1$ is shown in Fig. 9.6b. Take all straight lines through origin in two dimensions. Then declare all points on a line, except the origin, as equivalent. We may choose any point (not origin) on a line as a representative point. Draw a circle through the origin with the center

on the y-axis. Every line meets this circle at a point (again exclude the origin) which can be taken as a representative point of the line. There is therefore one-one correspondence between the points on the circle and the lines through the origin. The excluded point on the circle (the origin of $\mathbb{R}^2$) can then be included as the representative point for the x-axis. Hence the topological equivalence of $\mathbb{R}P^1$ and S^1.

In contrast, $\mathbb{R}P^2$ is not simple. In the Euclidean case any two straight lines intersect at one and only one point, unless they are parallel. Parallel lines do not intersect. In the real projective plane, any two straight lines always intersect either in the finite plane or at infinity[9].

9.4.3 Crystals

Take the case of a crystal which has broken translational symmetry. If we move an infinite crystal by a lattice vector, the new state is indistinguishable from the old one. For concreteness let us take the crystal to be a square lattice of spacing a in the xy plane. Consider the atoms to be slightly displaced from the chosen square lattice. Now, if one atom is at $\mathbf{r}_0$ from one particular lattice site, it is at $\mathbf{r}_{mn} = \mathbf{r}_0 + ma\hat{i} + na\hat{j}$ from a site at (m, n). All of these are equivalent. The order parameter space is then the real plane $\mathbb{R}^2$ under the equivalence condition of translation of a square lattice. There is now an equivalence relation that any point $\mathbf{r}$ is equivalent to a point $\mathbf{r} + ma\hat{i} + na\hat{j}$ for $m, n \in \mathbb{Z}$. The order parameter space is therefore a torus. Note that this is a generalization of the one dimensional case of Fig. 9.2 to 2-dimensions, except that we are now going from Fig 9.2c to Fig. 9.2b.

9.4.4 A few Spaces in Quantum mechinics

We consider the forms of a few finite dimensional Hilbert spaces.

Complex projective plane $\mathbb{C}P^n$

In quantum mechanics, the square integrable wave functions form a Hilbert space. Any state $|\psi\rangle$ can be written as a linear combination of a set of orthonormal basis set $\{|j\rangle\}$ as $|\psi\rangle = \sum_j c_j |j\rangle$. Let's keep the number of basis vectors finite, $n < \infty$. The set of n complex numbers $\{c_j\}$ is an equivalent description of the state so that the state space is an n-dimensional complex space $\mathbb{C}^n$. Since only normalized states matter, $\{c_j\}$ and $\{\lambda c_j\}$ for any complex number λ represent the same state. Hence there is an equivalence relation $\{c_j\} \sim \{\lambda c_j\}$ in $\mathbb{C}^n$. The relevant space for wave functions is then $(\mathbb{C}^n \backslash \{0\})/\sim = \mathbb{C}P^{n-1}$, a complex projective space in analogy with real projective spaces.

[9] In paintings, for proper perspective, parallel lines are drawn in a way that gives the impression of meeting at infinity.

Two level system

A particular case, probably the simplest, is a two level system (a qubit), like a spin 1/2 state with $|+\rangle$ and $|-\rangle$ states. The space of states is therefore the one-dimensional complex projective plane $\mathbb{C}P^1$. Any normalized state can be written as

$$|\psi\rangle = \cos\frac{\theta}{2}\ |+\rangle + e^{i\phi}\sin\frac{\theta}{2}\ |-\rangle, \quad \text{with } 0 \le \theta \le \pi, 0 \le \phi \le 2\pi. \tag{9.3}$$

The two angular parameters θ and ϕ allow us to map $\mathbb{C}P^1$ to S^2, called the *Bloch sphere*, though there are problems with the representation of $|+\rangle$ and $|-\rangle$. To get rid of this problem one actually needs two maps. From the equivalence relation $(c_1, c_2) \sim (\lambda c_1, \lambda c_2)$, we may choose λ to write $(c_1, c_2) \sim (1, z)$ or $(c_1, c_2) \sim (1/z, 1)$ so that the two original basis vectors $|\pm\rangle$ come from $z = 0$ or $z = \infty$. The addition of the point at infinity to the complex plane gives us the *Riemann sphere*, also called *one point compactification of the complex plane.*

The sphere allows us to define a metric in terms of θ, ϕ, which then acts as a metric, the Fubini-Study metric, for $\mathbb{C}P^1$.

Space of Hamiltonians for a two level system

The Hamiltonian for a two state system is a 2×2 Hermitian matrix. Therefore we may consider the space of all such Hamiltonians. Any Hermitian 2×2 matrix can be expressed in terms of the Pauli matrices[10]

$$H = \begin{pmatrix} \epsilon_1 & a - ib \\ a - ib & \epsilon_2 \end{pmatrix} = \frac{\epsilon_1 + \epsilon_2}{2}\mathbb{I} + a\ \sigma_x + b\ \sigma_y + \frac{\epsilon_1 - \epsilon_2}{2}\ \sigma_z. \tag{9.5}$$

In general, the space of the 2×2 Hermitian Hamiltonians is a real four dimensional space with $(\mathbb{I}, \sigma_x, \sigma_y, \sigma_z)$ as the basis vectors.. It has to be a real space because hermiticity requires all the four numbers, $\epsilon_1, \epsilon_2, a, b$ to be real.

In some situations a further reduction in the dimensionality of the space is possible. By a shift of origin, we may set $\epsilon_1 = -\epsilon_2 = \epsilon$ to make H traceless. In this situation,

$$H = \mathbf{d} \cdot \boldsymbol{\sigma} = |d|\mathbf{n} \cdot \boldsymbol{\sigma}, \tag{9.6}$$

with $\mathbf{n} = \mathbf{d}/|d|$, a unit vector. The set of all such traceless Hamiltonians can be described by the vector $\mathbf{n}$, *provided* $|d| \neq 0$. Therefore, the space of the Hamiltonians of any two level system is S^2. The center of the sphere corresponds to the degenerate case, $|d| = 0$, when the two energy eigenvalues are same.

A practical example is a spin-1/2 particle in a magnetic field with $H = -\mathbf{h} \cdot \boldsymbol{\sigma}$, where h may depend on some external parameters including time.

[10] Pauli matrices are taken in the standard form, where σ_y is complex, as

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{9.4}$$

Another example is a two band system. For a one dimensional lattice, consider two bands $\epsilon_1(k), \epsilon_2(k)$ with some symmetry such that $\epsilon_1 + \epsilon_2 =$const for all k. Choosing the constant to be zero, we now have Hamiltonian of the type Eq. 9.6 with $\mathbf{d}(k)$ a function of the quasimomentum k, where k is in the first Brillouin zone, $-\pi \leq k \leq \pi$. We therefore have a map $S^1 \to S^2$. In two dimensions, the Brillouin zone is a torus and therefore we need to study the map $\mathbb{T}^2 \to S^2$. Some aspects of these maps are considered in Sec. 9.7.7.

Problem 9.4.6: Construct the topological space for the Hamiltonian of a three level system. Explain why it is reasonable to expect $SU(3)$ and not a spin $s = 1$ state. Generalize it to m-level system for any m.

Problem 9.4.7: The Bloch sphere describes the pure states. The density matrix of a state $|\psi\rangle$ is $\rho = |\psi\rangle \langle\psi|$, with $\rho^2 = \rho$, $\mathrm{Tr}\ \rho = 1$. These two conditions on ρ can be taken as the definition of a pure state without any reference to wave functions. In this scheme, mixed states are those for which $\mathrm{Tr}\ \rho = 1$, but $\rho^2 \neq \rho$. This means $P = \mathrm{Tr}\ \rho^2 < 1$. P is called the purity of the state. For the two state system, mixed states are given by 2×2 Hermitian, positive semidefinite[11] matrices with trace 1. Show that these mixed states are points inside the Bloch sphere. The relevant space is now a 3-ball (a solid sphere).

9.5 Disconnected space: Domain walls

Of all the order parameter space for a magnet defined in Sec 9.4.1, the Ising class is special because here the space is disconnected. i The same result is obtained by using the ϕ^4 theory with an energy functional

$$E[\phi(x)] = \int_{-\infty}^{\infty} dx \left[\frac{1}{2} \left(\frac{d\phi}{dx} \right)^2 + \frac{1}{2} K \left(\phi(x)^2 - \phi_0^2 \right)^2 \right], \tag{9.7}$$

so that the minimum energy states correspond to $\phi(x) = \pm\phi_0$. For finite energy states, we require $\phi(x)$ to be nonconstant but going to $\pm\phi_0$ as $x \to \pm\infty$. The requirement at infinity gives us four possibilities, shown in a tabular form below.

	$\phi\|_{x\to-\infty}$	$\phi\|_{x\to\infty}$
(a)	ϕ_0	ϕ_0
(b)	$-\phi_0$	$-\phi_0$
(c)	$-\phi_0$	ϕ_0
(d)	ϕ_0	$-\phi_0$

Table 9.1: Boundary conditions at $\pm\infty$.

[11] Positive semidefinite means all the eigenvalues, λ_i's satisfy $\lambda_i \geq 0$. For a density matrix we need $0 \leq \lambda_i \leq 1, \sum_i \lambda_i = 1$.

These four cases are distinct because there is no continuous transformation that would change one to the other.

For cases (a) and (b), local changes (like spin flipping) can reduce the energy to zero and these represent small deviations from the fully ordered uniform state of ϕ_0 or $-\phi_0$. These two states are related by symmetry but they are distinct.

For cases (c) and (d), no continuous local transformation can change the boundary conditions to the uniform state. Therefore, they represent different types of states. These finite energy states are called topological excitations because their stability is protected by topology. This is a domain wall or interface separating the two possible macroscopic state $\pm\phi_0$. These topological excitations are called kink for (c) and anti-kink for (d).

More generally for any discrete or disconnected configuration space, i.e., if its π_0 (zeroth homotopy) is nontrivial, there will be domain walls. A better description of a disconnected space is via the zeroth homology, H_0, for which we refer to Ref. [6].

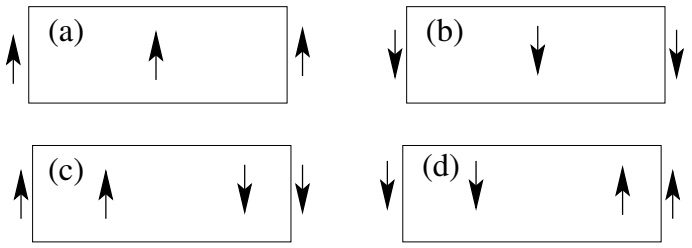


Figure 9.7: Four possible boundary conditions for an Ising magnet with spins=± 1. The up (down) arrow indicates spin $+1(-1)$. An interface exists for (c) and (d).

One may see this boundary condition induced domain walls in the ordered state of a two dimensional Ising model on a square lattice. If we take a long strip with four different boundary conditions as in Fig. 9.7 along one direction, we force domain walls in cases with opposite boundary conditions as in Fig. 9.7c,d. The energy of the interface is obtained by subtracting the free energy of (a) or (b) from (c) or (d). This is ensured in Eq. (9.7) by taking the energy at infinity to be zero.

Problem 9.5.1: Use the energy functional of Eq. (9.7) to determine the domain wall energy.

Problem 9.5.2: The previous discussion allows for only one type of domain wall to exist in a configuration. A generalization would be to consider a case of several disconnected pieces of the configuration space, as in the Potts model. In this model, each "spin" can take q discrete values. A lattice Hamiltonian with nearest neighbour interaction can be of the form $H = -J\sum_{nn}\delta(s_i, s_j), J > 0$. Show

that the ground state is q-fold degenerate. Discuss the nature of domain walls or kinks/antikinks in the Potts model.

Problem 9.5.3: The boundary conditions in Fig. 9.7 can be classified as periodic (a,b) and antiperiodic (c,d). If we join the two vertical edges (equivalence relation) in a way that matches the arrows, show that we get a cylinder for (a) and (b) while a Möbius strip for (c) and (d). See Appendix A for a problem on flux through such surfaces.

9.6 Continuous functions

So far our focus has been on the topological spaces defined for various sets. In the process functions are also defined as maps between two given spaces. It is necessary to define a continuous function in topology without invoking the ϵ, δ definitions of calculus.

The topological definition of a continuous function is in terms of its inverse function. A function $f : A \to B$ is continuous if f^{-1} maps open sets of B to open sets of A. The definition in calculus is that given any ϵ no matter how small, if we can find a $\delta(\epsilon)$, which depends on ϵ, such that $|f(x+\delta) - f(x-\delta)| < \epsilon$, then $f(x)$ is continuous at x. In this ϵ-δ definition, continuity is linked to closeness as measured by a distance-like quantity. The topological definition replaces the neighbourhoods by the open sets, the constituent blocks of the space, but, in addition, it involves the inverse function. That should not be a surprise if we recognize that, by specifying ϵ for f and then finding δ for x is like generating the inverse function.

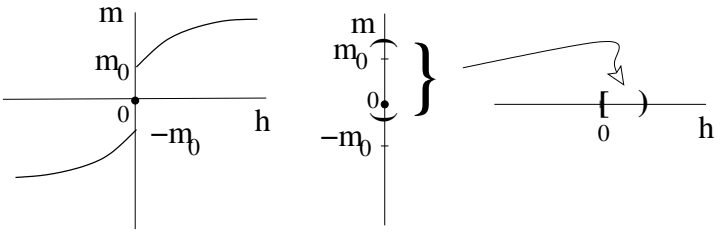


Figure 9.8: A discontinuous function. Magnetization vs magnetic field for a ferromagnet. At $h = 0, m = 0$ and $\lim_{h\to 0\pm} m(h) = \pm m_0$. The example on the right shows an open interval $(-\epsilon, m_1), \epsilon > 0, m_1 > m_0$, maps on to a semiclosed interval $[0, h_1)$ for h, with $m(h_1) = m_1$. The inverse function maps an open interval of m to a semi-open interval of h.

We illustrate this with the help of a known physical example, namely the magnetization of a ferromagnet in a magnetic field, $m(h)$ with $m, h \in \mathbb{R}$. See Fig 9.8. Take any open set (h_i, h_f), the corresponding values of m form an open

set $(m(h_i), m(h_f))$, missing the "discontinuous" nature at the origin, whereas the inverse image of $(-\epsilon, m_1)$ maps to the semiopen set $[0, h_1)$.[12]

The definition of continuity in terms of the inverse image has implications in physical situations too. The response function, susceptibility, defined as $\chi = \frac{\partial m(h)}{\partial h}$ loses its significance at $h = 0$. The relevant quantity in this situation is the inverse susceptibility χ^{-1} which may be defined as $\frac{\partial h(m)}{\partial m}$ in terms of the inverse function $h(m)$. In thermodynamics or statistical mechanics, the inversion is done by changing the ensemble. In a fixed magnetic field ensemble, the free energy $F(h)$ is a function of h while in a fixed magnetization ensemble the free energy $\mathcal{F}(m)$ is a function of the magnetization. These two free energies are related in thermodynamics by a Legendre transformation. By generalizing to free energy functionals, the two response functions are defined as

$$\chi(\mathbf{r}, \mathbf{r}') = \frac{\partial m(\mathbf{r})}{\partial h(\mathbf{r}')} = -\frac{\delta^2 F}{\delta h(\mathbf{r})\, \delta h(\mathbf{r}')}, \quad \chi^{-1}(\mathbf{r}, \mathbf{r}') = \frac{\partial h(\mathbf{r})}{\partial m(\mathbf{r}')} = \frac{\delta^2 \mathcal{F}}{\delta m(\mathbf{r})\, \delta m(\mathbf{r}')}, \tag{9.8}$$

such that $\int \chi^{-1}(\mathbf{r}, \mathbf{r}')\ \chi(\mathbf{r}', \mathbf{r}'') d\mathbf{r}' = \delta(\mathbf{r} - \mathbf{r}'')$. Such inverse response functions lead to vertex functions in field theories.

Problem 9.6.1: Consider a field theory like the ϕ^4 theory of Eq. (9.7), defined in a d-dimensional space. Show that the two point vertex function for this field theory corresponds to an inverse response function.

9.7 Quantum mechanics

A few examples of use of topology in elementary quantum mechanics are now discussed. These examples are not to be viewed in isolation but in totality with all other examples discussed in this chapter We shall mix classical examples here too to show the broadness of the topological concepts and topological arguments. Later on these ideas in the quantum context will be connected to another, completely classical, arena of biology involving DNA.

9.7.1 QM on multiply-connected spaces

We now consider quantum mechanics on topologically nontrivial spaces in quantum mechanics, for which we need to reexamine two traditional statements, namely,

1. Wave functions are single valued.
2. An overall phase factor, $\psi(x) \to e^{i\phi}\psi(x)$ does not matter.

The singlevaluedness criterion gets translated in the path integral formulation as the sum over all paths in the relevant configurational space with weights

[12] In this one-dimensional example, an interval has two boundaries. If one boundary is a member of the set but not the other one, then it is called a *semi-open* set.

determined by the action along the path. These are actually valid for simply-connected spaces, but not necessarily for a multiply-connected space. An example of such a case is the problem of a single particle on a ring which is discussed in some detail, avoiding a full fledged general analysis. The result can be extended to the case of a plane with a hole (See Prob. 2.13) as we shall do below.

9.7.2 Particle on a ring

A particle is constrained to move on a ring (a circle) of circumference L. We use the coordinate x to denote the position on the ring.

For a circle S^1 and its universal cover $\mathbb{R}$, refer to Fig. 9.1b,c and Fig 9.2b,c. To maintain generality, we use notations $\mathbb{F}$ and $\widetilde{\mathbb{F}}$ as the topological space in question, and its universal cover respectively. These are related by $\mathbb{F} = \widetilde{\mathbb{F}}/G$, where G is a discrete group expressing the equivalence relations. For the ring, $\mathbb{F} = S^1$, and $\widetilde{\mathbb{F}} = \mathbb{R}$, and $G = \mathbb{Z}$ (See Sec. 9.2.1). The requirement that $\pi_1(\widetilde{\mathbb{F}}) = 0$, sets $\pi_1(\mathbb{F}) = G$.

A trivial loop in $\mathbb{F}$, see Fig. 9.2, maps to a loop in $\widetilde{\mathbb{F}}$, whereas all nontrivial loops in $\mathbb{F}$ become open paths in $\widetilde{\mathbb{F}}$. A point q in $\mathbb{F}$ maps to many points (all equivalent) in $\widetilde{\mathbb{F}}$. Let us choose one such point $\widetilde{q}_0$ arbitrarily. A closed loop C in $\mathbb{F}$ from q_0 to q_0 maps to a unique path $\widetilde{C}$ in $\widetilde{\mathbb{F}}$ from $\widetilde{q}_0$ to another equivalent point $\widetilde{q}_0'$. The end point is the same for all paths homotopic to C (i.e., deformable to C). By denoting all such homotopic paths by $[C]$ or $[\widetilde{C}]$ as the case may be, we write $\widetilde{q}_0' = [C]\ \widetilde{q}_0$, without using any tilde on C.

It is now reasonable to expect

$$\widetilde{\psi}([C]\widetilde{q}) = a([C])\ \widetilde{\psi}(\widetilde{q}), \tag{9.9}$$

with a as a phase factor. Two loops C, C' in $\mathbb{F}$ based at q_0 can be combined into one[13], $[C''] = [C']\ [C]$. In $\widetilde{\mathbb{F}}$ the corresponding paths $\widetilde{C}$ connects $\widetilde{q}_0$ to $[C]\widetilde{q}_0$ while the subsequent $\widetilde{C}'$ connects $[C]\widetilde{q}_0$ to $[C'][C]\widetilde{q}_0$ which is also $[C'']\widetilde{q}_0$. For the wavefunction, we get the combination rule for the phase factors as

$$a([C'])a([C]) = a([C']\ [C]) = a([C'']), \tag{9.10}$$

i.e., a's follow the group multiplication rules of π_1. These a's therefore constitute a one-dimensional representation of the fundamental group.

A simple path in $\mathbb{F}$ connects any point q to q_0. Simple here means the path shrinks to a point as $q \to q_0$. Such paths allow us to map all the points of $\mathbb{F}$ to a domain containing $\widetilde{q}_0$ in $\widetilde{\mathbb{F}}$. For the circle case, this is reminiscent of a unit cell in $\mathbb{R}$. The equivalence relation or the nontriviality of G suggests that there are other equivalent domains, as many as the number of elements of G. As an example, in Fig. 9.1, the domains can be chosen as $(-\pi, \pi], (\pi, 3\pi]$..., an

[13]This rule, in fact, generates the fundamental group $\pi_1(\mathbb{F})$.

infinite of them as $\mathbb{Z}$ is countably infinite. The wavefunction is single valued in each of these domains but those in two different domains differ by a phase factor. As any of these domains is isomorphic to $\mathbb{F}$, any one of these wavefunctions can be taken as the wavefunction on $\mathbb{F}$. We end up with a multivalued wavefunction on $\mathbb{F}$ whose branches are the wavefunctions on the "unit cells" of $\widetilde{\mathbb{F}}$. In short, quantum mechanics on a multiply connected space requires a multivalued wavefunction, unlike the simple cases studied in Euclidean spaces.[14] This, fortunately, is not the end of the story. With the help of examples, we shall see that we may still choose single valued wavefunction, at the cost of an extra phase though. This is Berry's phase which goes beyond topology and appears in many problems as a geometrical phase

Let us consider a few special cases.

$a = 1$: single-valued wavefunction

The identity representation is the trivial representation of any group. Let us choose $a = 1$ for all elements of $G = \mathbb{Z}$. The free particle Hamiltonian

$$H = \frac{p^2}{2m}, \quad \text{with } H\psi(x) = E\ \psi(x), \quad \text{and} \quad \psi(0) = \psi(L). \tag{9.11}$$

The periodic boundary condition, which incorporates our requirement of $a = 1$, gives the known energy eigenfunctions and eigenvalues as

$$\psi_k(x) = e^{ikx}. \quad k = \frac{2\pi n}{L}, n \in \mathbb{Z}, \text{ and } E_n = \frac{2\pi^2\hbar^2 n^2}{mL^2}. \tag{9.12}$$

Importantly, the wavefunction is single-valued.

$a_n = e^{in\theta}$: multi-valued wavefunction

Let us now consider the case of multivalued wavefunction. By using gauge transformation, the multivaluedness is linked to the behaviour of a particle when the ring is threaded by a magnetic flux. A connection between the two problems is then obtained via Berry's phase.

A. Multi-valuedness

Let us choose, respecting Eq. (9.10), a unitary representation

$$a_n = e^{in\theta}, n \in \mathbb{Z}, \tag{9.13}$$

It is, as per Eq. (9.9), equivalent to a twisted boundary condition

$$\psi(0) = e^{-i\theta}\psi(L), \tag{9.14}$$

[14] An analogy: In the complex plane, $f(z) = \sqrt{z}$ is multivalued but it is single-valued on the extended Riemann sheets. Each sheet defines one branch of $f(z)$. Compare this with multivalued $\psi(q)$ on $\mathbb{F}$ but single-valued $\widetilde{\psi}$ on $\widetilde{\mathbb{F}}$.

thereby making the wavefunction multivalued.

The energy eigenvalues and eigenfunctions are still given by Eq. (9.12) but with

$$k = \frac{2\pi n + \theta}{L}, \text{ and } E_n = \frac{2\pi^2\hbar^2(n+n_0)^2}{mL^2}, \quad n \in \mathbb{Z}, \tag{9.15}$$

where $n_0 = \theta/(2\pi)$.

B. Gauge transformation, magnetic flux

We may do a gauge transformation for the wavefunction, $\Psi(x) = e^{i\Theta(x)}\psi(x) = U\psi(x)$, where U is the unitary transformation operator. Such a transformation changes the boundary condition to

$$\Psi(L) = e^{i\Theta(L)}\psi(L) = e^{i\Theta(L)}e^{i\theta}\psi(0) = e^{i\Theta(L)}e^{-i\Theta(0)}e^{i\theta}\Psi(0). \tag{9.16}$$

The choice

$$\Theta(L) - \Theta(0) = -\theta, \text{ or }, \Theta(x) = -\frac{\theta}{L}x, \tag{9.17}$$

gives us a θ-independent boundary condition, $\Psi(L) = \Psi(0)$ as in Sec. 9.7.2. Moreover, a direct substitution shows that $\Psi(x)$ is the eigenfunction of a transformed Hamiltonian[15] but with the same energy,

$$H_\theta = e^{i\Theta(x)}He^{-i\Theta(x)} = \frac{1}{2m}\left(p + \frac{\hbar\theta}{L}\right)^2, \quad \text{and} \quad H_\theta\Psi_n(x) = E_n\Psi_n(x), \tag{9.18}$$

where E is given by Eq. (9.15) The quantum problem turns out to be equivalent to a charged particle in a magnetic vector potential, but with a *θ-independent* periodic boundary condition for the wave function.

Suppose there is a thin solenoid of radius b carrying a magnetic field B at the center of the ring in a direction perpendicular to the plane of the ring . There is no magnetic field on the ring but there exists a vector potential

$$A = \frac{B\pi b^2}{2\pi r}, \tag{9.19}$$

on a circle of radius r in the angular direction. The Hamiltonian of a particle of charge q is then $H_{mag} = \frac{1}{2m}(p - qA/c)^2$, c being the velocity of light. Comparing this form with Eq. (9.18), we see that $\theta = 2\pi\frac{\Phi}{\Phi_0}$, where $\phi_0 = 2\pi\hbar q/c$, the standard flux quantum if the charge q is the electronic charge e.

It needs to be recognized at this point that the circular ring per se is not special here; geometry does not matter. The analysis is valid for a two dimensional plane threaded by an impenetrable thin flux line, which is like a hole (See Prob. 2.13). Phase θ is independent of path, allowing us to claim that it is topological in nature. This phase is known as the Aharonov-Bohm phase.

[15] Under a unitary transformation $|\psi'\rangle = U|\psi\rangle$, the average of an operator A remains the same so that $\langle\psi|A|\psi\rangle = \langle\psi'|A'|\psi'\rangle = \langle\psi'|U^\dagger AU|\psi'\rangle$, identifying the transformed operator $A' = U^\dagger AU$.

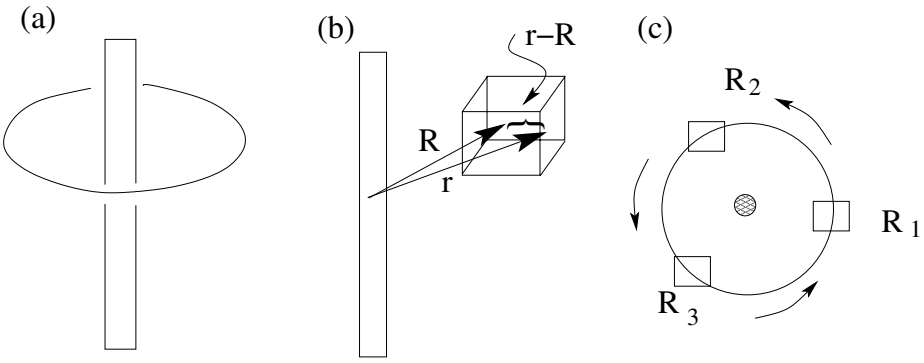


Figure 9.9: (a) A particle on a ring threaded by a flux line through origin. (b) A box, confining the particle, centered on **R** on the circle. The position of the particle inside the box is **r**. (c) The boxed is taken around the full circle adiabatically in a continuous manner, or, say, in three steps so as to enclose the flux.

9.7.3 Topological/Geometrical phase

The extra phase we derived in the preceding case is not the only one. Such phases occur in many situations, classical or quantum. For waves in classical physics, such a phase is called the Pancharatnam phase while in quantum mechanics it is Berry's phase. As an angle it can be found in classical problems as Hannay angle, writhes in DNA and so. Often the angle may be geometric in origin, not necessarily topological. What it means is the the angle one gets may depend on the path chosen unlike the path independence of a topological quantity. In this respect it is important to distinguish between a geometrical phase and a topological phase.

Berry's phase

We established the equivalence of two QM problems, viz.,

1. a charged particle on a ring enclosing a constant flux, with *single valued* wave function,

2. a particle on a ring in zero flux but with a *multivalued* wave function.

This equivalence raises the following question,

> What is the analogue of the extra phase (case 2) responsible for the multivaluedness in the singlevalued version of case 1?

The answer lies in Berry's phase.

Phase – an Angle: two formulas

In order to see the emergence of the phase, let us take a wavepacket localized on the ring. One way to achieve this is to enclose the particle in a small box centered on a point on the ring, and then move the box around the circle. Let the box position (centre of the box) be $\mathbf{R}$ on the ring. The particle is now described by

$$\left[\frac{1}{2m}\left(p-\frac{qA(\mathbf{r})}{c}\right)^2+V(\mathbf{r}-\mathbf{R})\right]\chi(\mathbf{r},\mathbf{R})=\mathcal{E}\chi(\mathbf{r},\mathbf{R}), \tag{9.20}$$

where $A(\mathbf{r})$ is given by Eq. (9.19). Within the finite-sized box the wavefunction is single valued.

Now this box is taken adiabatically around the circle through a complete turn as shown in Fig 9.9. We assume that the wave function changes very slowly, though, in fact, we shall see that the speed doe not matter. For two positions $\mathbf{R}_1, \mathbf{R}_2$, the phase difference is defined as

$$\exp(-i\phi_{12})=\frac{\langle\chi(\mathbf{r},\mathbf{R}_1)|\chi(\mathbf{r},\mathbf{R}_2)\rangle}{|\langle\chi(\mathbf{r},\mathbf{R}_1)|\chi(\mathbf{r},\mathbf{R}_2)\rangle|},\ \text{ or }\ \phi_{12}=-\text{Im }\ln\langle\chi(\mathbf{r},\mathbf{R}_1)|\chi(\mathbf{r},\mathbf{R}_2)\rangle. \tag{9.21}$$

This phase between two points is somewhat arbitrary because it can be changed by a gauge transformation at any one of the points $\mathbf{R}_1$ or $\mathbf{R}_2$. In spite of this arbitrariness, it can be combined with the phases from the remaining steps to define an overall phase as ($\mathbf{r}$ is suppressed in the notation)

$$\gamma=-\text{Im}\ln\left(\langle\chi(\mathbf{r},\mathbf{R}_1)|\chi(\mathbf{r},\mathbf{R}_2)\rangle\langle\chi(\mathbf{r},\mathbf{R}_2)|\chi(\mathbf{r},\mathbf{R}_3)\rangle.\langle\chi(\mathbf{r},\mathbf{R}_3)|\chi(\mathbf{r},\mathbf{R}_1)\rangle\right) \tag{9.22}$$

Note that the arbitrariness of phases at the intermediate points get cancelled in the product. Consequently the total phase γ modulo 2π is a phase that cannot be removed by a gauge transformation, and, as a gauge invariant quantity, must have physical consequences. This phase is an example of what is called *Berry's phase*.

We may take a continuum limit where the closed loop is traversed by infinite number steps, with $\mathbf{R}_{i+1}=\mathbf{R}_i+d\mathbf{R}$. Using continuity, $\chi\mathbf{R}+d\mathbf{R})=\chi(\mathbf{R})+\nabla_R\chi(\mathbf{R})\cdot d\mathbf{R}$, and then expanding the logarithm, the phase factor can be written as an integral (taking the wavefunctions to be normalized)

$$\gamma=i\oint d\mathbf{R}\cdot\langle\chi(\mathbf{R})|\nabla_R\chi(\mathbf{R})\rangle, \tag{9.23}$$

taking the wavefunctions to be normalized.

Whenever a Hamiltonian depends on a parameter (no quantum evolution of this parameter) and the parameter goes through a cyclic path in the parameter space, the parameter-dependent wavefunction develops a phase given by either Eq. (9.22) or Eq. (9.23). In numerical computation where the eigenfunctions are determined numerically – and therefore with unknown phases – Eq.

(9.22) is preferable because, by construction, the intermediate unknown phases cancel out. In many analytical approaches for which a continuous wavefunction is known, Eq. (9.23) is useful.

Berry's phase and the Aharonov-Bohm phase

We now show that the unavoidable Berry's phase in the formulation with a magnetic flux is the extra phase in the multivalued formulation.

Eq. (9.20) can be solved by a gauge transformation $\chi = e^{ig(x)}\widetilde{\chi}$ so that

$$\left[\frac{p^2}{2m} + V(\mathbf{r}-\mathbf{R})\right]\widetilde{\chi}(\mathbf{r}-\mathbf{R}) = \mathcal{E}\widetilde{\chi}(\mathbf{r}-\mathbf{R}), \quad \text{with} \quad g(x) = \frac{q}{\hbar c}\int_{\mathbf{R}}^{\mathbf{r}} A(\mathbf{r}')d\mathbf{r}. \tag{9.24}$$

This is very similar to what we did earlier in Sec B2, but here the wave function $\widetilde{\chi}$ remains singlevalued, mainly because the interior of the box is a simply-connected region.

To use Eq. (9.23), we need $\nabla_R\chi(\mathbf{r},\mathbf{R})$ which can be written as[16]

$$\nabla_R\chi(\mathbf{r},\mathbf{R}) = \nabla_R\{e^{ig}\widetilde{\chi}(\mathbf{r}-\mathbf{R})\} = -i\frac{q}{\hbar c}A(\mathbf{R})\chi - e^{ig}\nabla_r\widetilde{\chi}(\mathbf{r}-\mathbf{R})\}, \tag{9.25}$$

so that (taking normalized $\widetilde{\chi}$)

$$\langle\chi(\mathbf{R})|\nabla_R\chi(\mathbf{R})\rangle = -i\frac{q}{\hbar c}A(\mathbf{R}) - \langle\widetilde{\chi}(\mathbf{r}-\mathbf{R})|\nabla_r\widetilde{\chi}(\mathbf{r}-\mathbf{R})\rangle. \tag{9.26}$$

As the average momentum in the localized state in the box is zero, we obtain the overall phase on taking the box around the loop once

$$\gamma = \oint \frac{q}{\hbar c}A(\mathbf{R})\cdot d\mathbf{R} = \frac{q}{\hbar c}\int d\mathbf{S}\cdot\nabla_R\times A(\mathbf{R}) = \frac{q}{\hbar c}\Phi = 2\pi\,\frac{\Phi}{\Phi_0} = \theta, \tag{9.27}$$

precisely the same angle we saw in Eq. (9.14). The line of arguments here indicates that the angle is independent of the size and shape of the loop in the plane so long it encloses the origin once. The answer is ultimately determined by the number of times (=1) the flux tube pierces the surface used in the surface integral. The Aharonov-Bohm phase, viewed as Berry's phase, is therefore topological in essence.

To summarize, in a topologically nontrivial space, we may either use multivalued wavefunctions or use a gauge transformation to a magnetic field like problem with singlevalued wavefunction that admits a geometric phase, Berry's phase. A generalization, without proof, is that a topological phase (Berry's phase) occurs[17] if (i) a parameter, R, defined on a multiply-connected space, is taken around a nontrivial loop, and (ii) the space of the wavefunctions remains the same as the parameter is changed, i.e., the Hilbert space is independent of R.

[16] Note that $\nabla_R f(\mathbf{r}-\mathbf{R})) = -\nabla_r f(\mathbf{r}-\mathbf{R}))$

[17] With complex wavefunctions, Berry's phase may occur in simplyconnected spaces too.

Another lesson we learnt from this is that a hole or impenetrable region can be replaced by a vector potential or a gauge term where the θ parameter determines the effective flux. Such assignments of flux tubes become useful in many situations like anyons.

9.7.4 Generalization – Connection, curvature

Proper settings for generalization of the above ideas require the concepts of fibre bundles and differential forms. Without going into those, let us define the terms – with a little abuse of definitions – connection and curvature. The vector potential $\mathbf{A}$ we defined is called Berry connection, though actually it should be a 1-form meaning something like $\mathbf{A}\cdot d\mathbf{r}$. The "magnetic field" is the curvature. Again, curvature should be a 2-form meaning objects like $\mathbf{B}\cdot d\mathbf{s}$ with $d\mathbf{s}$ as the area element. In our convention, the integrals will have the infinitesimals dr, ds explicitly.

The general formulas for connection and curvature for a state, labelled by m, and, dependent on a set of parameters R_μ, are

$$\text{Berry connection}: \; A_\mu = i\,\langle mR\,|\partial_\mu|\,mR\rangle\,, \quad \left(\partial_\mu \equiv \frac{\partial}{\partial R_\mu}\right), \tag{9.28}$$

$$\text{Berry curvature}: \; \Omega_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \tag{9.29}$$

In three dimensions, i.e., if R has three components, the curvature tensor can be written as a vector,

$$B_\lambda = \frac{1}{2}\, i\, \varepsilon_{\lambda\mu\nu}\; \left(\partial_{R_\mu}\langle mR|\right)\; \left(\partial_\nu|mR\rangle\right), \tag{9.30}$$

where $\varepsilon_{\lambda\mu\nu}$ is the usual antisymmetric tensor. The state index m has been omitted from the notation of A, Ω. The connection is like the vector potential, while, from Eq. (9.30), $\mathbf{B} = \nabla\times\mathbf{A}$ is like a magnetic field. For generality, instead of linking these to electromagnetism, we call $\Omega_{\mu\nu}$ as the Berry curvature per unit area. An integral of the curvature over an open surface S (with boundary) gives the phase (Berry's phase) associated with the closed loop, the boundary of S.

For a given Hamiltonian $H(R)$, every eigenstate will have its own Berry connection and Berry curvature. Defining the eigenvalue equation as $H(R)|nR\rangle = E_n|nR\rangle$, with no degeneracy, a straightforward manipulation shows that the Berry curvature for the nth state is (see problem)

$$\Omega^n_{\mu\nu} = i\sum_{p\neq n} \frac{\langle nR|\partial_\mu H(R)|pR\rangle\langle pR|\partial_\nu H(R)|nR\rangle}{[E_p(R)-E_n(R)]^2} - \{\mu\leftrightarrow\nu\}, \tag{9.31}$$

which has the advantage that the derivatives are now of the Hamiltonian and not of the wavefunctions. It also follows from the antisymmetric nature that

$$\sum_n \Omega^n_{\mu\nu} = 0. \tag{9.32}$$

Eq. (9.31) shows that the Berry curvature is large for "near degeneracies" or, equivalently, a large Berry curvature can be taken as a signal for nearby eigenvalues. If there is a degeneracy, then one has to project out that part of the space leading to nonabelian issues.

The degeneracy points are singular points in the d-dimensional parameter space. The loops for Berry's phase need to enclose these singular points for a nonzero value. The loops actually tell us about the first Homotopy group of the allowed part of the space as the loop is not allowed to go through the singular point. Loops in 2-dimensions enclose point defects, in 3-dimensions line defects and so on, so that the singular points in d-dimensions must form a $(d-2)$-dimensional space (or manifold) for the first homotopy group of the allowed space to be nontrivial. This restriction provides a quick check when not to expect any topological phase.

9.7.5 Chern, Gauss-Bonnet

In the examples we consider, the space of the parameter R is an even dimensional closed surface $\mathcal{M}$. For concreteness, let $\mathcal{M}$ be a two-dimensional closed orientable surface that can be embedded in three dimensions. By orientable we mean at at each point we can define a unique normal to the surface. Simple examples are $S^2, \mathbb{T}^2$, etc. There are now two topological problems in hand. One is the topological characterization of $\mathcal{M}$ and the other one is that of the map from $\mathcal{M}$ to the manifold of wavefunctions. Two theorems are useful here, (i) the Gauss-Bonnet theorem involving the geometric curvature of $\mathcal{M}$ and Chern's theorem involving the Berry curvature.

Chern's theorem states that the integral of the Berry curvature (a geometric quantity) over the closed surface is equal to 2π times an integer, i.e.,

$$C_1 = \frac{1}{2\pi}\int_{\mathcal{M}} \boldsymbol{\Omega}\cdot d\mathbf{S} = n \in \mathbb{Z}, \qquad (\text{Chern's theorem}). \tag{9.33}$$

This number C_1 is called the *first Chern number*.[18] Two mappings (or states in this case) with different first Chern numbers cannot be continuously deformed into each other. In other words, to go from one to the other by tuning some parameter (not R), there has to be a topology change at some special value of the parameter. This corresponds to a phase transition or a quantum critical point. Of these, $C_1 = 0$ is called a trivial phase while $C_1 \neq 0$ are nontrivial topological phases.[19] As an analogy one may refer to Fig. 9.2c, where the oscillatory and the circular motions are separated by the special figure 8 space.

---

[18]The $1/(2\pi)$ factor actually comes from a general factor $2/K_d$, where $K_d = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ is the volume of S^{d-1} or the surface area of a d-dimensional sphere, which occurs for the theorem for a higher dimensional closed surface. This is applicable to the Gauss-Bonnet theorem too. For Eq. (9.33), put $d = 3$.

[19]Warning: Phase here means a state of the system like liquid, gas etc, and not the phase of a wavefunction!

The Chern number is a topological characteristic of the manifold of the energy eigenstate defined on $\mathcal{M}$ and is not just a topological property of $\mathcal{M}$. The topological characteristic of $\mathcal{M}$ comes from the Gauss-Bonnet theorem, which for a closed two dimensional surface states that the surface integral of the Gaussian curvature is a topological quantity, viz.,

$$\frac{1}{2\pi}\int KdS = \chi = 2(1-g), (\text{Gauss} - \text{Bonnet theorem}) \tag{9.34}$$

where χ is the Euler characteristic and g is the genus of the surface. Both χ and g are topological properties of any surface. For a sphere of radius r, the Gaussian curvature (=product of the two principal curvatures at a point) is uniform, $K = 1/r^2$, and its genus $g = 0$. Eq. (9.34) is then obviously satisfied.

9.7.6 Classical context: geometric phase

To show that the angle is not just a quantum mechanical issue, let us take a classical example.

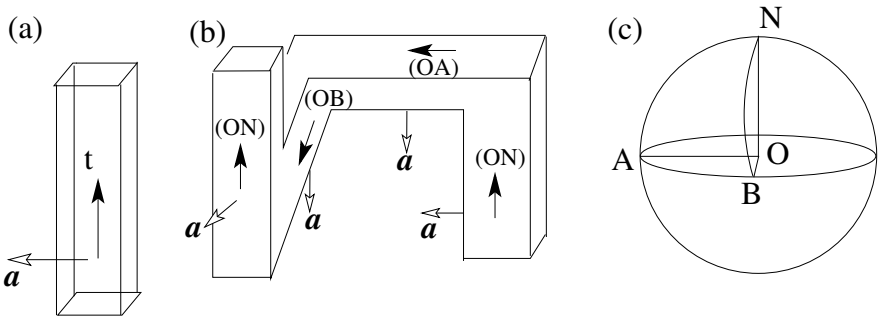


Figure 9.10: (a) A long bar of rectangular cross section. Unit vector **t** is along the axis, and **n** perpendicular to one side. (b) The bar is bent into a twisted form. The two vectors at intermediate positions are as shown. A change in the orientation of **n** by $\pi/2$ is visible as **t** returns to its original orientation. (c) S^2 as the space for vector **t**. As one moves along the bar, **t** goes from ON to OA, OA to OB, and then back to ON, a closed path on S^2. The solid angle subtended by the closed region NABN is $\pi/2$. In the quantum mechanics problem, there is an extra factor of 1/2 (see Eq. (9.37)).

Take a long bar of rectangular cross section so as to identify the sides easily. See Fig. 9.10. Define unit vectors **t** along the axis, and **a** perpendicular to one side. Orientations are to be kept fixed locally. The bar is bent into a twisted form as in Fig. 9.10b. The two vectors are monitored along the tube, keeping their orientations fixed locally. A $\pi/2$ change in the orientation of **a** is visible even when **t** gets back to its original orientation. To see this change, we note that vector **t**, as a unit vector, spans a sphere S^2 as in Fig. 9.10c. As one

moves along the bar, **t** goes from ON to OA, OA to OB, and then back to ON, a closed path on S^2. The solid angle subtended by the closed region NABN is 1/8 of the sphere, i.e., $\pi/2$ which is the change in orientation of **n**. One may straighten the bar in Fig. 9.10b to see that there is a twist though the axis may remain straight.

This is an example of a geometric phase, not necessarily topological, because the solid angle subtended by a closed loop depends on the details of the loop. The rotation of the Foucault pendulum as the earth rotates under it is another example of this geometric phase. This particular example and its generalizations are important in macromolecules like DNA and is called *twist*.

For a classical wave, polarized light, with **n** as the polarization direction and **t** as the direction of propagation, one may see a change in the direction of polarization. The angle appears as a phase there and is the classical analog of Berry's phase. The classical phase is called the *Pancharatnam phase*. In classical dynamics, such an angle also occurs and is known as the *Hannay Angle*. For details on these see Ref. [7]

Analogous to this classical example, Berry's phase is generally a geometrical phase, and in special situations, like the Aharonov-Bohm case, it becomes a topological phase. To repeat, a topological phase is independent of the details of the path (same for all homotopic loops) while a geometrical phase is, in general, dependent on the details of the loop.

9.7.7 Examples: Spin-1/2, Quantum two level system, Chern insulators

Spin-1/2 in a magnetic field

A counterpart of the problem discussed in Sec. 9.7.6 is a spin 1/2 in a magnetic field with a Hamiltonian $H = -\mathbf{d}.\boldsymbol{\sigma}$, where $\boldsymbol{\sigma}$ is the 3-d vector of the Pauli spin matrices (See Sec 9.4.4). For a given field **d**, there are two eigenstates, $\pm|d|$ with eigenvectors parallel or antiparallel to the direction of **d**. As a vector, the field of constant magnitude $|d|$ can be in any direction in 3-dimensions, $\mathbf{d} = |d|\ \ (\sin\theta\cos\ \phi, \sin\ \theta\sin\ \phi, \cos\ \theta)$ with θ, ϕ as the usual polar angles, spanning a sphere. The relevant space for the field is S^2. Let us choose the eigenstate for energy $+|d|$, (compare with Eq. (9.3)),

$$|u\rangle = \begin{pmatrix} \sin\frac{\theta}{2}e^{-i\phi} \\ -\cos\frac{\theta}{2} \end{pmatrix}, \tag{9.35}$$

upto an arbitrary phase factor. This $|u\rangle$ is a spinor, but, is, otherwise, fixed by the direction of the field **d**. If the magnetic field is now rotated, it forms a closed loop on the sphere. To get the phase acquired by the wavefunction as

given by Eq. (9.27), embed the sphere in three dimensions [20] to obtain

$$\mathbf{A} = \langle u|i\nabla_{\mathbf{d}}|u\rangle = \frac{\sin^2\frac{\theta}{2}}{|d|\ \sin\theta}\hat{\phi} = \frac{1}{2}\ \frac{1-\cos\theta}{|d|\ \sin\theta}\,\hat{\phi},\ \text{with}\ \mathbf{\Omega} = \nabla\times\mathbf{A} = -\frac{1}{2}\frac{\mathbf{d}}{|d|^3}, \tag{9.36}$$

in the radial direction. The line integral may be converted to a surface integral with the help of Stokes' theorem, by choosing the enclosed part of the sphere as the relevant surface. As the area vector is radial, we see

$$\gamma = \int \mathbf{\Omega}\cdot d\mathbf{s} = -\frac{1}{2}\int d\omega = -\frac{1}{2}\ \mho_c, \tag{9.37}$$

$d\omega$, being the angular part of the spherical integral and $\mho_c$ is the solid angle formed by the closed loop. The similarity with the classical case of Fig 9.10 is to be recognized, except for the factor of $1/2$ which is purely a quantum mechanical contribution. A solid angle is a geometrical quantity, dependent on the closed path, and so the phase here, unlike the Aharonov-Bohm phase, is not topological. The form of the "magnetic field" $\mathbf{B}$ shows that there is a singularity at $d = 0$, the degeneracy point and the functional form of $\mathbf{\Omega}$ satisfies $\nabla\cdot\mathbf{\Omega} = 4\pi\ \delta(\mathbf{r})$ with $g = -1/2$ and $\delta(\mathbf{r})$ as the Dirac δ-function. With $\mathbf{\Omega}$ as a "magnetic field", it looks like there is a magnetic monopole $g = -1/2$ at the center, and the Berry phase is the flux due to this monopole through the loop.

The magnetic monopole interpretation helps us in identifying a topological invariant, using the equivalent of the Gauss theorem. An integration over the whole sphere is the flux through the closed surface and it counts the number of monopoles enclosed by the sphere. In other words

$$C_1 \equiv \frac{1}{2\pi}\ \int_{S^2}\mathbf{\Omega}\cdot d\mathbf{s} = n \in \mathbb{Z}. \tag{9.38}$$

This is the *first Chern number* as defined in Eq. (9.33).

It is important to see from a different view point why $C_1 \neq 0$. Let us divide the surface integral into parts at the equator. The closed loop integral $\oint \mathbf{A}\cdot d\mathbf{l}$ can be evaluated with the help of Stokes' theorem with either the upper hemisphere containing the North pole or the lower hemisphere containing the South pole, provided $\mathbf{A}$ can be defined uniquely on the chosen surface and the equator. Since Eq. (9.35) is valid for the part of the sphere that encloses the North pole, we get, for the upper hemisphere (uh), $\oint \mathbf{A}\cdot d\mathbf{l} = \int_{\text{uh}} d\mathbf{s}\cdot\mathbf{\Omega}$ which evaluates to $+\pi$. However, this choice of $|u\rangle$ cannot be used for the lower hemisphere (lh), as the spinor, Eq. (9.35), has undefined phase at the South

[20] We are embedding to take advantage of the vector notation. In spherical polar coordinates, for any vector $\mathbf{A} = A_r\hat{r} + A_\theta\hat{\theta} + A_\phi\hat{\phi}$,

$$\nabla = \hat{r}\frac{\partial}{\partial r} + \hat{\theta}\ \frac{1}{r}\ \frac{\partial}{\partial\theta} + \hat{\phi}\frac{1}{r\sin\theta}\ \frac{\partial}{\partial\phi},\quad \nabla\times\mathbf{A} = \begin{pmatrix} (r^2\sin\theta)^{-1}\hat{r} & (r\sin\theta)^{-1}\hat{\theta} & r^{-1}\hat{\phi} \\ \partial/\partial r & \partial/\partial\theta & \partial/\partial\phi \\ A_r & rA_\theta & r\sin\theta A_\phi \end{pmatrix}.$$

pole ($\theta = \pi$) where $|u\rangle = (e^{i\phi}\ 0)^T \equiv (1\ \ 0)^T$. A possible choice ("choice of gauge") is $|u_{lh}\rangle = (\sin(\theta/2), -\cos(\theta/2)e^{i\phi})^T$ which is now defined everywhere on the sphere except the North pole ($\theta = 0$). With this choice $\mathbf{A} = -\frac{1}{2}\,\frac{1+\cos\theta}{|d|\,\sin\theta}\hat{\phi}$, though the curvature ($\Omega$) remains the same as Eq. (9.36). A direct line integral over the equator ($\theta = \pi/2$) gives $-\pi = \pi - 2\pi$. In other words it differs from the upper hemisphere result by -2π which does not matter as an angle. If Stokes' theorem is used, $\oint \mathbf{A}\cdot d\mathbf{l} = -\int_{lh} \mathbf{\Omega}\cdot d\mathbf{s}$ with a minus sign coming from the direction rule of the theorem. Since it is the same line integral, We must have $(\int_{uh} + \int_{lh})\mathbf{\Omega}\cdot d\mathbf{s} = 0$ upto $2\pi n, n \in \mathbb{Z}$. In this particular case, we find -2π whose origin lies in the nonuniformity of the gauge choice. If a single gauge choice can be done over the whole surface, then the surface integral would have been zero. Thus the zero Chern number corresponds to the trivial case, while a nonzero Chern number tells us that more than one map is needed to cover the whole closed surface. In this case we need two.

For the S^2 case, the state is given by $-\hat{n}$ which is also the normal to the surface. The Berry curvature, Eq. (9.29) is given by

$$\Omega_{lm} = -\frac{1}{2}\hat{n}\cdot\left(\frac{\partial\hat{n}}{\partial R_l}\times\frac{\partial\hat{n}}{\partial R_m}\right), \tag{9.39}$$

where the derivatives are now the angular derivatives, and it is same as $\mathbf{B}$ in Eq. (9.36). The Chern number can therefore be written as

$$C_1 = -\frac{1}{4\pi}\int_{S^2} ds\ \ \hat{n}\cdot\left(\frac{\partial\hat{n}}{\partial\theta}\times\frac{1}{\sin\theta}\frac{\partial\hat{n}}{\partial\phi}\right) = -1. \tag{9.40}$$

The analysis for Berry's phase can be done for the other state $|v\rangle$ with eigenvalue $-|d|$. In that case the monopole will be of charge $+1/2$ so that the total of all the states is zero, consistent with Eq. (9.32).

Case of two bands: Chern insulators

The results of the two level system in the previous section finds a ready use in a two band system. This situation arises for a band insulator where we may focus on the last occupied band and the next unoccupied band.

The bands are described by the pseudomomentum $\mathbf{k}$ in the first Brillouin zone. The Hamiltonian can be written as $H = -\mathbf{d}(\mathbf{k})\cdot\boldsymbol{\sigma}$, with the bands given by $\pm|d(\mathbf{k})|$. For a two dimensional problem, the k-space (or the first Brillouin zone) is a torus $\mathbb{T}^2$, while unit vector $\hat{n}(\mathbf{k}) = \mathbf{d}(\mathbf{k})/|d(\mathbf{k})|$ maps out a sphere S^2. It may be visualized as a three component vector $\mathbf{n}$ attached to every point of the Brillouin zone torus. One may compare with the Heisenberg magnet example of Sec. 9.4.1, where we looked at the arrangements of three-vectors in Euclidean space. Here, instead of the Euclidean space, we now have a torus. All possible insulators can now be characterized by the "spin arrangements" on the torus. In the real space case of magnets, we found $\pi_1(S^2) = 0$, and so there is no topological distinction among the spin configurations in space. In the present

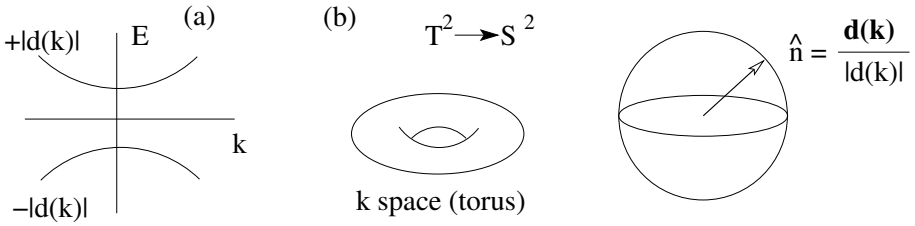


Figure 9.11: (a) Schematic diagram of two bands vs k. The band energies are $\pm|d(\mathbf{k})|$. (b) In two dimensions the k-space is a torus $\mathbb{T}^2$ while unit vector $\hat{n}(\mathbf{k}) = \mathbf{d}(\mathbf{k})/|d(\mathbf{k})|$ lies on a sphere. This gives a $\mathbb{T}^2 \to S^2$ map.

case, the mapping is nontrivial, and that tells us that all insulators are not topologically equivalent. In other words, there are topologically inequivalent classes of insulators, identified by the arrangements of the gap vector $\mathbf{d}$ on the Brillouin Zone, which is a torus. The nontrivial ones are called topological insulators, even though the band structure looks the same.

As we move along the torus, the $\mathbf{k}$-vector changes, and it acts as the parameter for Berry's phase or equivalently the curvature. The topological invariant is then the first Chern number that tells us how many times (with sign) vector $\hat{n}$ goes around the sphere as one traverses the torus. This number comes from the Chern formula

$$C_1 = \frac{1}{4\pi} \int\int d^2k \;\; \hat{n} \cdot \left(\frac{\partial \hat{n}}{\partial k_x} \times \frac{\partial \hat{n}}{\partial k_y} \right). \tag{9.41}$$

We have talked about the homotopy groups which allows one to explore a space by spheres S^n for various integer n. But, instead of spheres, we may explore the space by tori as well. For the problem in hand, we are classifying the configurations of 3-component vectors on the torus. This is described by homotopy$[\mathbb{T}^2, S^2]$ which is known to be $\mathbb{Z}$. The first Chern number is a concrete way of getting this integer for a particular case.

Let us choose an example,

$$\mathbf{d}(\mathbf{k}) = (\sin k_x, \sin k_y, r + \cos k_x + \cos k_y)). \tag{9.42}$$

If $|r|$ is very large, $\hat{n}$ is nearly along $(0, 0, \pm 1)$, and so does not wind completely the sphere. The Chern number is zero as can be checked from Eq. (9.41) for $|r| > 2$. One can easily check that $C_1 = 1$ if $-2 < r < 0$, while $C_1 = -1$ if $0 < r < 2$. There is topology change but the change is not obvious from the band structure. The Chern number is not defined at $r = 0, \pm 2$, the transition points. Exactly at $r = -2$, we see $|d| = 0$ at $k_x = k_y = 0$; this means the gap closes (no longer an insulator) at one point. For $r = 2$, the gap closing occurs at (π, π), while for $r = 0$, it occurs at $(\pi, 0)$ and $(0, \pi)$. The mapping to a sphere fails when a gap closes, and, consequently, a gap closing is important for a change of topology.

Band insulators with first Chern number=0 are called trivial insulators while those with $C_1 \neq 0$ are called Chern insulators. [9].

Problem 9.7.1: If the Hamiltonian for a two level system is real, i.e., $d_y = 0$, show that Berry's phase can be 0 or π.

Problem 9.7.2: General spin case: A spin-J particle is in a magnetic field with Hamiltonian $H = -\mathbf{d}\cdot\mathbf{J}$. Consider any of the eigenstates, say, state $|J\ M_J\rangle$ which is an eigenstate of the Hamiltonian. Calculate the Berry phase when $\mathbf{d}$ rotates as in Sec. 9.7.7. It is easier to use Eq. (9.31). The answer is $-M_J\mho_c$, generalizing Eq. (9.37).

Problem 9.7.3: Take the problem of a particle on a ring of radius 1, in presence of a magnetic flux threading the ring. This is the problem discussed above but there is now a potential $V(x) = v\delta(x)$ on the ring so that the Hamiltonian is

$$H = \frac{1}{2m}(p+\theta)^2 + v\delta(x), \tag{9.43}$$

under periodic boundary condition $\psi(x) = \psi(x+2\pi)$. Use topological arguments and necessary gauge transformations to identify this problem as the Dirac comb problem (one dimensional Kronig Penny model). Use Bloch's theorem to show that θ plays the role of the quasi-momentum.

Problem 9.7.4: For a one dimensional model, the Brillouin zone is a circle. The two band problem then corresponds to a mapping $S^1 \to S^2$. Discuss the nature of this mapping.

Discuss the general d-dimensional case, $T^d \to S^2$.

Problem 9.7.5: Complete the calculations for the Berry phase, Berry curvature and the monopole, counterparts of Eqs. (9.36)- (9.40), for the eigenstate with eigenvalue $-|d|$.

Problem 9.7.6: Take the example of Eq. (9.42) for three different values of r, $r = -1, 1, 3$. Draw the energy bands (3d plot) against k_x, k_y. Separately, map out the region on a sphere traced out by $\mathbf{n} = \mathbf{d}(\mathbf{k})/|d(\mathbf{k})|$ as $\mathbf{k}$ is taken over the Brillouin zone.

Problem 9.7.7: The two bands in Sec. 9.7.7 are taken to have a special symmetry so that the midpoint of the gap is independent of $\mathbf{k}$. In general, the form of the Hamiltonian should be $H = -\mathbf{d}(\mathbf{k})\cdot\boldsymbol{\sigma} + d_0(\mathbf{k})\mathbf{I}$, so that the space for the Hamiltonian is 4-dimensional. Show that the arguments of that section are not affected by $d_0(\mathbf{k})$. In other words, it is justified to consider a 3-dimensional subspace.

Problem 9.7.8: For the Aharonov-Bohm geometry, we saw the importance of winding around the hole. Consider the free particle case in a plane with a hole at origin. By a transformation $t = i\tau$ (imaginary time transformation), the Schrödinger equation can be converted to a diffusion equation which describes a Brownian particle in continuum. The winding of the Brownian particle around the hole can

be measured by making the angle θ a real variable (refer to Fig. 9.1b,c). Show that for such a Brownian particle in a plane with a hole at origin, the probability distribution for winding angle $\theta(t)$ for large t is

$$P(n) = \frac{1}{\pi}\,\frac{1}{1+x^2}, \quad \text{where } x = \frac{2\theta(t)}{\ln t}. \quad \text{(Spitzer law)} \tag{9.44}$$

The infinite variance of this distribution is because of the large number of very small windings a particle can do around the hole.

9.8 DNA

A situation where the topology of two circles is needed is DNA. We saw the importance of two circles S^1 in Fig. 9.3. DNA involves a different type of topological problem.

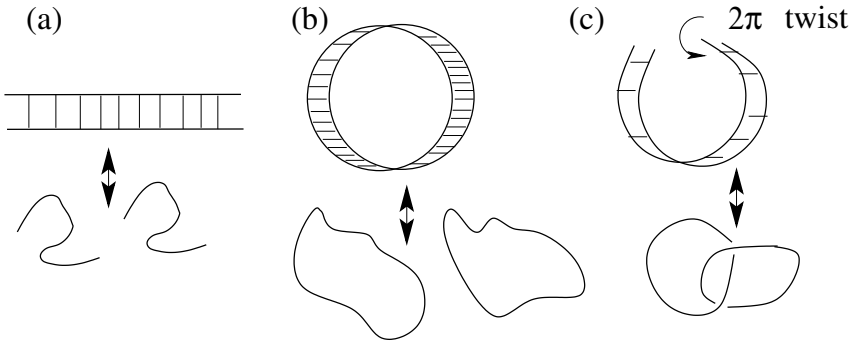


Figure 9.12: DNA. (a) open DNA with N base pairs. On melting it gives two separate strands. (b) A circular DNA on melting gives two separate rings. (c) If a 2π twist is given before forming the closed loop, then on melting there are again two nonpaired rings but now topologically linked. Helicity of the DNA is ignored here.

Double stranded DNA consists of two chains, called strands, connected together by base pairings. Suppose there are N base pairs, making the strand length proportional to N. See Fig 9.12. In the completely bound state, the energy is $E = -N\epsilon$, where $-\epsilon$ is the hydrogen bond energy and N is large. If the double stranded DNA has an entropy s_b per base pair, then the bound state free energy is given by $F_b = -N\epsilon - TNs_b$ at temperature T. On the other hand if all the hydrogen bonds are broken, then there are two nonpaired separate chains, each with entropy s_0 per base. Since there is no energetic contribution, the free energy of the unbound state $F_u = -2NTs_0$. Evidently, the entropy per base of a single strand is higher than that of a double stranded DNA, and so a phase transition from the bound to the unbound state is possible at $T_c = \epsilon/(2s_0 - s_b)$. This is called the melting transition of a DNA. A real melting

is a slightly more complicated phenomenon but this simple minded picture is sufficient here for our purpose. DNA strands can also be separated by force at $T < T_c$ and that is called unzipping transition.

DNA strands need to be separated because, as per semi-conservative replication, each of the new daughter molecules carry one of the original strands. DNA can be open or closed like a circular ring (or a ribbon). For example in many virus or bacteriophages, DNA is in an open state but after infection it closes to form a circular DNA. A circular DNA (with $N \to \infty$) will also undergo similar melting transition into two separate circular DNA as in Fig. 9.12b. It is possible that there are other events disrupting a smooth joining of the ends. Suppose there is a 2π twist of the ribbon before joining to form a twisted circular DNA (not a Möbius strip). This long DNA will also undergo a melting transition to give two nonbonded single stranded circles, but the two circles are linked topologically. From a thermodynamics ot statistical mechanical point of view, the slight change in the "boundary conditions" in the three cases shown in Fig. 9.12 do not matter but from a biological point of view, case (c) is dead or inactive because two the strands cannot be shared by the daughter virus or phages. We recognize the importance of topological constraints in biology though it does not affect the thermodynamical quantities like energy or entropy much.

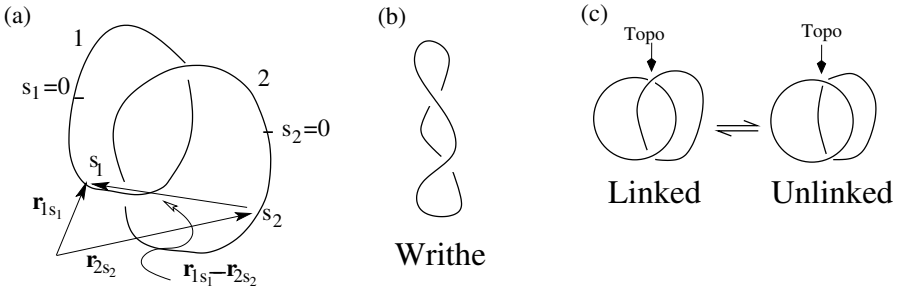


Figure 9.13: (a) Two loops with contour variables k_1 and k_2, and position vectors used in Eq. (9.45). (b)The coiling of the axis of the loop is called writhe. In DNA this leads to supercoiling. (c) Two linked DNA with TopoIV (Topoisomerase IV is a type of Topoisomerase II). TopoIV cuts both the starnds of a DNA, allows the other double stranded DNA to pass through the cut, and then rejoins the cut strands. Left to right: TopoIV unlinks the two DNA's; right to left: two unlinked DNA's get topologically linked.

9.8.1 Linking number

That cases (b) and (c) of Fig. 9.12 are different can be seen by studying a topological invariant called Gauss linking number. For simplicity we assume that each of the strand is free to cross itself ("phantom" chain) but they are

mutually avoiding. In other words a chain cannot cross the other. This prevents the two circles becoming disentangled in case (c). If $\mathbf{r}_{is_i}$ denotes the position vector of the point of chain $i(i = 1, 2)$ at position s_i measured along the chain from an arbitrarily chosen point $s_i = 0$, then the Gauss linking number is given by [8]

$$Lk = \frac{1}{4\pi} \oint ds_1 \oint ds_2 \left[\left(\frac{\partial \mathbf{r}_{1s_1}}{\partial s_1} \right) \times \left(\frac{\partial \mathbf{r}_{2s_2}}{\partial s_2} \right) \right] \cdot \frac{\mathbf{r}_{1s_1} - \mathbf{r}_{2s_2}}{|\mathbf{r}_{1s_1} - \mathbf{r}_{2s_2}|^3}, \tag{9.45}$$

where the integrals are from $s_i = 0, N$, assuming both the chains to be of same length. Lk is zero for cases (a) and (b), but nonzero for (c) for which it is 1. Since each chain can cross itself, one of the two, say 1, can be flattened into a planar circle. The linking number then counts the number of times chain 2 pierces loop 1. If we put a direction on each of the two strands in the direction of increasing s_1, s_2 as if there are currents in the loops, then the counting rule can be proved by using Ampere's law in magnetism (see problem). Because of the directions on the loops, Lk can be positive or negative.

The two integrals over the closed loops of the contour variables s_1, s_2, can be rewritten as a two dimensional integral over a torus, $\mathbf{k} = (s_1, s_2)$, as $S^1 \times S^1 = \mathbb{T}^2$. If we define a unit vector

$$\hat{n}(\mathbf{k}) = \frac{\mathbf{r}_{1s_1} - \mathbf{r}_{2s_2}}{|\mathbf{r}_{1s_1} - \mathbf{r}_{2s_2}|},$$

then the integral in Eq. (9.45) can be written as

$$Lk = \frac{1}{4\pi} \iint_{\mathbb{T}^2} d^2k \;\; \hat{n} \cdot \left(\frac{\partial \hat{n}}{\partial s_1} \times \frac{\partial \hat{n}}{\partial s_2} \right), \tag{9.46}$$

which is identical in form as the Chern number integral for insulators, Eq. (9.41). Since $\hat{n}$ maps out a sphere, the linking number actually counts how many times $\hat{n}$ winds around the sphere as one covers the whole torus. This is the same number one looks for Chern insulators, but now the context is vastly different.

9.8.2 Twist and Writhe

As mentioned already, the linking number remains invariant under melting, which does not allow chain breaking. For the unbound phase, this is the only relevant topological invariant of interest, but for the bound phase there can be other geometrical quantities. Refer to Fig. 9.10. Consider a ribbon like object by making the cross-section very thin. The unit vector along the axis, $\mathbf{t}(s)$ represents the local DNA orientation in space, while $\mathbf{a}(s)$ is the hydrogen bond vector at contour position s measured along the axis of the ribbon. With the third direction, we have a triad, a local coordinate system. The ribbon or the bar is closed so that at the end of the loop, $\mathbf{t}$ comes back to its original position $(\mathbf{t}(0) = \mathbf{t}(L)$, thereby forming a closed loop on S^2 as in Fig. 9.10c. The solid

angle formed by the loop is the change in the orientation of **a**. The twist can be defined by

$$Tw = \frac{1}{2\pi} \oint ds \;\; \mathbf{a}(s) \cdot \left(\frac{\partial \mathbf{a}}{\partial s} \times \frac{\partial \mathbf{t}}{\partial s} \right), \tag{9.47}$$

which need not be an integer. There is another geometric quantity that determines the twisting of the axis as in Fig. 9.13b. It is given by a formula similar to the linking number formula except that both the integrations are over the same loop formed by the axis,

$$Wr = \frac{1}{4\pi} \oint ds \oint ds' \left[\left(\frac{\partial \mathbf{r}_s}{\partial s} \right) \times \left(\frac{\partial \mathbf{r}_{s'}}{\partial s'} \right) \right] \cdot \frac{\mathbf{r}_s - \mathbf{r}_s}{|\mathbf{r}_s - \mathbf{r}_{s'}|^3}. \tag{9.48}$$

Note that $s = s'$ does not pose a problem because the crossproduct is also zero. Wr is a continuous variable and can be changed by deforming the loop. It depends only on the shape but not on the scale. The sign of Wr tells us the overall handedness of the coil – right or left handed. An important theorem by Călugăreanu connects the topological invariant Lk to the two geometrical quantities as

$$Lk = Tw + Wr. \tag{9.49}$$

If we now consider the ensemble of all possible closed configurations of the double stranded DNA, with the relaxed state (energetically minimum state) with zero twist and zero writhe, we have $\langle Lk \rangle = \langle Tw \rangle = \langle Wr \rangle = 0$. Intuitively, twist and writhe are independent. Therefore, the variances are additive, $\langle Lk^2 \rangle = \langle Tw^2 \rangle + \langle Wr^2 \rangle$. For a real DNA of length L with a helical pitch of γ, the normal relaxed state will have $Tw_0 = L/\gamma$ In that case the averages satisfy $\langle Tw - Tw_0 \rangle = \langle Wr \rangle = 0$, and $\langle (\Delta Lk)^2 \rangle = \langle (\Delta Tw)^2 \rangle + \langle (\Delta Wr)^2 \rangle$ where Δ denotes deviation from the average. The variances of Tw, Wr can be related to the elastic constants of the DNA, allowing us to link topological or geometrical features to the elastic constants. Although, Lk remains constant at melting but Tw, Wr lose their meaning in the unbound phase, or even in partially unzipped state.

9.8.3 Problem of Topoisomerase

Biological processes require trivial L. Two closed loops with different L cannot be deformed into one another and therefore belong to topologically different classes. Note that this is true only in three dimensions (**r**'s are 3-dimensional vectors). An example is shown in Appendix B on opening up this link in four dimensions without any cut-paste.

There are enzymes called Topoisomerase II (topoisomerase IV to be precise) that can cut a double stranded DNA at a crossing, change the value of Lk as shown in Fig. 9.13c, and rejoin the cut DNA. This is needed in biology to separate two circular DNAs after replication. What is surprising is that the topoisomerase can locally do a cut-paste to make the change. In the figure only minimal number of crossings are shown. There could be many trivial crossings.

As the topological feature is a global one (the integrations over the two loops are equivalent to scanning the whole chains), how an object, acting locally, can achieve this is a big puzzle. It is easy to see that, in thermal equilibrium, any crossing changed at random can produce a link as often as it may open it up. One wonders if Topoisomerase knows of the fourth dimension!

Problem 9.8.1: Show the equality of the two expressions in Eqs. (9.46) and (9.45).

Problem 9.8.2: Use the standard formula of magnetic field $d\mathbf{B}(\mathbf{r})$ at $\mathbf{r}$ due to a small current element $d\mathbf{l}$ at $\mathbf{r}'(l)$, $d\mathbf{B}(\mathbf{r}) \propto d\mathbf{l} \times \nabla \frac{1}{|\mathbf{r}-\mathbf{r}'(l)|}$ and Stokes' theorem to prove that Lk in Eq. (9.45) is an integer..

Problem 9.8.3: Show that Tw is additive, i.e., it can be computed by adding twists for pieces. Show that writhe is zero for a planar figure.

Problem 9.8.4: Prove Călugăreanu's theorem.

Problem 9.8.5: Show that for a double stranded DNA loop, Topoisomerase IV changes writhe by ± 2.

9.9 Summary

In this chapter we explored several simple problems from classical mechanics, statistical mechanics, and quantum mechanics by using topological arguments. These, in turn, allowed us to explain some of the basic ideas of elementary topology in terms of the known physical phenomena, and the common link among diverse topics.

Appendix A: Möbius strip and Stokes' theorem

Let us now discuss a different type of problems involving integrals of vector fields. Stokes' theorem states that the surface integral of the curl of a vector field is equal to the line integral of the field over the boundary of the surface.

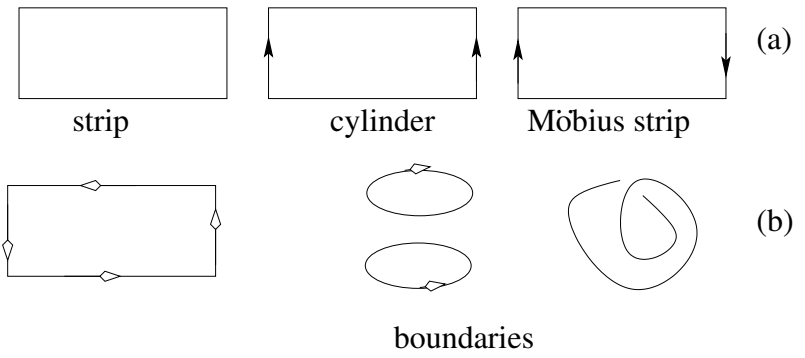


Figure 9.14: Cylinder, Möbius strip and Stokes' law.

Take a rectangular strip (say, a long piece of paper) of length $2\pi R$ and width $2L$ as in Fig. 9.14a. Now use the arrows to define "equivalence" condition, namely periodic boundary condition in the x-direction to get a cylinder and with a twisted boundary condition ("half-twist") to get a Möbius strip. These two in 3 dimensions can be described by

$$\text{cylinder}: \quad x(t,\theta) = R\cos\theta, y(t,\theta) = R\sin\theta, z(t,\theta) = t, \tag{9.50}$$

$$\text{Möbius}: \quad x(t,\theta) = \left(R - t\sin\frac{\theta}{2}\right)\cos\theta, y(t,\theta) = \left(R - t\sin\frac{\theta}{2}\right)\sin\theta,$$
$$z(t,\theta) = t\cos\theta, \tag{9.51}$$

for $-L \le t \le L$ and $0 \le \theta < 2\pi$.

The cylinder has two boundaries at $t = -L$ and $t = L$, but the Möbius strip has only one boundary. For the Möbius strip, we reach the same point by going around twice, so that the boundary can be described by Eq. (9.51) with $t = L$ and $0 \le \theta \le 4\pi$. It actually consists of the two original boundaries at $t = L, 0 \le \theta \le 2\pi$ and $t = -L, 0 \le \theta \le 2\pi$ joined together to form an unknot, as can be seen in Fig. 9.14b.

Suppose we have a vector field

$$\mathbf{A} = \hat{k} \times \frac{\vec{\rho}}{\rho^2},$$

where $\vec{\rho} = x\hat{\imath} + y\hat{\jmath}$ in the same three dimensional coordinate system. For this field $\nabla \times \mathbf{A} = 0$ everywhere except for the z-axis.

We now consider the three geometries of Fig. 9.14a separately.

1. Let us now put the strip in Fig. 9.14 say parallel to the z axis with the centre at $(R, 0, 0)$. The surface integral $\int\int d\vec{S} \cdot (\nabla \times \mathbf{A}) = 0$. It is easy to check that Stokes' theorem is valid by showing that the line integral $\oint \mathbf{A} \cdot d\mathbf{l} = 0$, where the integration is along the boundary, as shown in the left figure of Fig. 9.14b.

2. For the cylinder placed along the z-axis with the centre of the cylinder at the origin, there are two boundaries along which one has to do the line integral. Since the directions are opposite for the two rings at $z = \pm L$ (middle fig of Fig. 9.14b), the total line integral is zero. In other words Stokes' theorem is explicitly verified.

3. For the Möbius case, the line integral along the boundary curve is

$$\oint_{\text{boundary}} \mathbf{A} \cdot d\mathbf{r} = \int_0^{4\pi} \mathbf{A}(x(\theta), y(\theta), z(\theta)) \cdot \mathbf{r}' \, d\theta \neq 0,$$

where $\mathbf{r}' = d\mathbf{r}/d\theta$. For the parametrization used, $\mathbf{A} \cdot \mathbf{r}' = 1$ so that the integral is 4π. *There seems to be a violation of Stokes' law.*

This paradox is resolved by noting that the area vector $d\mathbf{S}$ cannot be defined on the Möbius strip. If we slide a small area element along the strip through 2π the area vector will not point in the same direction. Another way of seeing the difference is to colour the surfaces without any abrupt change. Two colours are needed to paint the surfaces of the strp and the cylinder, but one is enough for the Möbius strip. Such a surface, like the Möbius strip, is called a nonorientable surface and Stokes' theorem is not applicable there.[21]

Problem 9.9.1: Is it possible to generalize the Möbius strip construction so that the boundary curve is, say, a trefoil knot? (Hint: three half-twists)

Appendix B: Disentanglement via moves in 4-dimensions

Take two loops with linking number one. It is a common knowledge that the two loops cannot be taken apart if the chains do not cross each other. We now show a set of local moves in four dimensions (x-y-z-w space) that takes (b) to (c) in Fig. 9.15.[22]

Let t be a parameter, $t \in [0, 1]$, so that chain 1 can be parameterized in 4-dimensions by

$$f_0(t) = \begin{cases} (-a, -a, -1, 0) & (0 \leq t \leq \frac{3}{8}) \\ (-a + 8a(t - \frac{1}{8}), -a + 8a(t - \frac{1}{8}), -1, 0) & (\frac{3}{8} \leq t \leq \frac{5}{8}) \\ (a, a, -1, 0) & (\frac{5}{8} \leq t \leq 1) \end{cases} \quad (9.52)$$

[21] See Chapter 5 for a discussion on Stokes' theorem.

[22] This is based on the online mathjournal article by Jeff Boersema and Erica J. Taylor, https://www.rose-hulman.edu/mathjournal/archives/2003/vol4-n2/paper2/v4n2-2pd.pdf

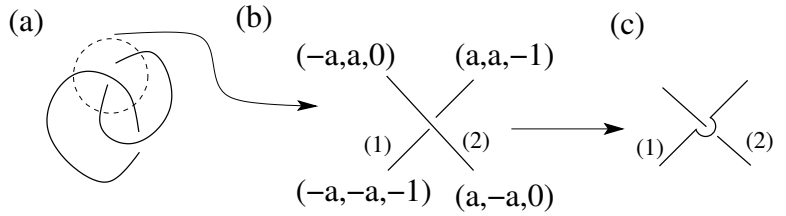


Figure 9.15: (a) Linked loops. The crossing indicted by the dotted circle is shown in (b). In three dimensions, chain 1 in (b) is in the plane $z = -1$ while the other one is at $z = 0$ so that chain 1 is below chain 2. The crossing point is taken to be at $x = y = 0$ when projected in the xy plane. If the chain configurations can be changed to type (c), then the two loops can be unlinked. This is possible in 4-dimensions but not in three.

We shall deform the chain in three mutually exclusive ways, (i) in the xy plane, or, (ii) in the z-direction, or, (iii) in the 4th w direction. The variable t is used to describe the configuration of the chain, while a second variable $s \in [0, 1]$ is to be used for the deformation so that the chain is described as $f_s(t)$.

First we deform in the x-y-w space keeping $z = -1$ fixed. The new chain at $s = 1/3$ is given by

$$f_{1/3}(t) = \begin{cases} (-a, -a, -1, 8t), & (0 \le t \le \frac{1}{8}) \\ (-a, -a, -1, 1), & (\frac{1}{8} \le t \le \frac{3}{8}) \\ (-a + 8a(t - \frac{3}{8}), -a + 8a(t - \frac{3}{8}), -1, 1), & (\frac{3}{8} \le t \le \frac{5}{8}) \\ (a, a, -1, 1), & (\frac{5}{8} \le t \le \frac{7}{8}) \\ (a, a, -1, 1 - 8(t - \frac{7}{8})), & (\frac{7}{8} \le t \le 1) \end{cases} \quad (9.53)$$

Next, we deform the z coordinate from -1 to 1 taking the chain above the other one as

$$f_{2/3}(t) = \begin{cases} (-a, -a, -1, 8t), & (0 \le t \le \frac{1}{8}) \\ (-a, -a, -1 + 16(t - \frac{1}{8}), 1), & (\frac{1}{8} \le t \le \frac{2}{8}) \\ (-a, -a, 1, 1), & (\frac{2}{8} \le t \le \frac{3}{8}) \\ (-a + 8a(t - \frac{3}{8}), -a + 8a(t - \frac{3}{8}), 1, 1), & (\frac{3}{8} \le t \le \frac{5}{8}) \\ (a, a, 1, 1), & (\frac{5}{8} \le t \le \frac{6}{8}) \\ (a, a, 1 - 16(t - \frac{6}{8}), 1), & (\frac{5}{8} \le t \le \frac{7}{8}) \\ (a, a, -1, 1 - 8(t - \frac{7}{8})), & (\frac{7}{8} \le t \le 1) \end{cases} \quad (9.54)$$

Finally, we bring back the fourth w coordinate to zero as,

$$f_1(t) = \begin{cases} (-a,-a,-1,0), & (0 \le t \le \frac{1}{8}) \\ (-a,-a,-1+16(t-\frac{1}{8}),0), & (\frac{1}{8} \le t \le \frac{2}{8}) \\ (-a,-a,1,0), & (\frac{2}{8} \le t \le \frac{3}{8}) \\ (-a+8a(t-\frac{3}{8}),-a+8a(t-\frac{3}{8}),1,0), & (\frac{3}{8} \le t \le \frac{5}{8}) \\ (a,a,1,0), & (\frac{5}{8} \le t \le \frac{6}{8}) \\ (a,a,1-16(t-\frac{6}{8}),0), & (\frac{5}{8} \le t \le \frac{7}{8}) \\ (a,a,-1,0), & (\frac{7}{8} \le t \le 1) \end{cases} \tag{9.55}$$

It is now straight forward to construct a map $f_s(t)$ linear and continuous in s – it is already linear and continuous in t – that goes from $f_0(t) \to f_{1/3}(t) \to f_{2/3}(t) \to f_1(t)$. For example, for $2/3 \le s \le 1$ a continuous map from $f_{2/3}(t)$ to $f_1(t)$ can be constructed by replacing the w values of $f_{2/3}$ by $3w(1-s)$. Note that we get back the same boundary points $(-a,-a,-1,0)$ at $t=0$ and $(a,a,-1,0)$ at $t=1$ for all $s \in [2/3,1]$.

The two loops can therefore be delinked via these local deformations in the 4th dimension.

References

[1] M. Nakahara, *Geometry, Topology and Physics* (Taylor & Francis, Boca Raton, FL, USA, 2003).

[2] R. Aldrovandi and J. G. Pereira, *An Introduction to Geometrical Physics* (World Scientific, SIngapore, 1995).

[3] C. Nash, S. Sen, *Topology and Geometry for Physicists* (Dover, Mineola, NY, 2011).

[4] See, e.g., the chapter by Somnath Basu and Atreyee Bhattacharya in this book.

[5] See, e.g., the chapter by Samik Basu and Soma Maity in this book.

[6] See, e.g., the chapter by Dheeraj Kulkarni in this book.

[7] *Geometric Phases in Physics*, Edited by: F Wilczek and A Shapere (World Scientific, Singapore, 1989)

[8] See, e.g., the chapters by Mahan Mj, and by P. Ramadevi in this book.

[9] For an example of an insulator with Chern number=2, see P. Titum, N. H. Lindner, M. C. Rechtsman, and G. Refael, Phys. Rev. Lett. **114**, 056801 (2015); P. Titum, N. Lindner, and G. Refael, Phys. Rev. B 96, 054207 (2017) arXiv:1702.02956.

[10] D. J. Thouless, *Topological Quantum Numbers in Nonrelativistic Physics*, (World Scientific, Singapore, 1997).

10

What is dimension?

Somendra M. Bhattacharjee

This chapter explores the notion of "dimension" of a set. Various power laws by which an Euclidean space can be characterized are used to define dimensions, which then explore different aspects of the set. Also discussed are the generalization to multifractals, and discrete and continuous scale invariance with the emergence of complex dimensions. The idea of renormalization group flow equations can be introduced in this framework, to show how the power laws determined by dimensional analysis (engineering dimensions) get modified by extra anomalous dimensions. As an example of the RG flow equation, the scaling of conductance by disorder in the context of localization is used. A few technicalities, including the connection between entropy and fractal dimension, can be found in the appendices.

10.1 Introduction

The purpose of this chapter is to explore the idea of "dimension" of a set. E.g., what does superscript 1 mean when we talk of S^1? The vector space idea of the minimum number of basis vectors is too restrictive to be applicable to even many subsets of standard Euclidean manifolds, which are not necessarily vector spaces. Generalizations of the definition of dimension consistent with the intuitions based on Euclidean spaces give a large number of possibilities, with each one describing some aspect of the set. These dimensions need not be integers anymore. If various physical phenomena are studied on these subsets, which "dimension" will matter?

Once released from the integer constraint, there is no restriction on the value of dimension, with examples of positive, negative, real, and complex, d,

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_10

even cases which require more than one d. Such varieties are not possible in ordinary Euclidean spaces $\mathbb{R}^d$ because all definitions give the same d for them.

It is amusing to note that the word "dimension" means several different things in physics. It is used in statements like "a line is one dimensional", and that is the "dimension" we are interested in this chapter.[1] The other one is the dimension in the context of "Units and Dimensions", as for example, force has dimension MLT^{-2}, where $\mathsf{M}, \mathsf{L}, \mathsf{T}$ are the dimensions of mass, length, and time respectively. With the available scales of a problem, like the interaction strength, thermal energy (k_BT), $\hbar$ (action), c (velocity), we may, if we wish, express all physical quantities in terms of length only.[2] The power of L we get by the dimensional analysis is to be called the *engineering dimension* of the quantity. One of the aims of this chapter is to bring these two usages of "dimension" in the same framework. In the process we shall argue that the framework allows ways to apparently "violate" dimensional analysis, and how extra "anomalous dimensions" emerge.

The definitions of dimensions are based on various *power laws* by which an Euclidean space can be characterized. This procedure opens up a new way of studying power laws, beyond geometrical objects, like the divergences of response functions, e.g., susceptibility, near critical points. The idea of renormalization group can be introduced in this framework, to show, as just mentioned, how the power laws determined by engineering dimensions get modified by extra anomalous dimensions.

Mostly well-known examples are considered in this chapter. For generalities and more mathematical issues, see Refs. [1–3].

10.2 Does "dimension" matter?

Many phenomena when viewed in a broad way, are found to depend on the dimensionality of the system. Let's take a few examples where d, the dimension of the space, occurs explicitly.

1. For noninteracting gases, classical or bosons or fermions, in d-dimensions, the thermodynamic fundamental relation is $PV = \frac{2}{d}\, U$ for pressure P, volume V, and total energy U.

2. Debye specific heat $c \sim T^d$ for a d-dimensional crystal at low temperatures T.

[1] Common usages like "love adds a new dimension to the conflict" or "the dimensions of this box are 10cm×20cm×30cm" can possibly be traced to the meaning of "dimension" as used in this chapter!

[2] As an example consider the path integral form of the propagator of a free particle in quantum mechanics , which involves a sum over all trajectories of $\exp(iS/\hbar)$, where $\hbar$ is the Planck constant divided by 2π, $S = (m/2)\int dt\ (d\mathbf{x}/dt)^2$ is the action, m and $\mathbf{x}$ are the mass and the position of the particle, t being time. Since $S/\hbar$ is necessarily dimensionless, we may define $\tau = \hbar t/m$ to write $S/\hbar = (1/2)\int d\tau\ (d\mathbf{x}/d\tau)^2$. With this form, x has dimension L while redefined time τ has dimension L^2.

3. The probability P_t of a random walker coming back to the starting point in time t satisfies $P_t \propto t^{-d/2}$.

4. A diffusing particle has the characteristic mean square displacement $R^2 \sim t$ in time t, in all *dimensions* so that the volume occupied is $t^{d/2}$.

5. In quantum mechanics, a particle in a short range attractive potential may not have a bound state if $d > 2$, but there is always a bound state if $d < 2$.

6. Take an Ising type model with short-range interactions. It is known that there is a phase transition (critical point) if $d > 1$. For vector spins, a critical point exists only if $d > 2$. The critical behaviour is mean-field like if $d > 4$. Inbetween, the critical exponents depend on d and a few other gross features.

These are just a few. What do we mean by d in these statements?

10.3 Euclidean and topological dimensions

That a square lattice is two dimensional is easy to see if its vector space property is known. A cube, consisting of vertices and edges, can be drawn on a piece of paper. Fig 10.1 shows possible constructions of hypercubes and hyperspheres S^n (drawn in $d = 2$). [3] That the cube is in some sense not a two dimensional object becomes clear if one wants to draw on a plane larger lattices or graphs with such cubes as units.

10.3.1 Euclidean dimension

A common procedure is to embed the lattice in an Euclidean space of large enough dimensions. From any point, draw a sphere of radius R and count the number N of points enclosed by the sphere. Our expectation is that, for a d-dimensional set, [4] $N \sim R^d$. Exploiting this intuition, a definition of d can be

$$d = \lim_{R\to\infty} \frac{\ln N}{\ln R}, \qquad \text{(Euclidean)} \tag{10.1}$$

the asymptotic slope of a log-log plot of N vs R. The above definition may be written in a more useful form as

$$R\,\frac{\partial N(R)}{\partial R} = d\,N(R), \tag{10.2}$$

[3] Convention: S^n denotes the surface of an $(n+1)$-dimensional sphere while B^n denotes an n-dimensional ball, i.e., a sphere with its interior and its boundary surface. For example S^1 is the set of points in two dimensions, $x^2 + y^2 = 1$, while B^2 is the set of points with $x^2 + y^2 \leq 1$. This is equivalent to saying S^1 is the surface or boundary of B^2.

[4] A symbol $\sim$ indicates the functional relation without worrying about dimensional analysis, prefactors etc, while $\approx$ will be reserved for approximate equality.

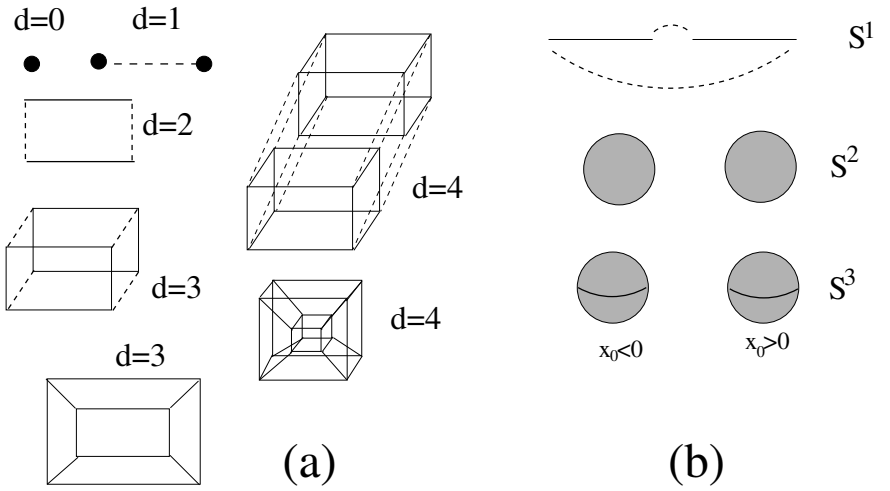


Figure 10.1: (a) Construction of a d-dimensional cube (hypercube) by combining pairs of $(d-1)$-dimensional cubes. The dotted lines are the new connections or bonds. (b) Hyperspheres may also be constructed that way. Take two closed intervals with the identification of the boundary points as indicated by the dashed lines to get S^1. Similarly, two full spheres B^d, unit d-balls, one for $x_0 < 0$ and another one for $x_0 > 0$ with the identification of the boundaries gives S^d. E.g., take two disks ($d = 2$) and place one top of the other both centered at origin. Since the boundaries are identified, puff the object (think of making *luchi* [https://en.wikipedia.org/wiki/Luchi]) giving an extra dimension as $x_0 = \pm\sqrt{1 - \sum_{i=1}^{d} x_i^2}$, the boundary being at $x_0 = 0$. The net result is S^d, $\sum_{i=0}^{d} x_i^2 = 1$. This construction is not possible for solid cubes or solid balls.

where the coefficient of the linear term on the right hand side corresponds to the dimension of the space.

An equation like Eq. (10.2) is suggestive of the form used in renormalization group approach and can be linked to dimensional analysis for various physical quantities. If a physical quantity A, on dimensional grounds, depends on length as L^c, then there is an equation equivalent to Eq. (10.2), viz.,

$$L\,\frac{\partial A}{\partial L} = c\,A, \tag{10.3}$$

where c will be called the engineering dimension of A.

d via analytic continuation

In many problems, especially in renormalization group calculations (ϵ-expansion), one generalizes the Euclidean dimension, Eq. (10.1), to a continuous

variable. This is more of an analytic continuation with the help of the metric than by any real construction of any space. A Gaussian integral in d dimensions can be written as

$$\int e^{-ar^2} d^d r = K_d \int_0^\infty e^{-ar^2} r^{d-1} dr, \quad \text{with} \quad K_d = \frac{2\pi^{d/2}}{\Gamma(d/2)}, \tag{10.4}$$

as the surface area of a unit sphere. An integration like the left hand side of Eq. (10.4) above is metric dependent. Once converted to a one-dimensional integral of a function where d appears as a parameter (like the right hand side of Eq.(10.4)), it is defined for any value of d allowing an analytic continuation to the whole complex plane of d. This is very useful to handle singularities or divergent integrals (dimensional regularization) in many problems. In this analytic continuation approach, there is no association of any space with non-integer d and so we do not get into detailed discussion on this approach in this chapter. (See Prob 3.1).

10.3.2 Topological dimension

It is possible to avoid any reference to an embedding space by using the intrinsic characteristics of the lattice. By lattice we mean a set of points connected by bonds and these bonds can be taken as a unit or a scale for the connectivity of the points. We take an L step path on the lattice from any one point and count all the new points visited that were not seen upto the $(L-1)$th step. This is like counting the boundary points of an intrinsically defined sphere of radius L. Based on the expectation that the boundary "area" grows like $N_b \sim L^{d-1}$, the definition of the dimension is

$$d_{\mathrm{t}} = 1 + \lim_{L\to\infty} \frac{\ln N_b}{\ln L}. \qquad \text{(topological)} \tag{10.5}$$

This will be called the *topological dimension* of the object. It is topological because this number does not change under continuous deformation of the space. In other words two homeomorphic spaces have the same topological dimension.

A formal definition of the topological dimension is via the covers. A crude definition is that if δ is the dimension of the space (or lowest dimension of all possible spaces) that separates our space into disconnected pieces, then the topological dimension of our space is $1+\delta$.

We make a convention that a null set $\emptyset$ has dimension -1 while a point set has dimension 0. All others follow from the above rule.

There are ambiguities. As an example, take a line. A line is broken into two pieces by removing a point. A point by definition is of zero dimension. Hence a line is a one-dimensional object. This is however a bit tricky. If we think of a line in three dimensions, it can be broken into two pieces by another line or by a plane etc. In such situations, we need to choose the smallest dimensional object to determine δ. Another example could be a figure $\overline{\bullet}$ (a line and a disk). Being

disconnected, a null set separates them, and, therefore, the dimension should be $1 + (-1) = 0$. But individually these are 1 (line) and 2 dimensional (disk) spaces. In such a situation we define a local dimension and choose the largest one. In this particular case it will be $d = 2$.

The above rules may be formalized by an iterative procedure with the basis sets of the space. (a) If the boundaries of the basis sets are of dimensions $\leq d-1$, then the space is of dimension $\leq d$. (b) If this is true for d but not for $d-1$, then the space has dimension d.

For the disk-bar example above, the basis sets for the bar has boundaries $d = 0$ while the disk has basis sets with boundaries $d = 1$, i.e., the dimension of the boundaries of the basis sets satisfy $d \leq 1$. Therefore $d \leq 2$. Invoke (b) to rule out $d = 1$ or any number greater than 2.

Let us take R with the usual topology defined by the open sets $(a, b), b > a$. The boundaries are 0-dimensional points for all such open sets. Hence R has dimension $d = 1$. With inherited topology for S^1, the basis sets (open arcs of a circle) have boundaries of dimensionality 0. Therefore S^1 has $d = 1$.

All the definitions used so far would give the dimension of R^d to be d. This number d happens to be the number of independent vectors needed to span the space when viewed as a vector space. The dimensionality of a topological space is a topological invariant in the sense that if there is a continuous mapping or homeomorphism that takes R^m to R^n, then $m = n$.

Problem 10.3.1: A problem on dimensional regularization. Show that the one-dimensional integral $I_1(x) = \int_{-\infty}^{\infty} \frac{dz}{\sqrt{x^2+z^2}}$ is divergent.

To tackle this divergence, generalize the integral to d dimensions as

$$I_d(x) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \frac{d^d r}{\mu^{d-1}\sqrt{x^2+r^2}} = K_d \int_0^{\infty} \frac{r^{d-1} dr}{\mu^{d-1}\sqrt{x^2+r^2}}, \tag{10.6}$$

where an arbitrary length μ is introduced to maintain the correct dimensions (engineering dimension!). Formally, I_d is I_1 for $d = 1$. The form on the right hand side can be defined for any d. With d as a continuous variable, the integral is divergent[5] for $d \geq 1$ but convergent for $d < 1$. This signals the possibility of a singularity in the complex d-plane at $d = 1$. By doing the integral in the convergent domain in the d-plane, show, by using Gamma functions and analytic continuation, that

$$I_d(x) = \left(\sqrt{\pi}\frac{x}{\mu}\right)^{d-1} \Gamma\left(\frac{1-d}{2}\right) = \frac{2}{\epsilon} - 2\ \ln(x/\tilde{\mu}) + O(\epsilon), \tag{10.7}$$

where $\epsilon = 1 - d$ is a small parameter for the expansion. Absorb some O(1) factor in μ to define $\tilde{\mu}$.

This particular example appears in the calculation of the electrostatic potential due to an infinitely long uniformly charged wire [4].

[5]Important here is the behaviour at the upper limit, which we may see by putting a cutoff L as $I_d^{(L)} \sim \int^L r^{d-2} dr \sim L^{d-1}/(d-1)$, for $d \neq 1$. For $d = 1, I_{d=1}^{(L)} \sim \ln L$. Therefore $I_{d\geq 1} \to \infty$ as $L \to \infty$. In such analytic continuations of integrals, log divergences are always very special.

Problem 10.3.2: (Mathematical) Show that the definition of d obtained by using a basis is independent of the basis chosen.

One way of doing it is to define d without any basis but with the help of the open sets. The iterative definition of the dimension of a topological space T for a set X would be as follows: The dimension is d if for any point $x \in X$ and any set $U \in T$ containing x, there exists an open V with $x \in V$ such that the closure of V is contained in U with $\dim(\partial V) \leq d-1$. If this is satisfied for d but not for $d-1$, then the dimension of the space is d. It is assumed that $\dim(T)$=-1 if $X = \emptyset$.

Problem 10.3.3: (Mathematical) Prove that if R^m is topologically equivalent (i.e., homeomorphic) to R^n, then $m = n$. This means the dimension is a topological invariant.

10.4 Fractal dimension: Hausdorff, Minkowski (box) dimensions

Let us now look at some nontrivial examples. We shall see that topological dimension (d_{t}) is not enough; a few others are needed. First among these is the Hausdorff dimension (d_{f}) that one gets by embedding the set in a real space R^n of appropriate dimension n.

In the following the Cantor set (defined below) is taken as a paradigmatic example because of its apparent simplicity. It is an example of a space of topological dimension 0 but it is not just a finite collection of points.

10.4.1 Cantor set: $d_{\mathrm{t}} = 0, d_{\mathrm{f}} < 1$

There are many ways to define Cantor sets. The middle 1/3 rule used below is historically the first one defined by Cantor and will be called *the* Cantor set.

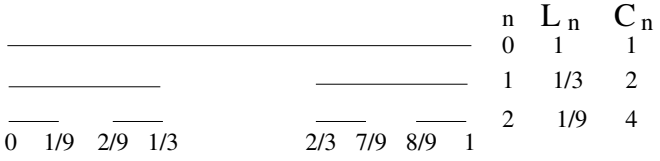


Figure 10.2: Construction of a cantor set

The set is constructed iteratively by taking a closed interval 0 to 1, and then removing the middle 1/3 to get two *closed* intervals. In the next step, the middle one third of the two branches are removed leaving us with 4 intervals (Fig. 10.2). This iterative process leaves a set of disconnected points,

$$\left.\begin{array}{rcl} S_0 & = & [0,1]\,, \\ S_1 & = & \left[0,\frac{1}{3}\right] \cup \left[\frac{2}{3},1\right], \\ S_2 & = & \left[0,\frac{1}{9}\right] \cup \left[\frac{2}{9},\frac{3}{9}\right] \cup \left[\frac{6}{9},\frac{7}{9}\right] \cup \left[\frac{8}{9},1\right], \\ \ldots & & \end{array}\right\} \quad \mathcal{S} = \cap_{n=0}^{\infty} S_n. \tag{10.8}$$

First note that the lengths of intervals removed are successively $\frac{1}{3}, \frac{2}{3^2}, \ldots$ so that the total length removed is $\sum_n \frac{2^{n-1}}{3^n} = 1$. Since the total length we started with is 1, the remaining points have no "length" and therefore cannot be one-dimensional. The set is disconnected. Therefore the *topological dimension* is $d_{\rm t} = 0$. Despite this 0-dimensionality, there is still a nontrivial structure like self-similarity. If any one part of the nth iterate is multiplied by 3, we get back the state in the $(n-1)$th iterate. This geometric structure is described by a different dimension, to be called the self-similarity dimension or box dimension or Hausdorff dimension.

For a regular object like the Cantor set, the pattern is obtained by scaling the nth iterate by a scale $b < 1$ and combining C_b of them. Here, $b = 1/3$ and $C_b = 2$. On successive rescaling how the number grows is given by the Hausdorff dimension

$$d_{\rm H} = \frac{\ln(C_b)}{\ln(1/b)}. \qquad \text{(Hausdorff)} \tag{10.9}$$

For the Cantor set, $d_{\rm H} = \frac{\ln 2}{\ln 3} \approx 0.63 < 1$. We have assumed a power law dependence, $C_b \sim b^{-d_{\rm H}}$.

Another practical procedure is to cover the set by boxes. Consider the set as a part of R^d. Cover it by boxes of length L and count the number of boxes occupied by the set. Let this be B_L. If the length is changed to bL by a scale factor $b < 1$, the number changes to B_{bL}. The *box dimension* is then defined as

$$d_{\rm f} = \lim_{n\to\infty} \frac{\ln(B_{b^n L}/B_L)}{\ln(1/b^n)}. \qquad \text{(Box or Minkowski)} \tag{10.10}$$

By using successive generations, we may also write the above equation as

$$d_{\rm f} = \lim_{n\to\infty} \frac{\ln(B_{b^n L}) - \ln(B_{b^{n-1} L})}{\ln(1/b^n) - \ln(1/b^{n-1})}, \qquad \text{(Box or Minkowski)} \tag{10.11}$$

which is the discrete version of Eq. (10.2).

For the Cantor set, if we choose $b = 1/3$, then $B_{b^n} = 2^n$, and, therefore, $d_{\rm f} = \frac{\ln 2}{\ln 3} = \log_3 2 = d_{\rm H}$. The box dimension is also called the *Minkowski dimension*.

The occurrence of $\ln 3$, or log base 3, is not accident. It comes from the scale 1/3 under which the Cantor set is scale invariant. If we choose an arbitrary scale factor, say $b = 1/2.9$, the scale invariance of the Cantor set is not obvious. This existence of a special scale, here 1/3 or its powers, is an example of *discrete scale invariance*. In Appendix B we discuss how a discrete scale invariance leads to complex dimensions.

Whenever the Hausdorff dimension is different from the topological dimension, the set is called a *fractal* [5]. For most regular fractals, $d_{\rm H} = d_{\rm f}$, but there are cases where they may differ. When they are same, they may be called the fractal dimension or the scaling dimension.

A subtle difference in the way the Hausdorff and the Minkowski dimensions are defined may be noted here. The Hausdorff dimension is obtained by

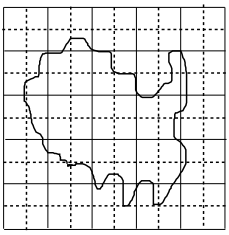


Figure 10.3: Covering by boxes. Count the number of boxes required to cover the object for different box sizes like those represented by the solid grid and the dotted grid. The dotted grid is finer and therefore would involve larger number of boxes.

going to the large size limit by scaling up the structure, while the Minkowski dimension is from the opposite limit. In the latter case, one explores to the short scale behaviour by using progressively smaller boxes. From an experimentalist's point of view, the Hausdorff dimension is obtained when probed in the long wavelength limit, called the infrared limit, while the Minkowski (box) dimension is obtained when probed using shorter wavelengths, called the ultraviolet limit.

When Hausdorff≠ Minkowski

An example of a case of different Hausdorff and Minkowski dimension is the set of rational numbers in $(0, 1)$. If we want to cover this set by linear "boxes", every box will contain some points. It follows from the fact that the rationals form a dense set in $(0, 1)$. Therefore the box dimension or Minkowski dimension is one. On the other hand, each rational number is an isolated point with dimension zero. A *countable union* of points will then have a Hausdorff dimension of zero. Generally if F is a dense subset of an open region of $\mathbb{R}^n$, then its box dimension is n. Note however that Cantor set does not belong to this category because it is *uncountable*.

Inhomogeneous scaling

Let us write the definition of $d_{\rm H}$ in Eq. (10.9) in a different way as $C\lambda^{d_{\rm f}} = 1$, where $\lambda(= 1/b) < 1$ is the scale factor and C is the number of such scaled objects combined to generate the next generation. This equation may now be extended to a situation of inhomogeneous scaling where each of the C objects has its own scale factor $\lambda_i < 1$. The fractal dimension is then the solution of the equation $\sum_{i=1}^{C} \lambda_i^{d_{\rm f}} = 1$.

Fractal dimension as a Continuous variable

Instead of the 1/3 rule of the Cantor set, we may remove any open interval $(x, 1-x), x < 1/2$. After that we remove the appropriate $(1-2x)\mathcal{L}$ part from each of the remainders of length $\mathcal{L}$. Let's call it th x-rule. For $b = x$, $C_{b^n} = 2^n$. The fractal dimension is

$$d_{\rm f}(x) = \frac{\ln 2}{\ln(1/x)}, \qquad \text{(continuous)} \tag{10.12}$$

with $d_{\rm f}(x=0)=0, d_{\rm f}(x=1/2)=1$ as expected. It is therefore possible to construct a family of fractals with fractal dimension $d_{\rm f}$ as a continuous variable in the range $[0,1]$.

Fractals of dimension $d_{\rm f}>1$ can be constructed by taking advantage of product spaces. Given a $1<d_{\rm f}<m$, where m is an integer (>1), use the x-rule to construct a fractal $\mathcal{F}$ of $d'_{\rm f}=d_{\rm f}/m<1$. Then construct the direct product space $\mathcal{F}\times ...\times\mathcal{F}$. For the scale factor b, we now need m-dimensional boxes so that the number of boxes covered is C^m_{bL}. The fractal dimension, by Eq. (10.10), is $m\ d'_{\rm f}=d_{\rm f}$.

Since m can be any integer greater than the chosen value of $d_{\rm f}$, we end up with many different product spaces all of the same fractal dimension $d_{\rm f}$. Therefore, $d_{\rm f}$ is a necessary but not a sufficient characterization of the space.

Configuration space of the Ising model

We raised the question of the configuration space of the Ising model which consists of N spins $s_i=\pm1$. For each spin we have a discrete topological space $\{0,1\}$ so that for infinitely many spins arranged in a one-dimensional lattice, the total configuration space is a product space $\{0,1\}^Z$. We now show that the configuration space can be mapped on to a set of real numbers $\in[0,1]$, whose fractal dimension can be determined.

Let us first consider a particular case using ternary expansion of numbers. By construction, any member of the Cantor set can be expressed as

$$y=\sum_{n=1}\frac{a_n}{3^n},\quad \text{where}\quad a_n=0 \text{ or } 2. \tag{10.13}$$

This is because every point in S_n exists in all the previous generations. Therefore any point can be tracked as belonging to either the 0th or the 2nd interval of the previous step.[6] Equivalently, the points can be represented by an infinite string $(a_1a_2a_3....a_n....)$, like (02220220002....) , or by dividing by 2, an infinite string of 0's and 1's. For an infinitely long chain of Ising spins, a configuration $(s_1s_2...s_n...)$ can be converted to a real number

$$x=\sum_{n=1}^{\infty}\frac{2s_n}{3^n},\quad \text{with}\quad s_n=0,1, \tag{10.14}$$

similar to Eq. (10.13). In other words, the topological space of the Cantor set can be mapped on to $\{0,1\}^Z$, or, equivalently, can be mapped onto the configuration space of an infinitely long Ising chain.

[6] Warning: there are two possible representations of some numbers like $\frac{1}{3}=\sum_n\frac{2}{3^n}$, i.e., in base 3 notation, (10000...) and (2222...) denote the same number. We omitted the decimal point in front of the numbers. Similarly $\frac{1}{4}=\frac{2}{9}+\frac{2}{9^2}+...=\frac{1}{3}-\frac{1}{3^2}+\frac{1}{3^3}-\frac{1}{3^4}+....$ In such cases we choose the representation involving 0 and 2. This restriction to 0 and 2 makes the binary string to Cantor set a one-to-one and onto mapping.

The entropy per spin of the Ising model in the high temperature limit is $k_B \ln 2$ where k_B is the Boltzmann constant What we learn from this analysis is that the entropy (in this case the high temperature entropy) of the Ising system (or, for that matter, any two state model) is determined by the dimension $\log_3 2$ of the set of real numbers equivalent to the configuration space . The connection between the dimension of the equivalent set of real numbers and the entropy is discussed in Appendix A where we also show that base 3 is nothing special.

Problem 10.4.1: The Ising two state problem can be mapped on to the Cantor set as the collection of all infinite strings of 0 and 1 occurring with equal probability. Suppose, instead, 0 occurs with probability p, and 1 with $q = 1 - p$. The physical entropy per spin is known to be $s = -k_B(p \ln p + q \ln q)$, where k_B is the Boltzmann constant. Show that the fractal dimension of the set of real numbers is $d_{\rm f} \propto S$, because $d_{\rm f} = -(p \ln p + q \ln q)/\ln 3$.

Problem 10.4.2: A generalization of the above problem is to consider the set of all strings of base m numbers, i.e., strings of $0, 1, ...m-1$. A string $\{a_n\}, a_n = 0, 1, ...m-1$, corresponds to a real number $x = \sum_1^\infty a_n/m^n$. If the digits $0, 1...m-1$ occur with probabilities $p_n, n = 0, ..., m-1$, then the fractal dimension of the et consisting of x's is $d_{\rm f} = -(\sum_n p_n \ln p_n)/\ln m$.

A consequence of this result is that if only $s_n = 0, 1$ occur with probabilities $p, q = 1 - p$, then $d_{\rm f} = -(p \ln p + q \ln q)/\ln m$,, e.g., for the set $x = \sum_n (m-1)s_n/m^n$ (a generalization of the standard Cantor set).

Problem 10.4.3: Consider the infinitely long Ising chain configurations, but now with a restriction that no two 0's can be adjacent (or nearest neighbours). This occurs in nonabelian anyon chains discussed in later chapters. (See Ref. [6]). For N spins, the total number of configurations is not 2^N any more. If C_N is the number of N spin configurations under this restriction, then show that $C_N = C_{N-1} + C_{N-2}$. Hint: If the first spin is 1, then the second onwards can be any of the allowed $N-1$ spin configurations. This is C_{N-1}. If the first one is 0, then by restriction, the next one has to be 1 but the spins are free after that. The number of such configurations is C_{N-2}. Note that $C_1 = 1, C_2 = 3$. This is the Fibonacci sequence.

If $C_N \sim \tau^N$ for large N, then show $\tau = (1 + \sqrt{5})/2$ (golden mean), as expected for the Fibonacci numbers. Show that the fractal dimension of the corresponding real number set is $\propto \ln \tau$ (see problems 4.1 and 4.2).

A number like τ here, different from the standard value 2, is often called the *quantum dimension* of the anyonic chain, . Suppose we consider spin-1/2 particles. For each spin the Hilbert space is 2 dimensional. Then the Hilbert space for N spins is the tensor product with dimensions $2 \times 2 \times 2... = 2^N$. In contrast for the (Fibonacci-) anyon chains, even though individually the spaces are two dimensional, the dimension of the N-anyon Hilbert space is τ^N for large N. This is as if the effective dimension of individual Hilbert space is τ. To recognize this difference, this dimension is called "quantum dimension". It is interesting to note that the topological entanglement entropy is determined by $\ln \tau$. This goes beyond the scope of this chapter.

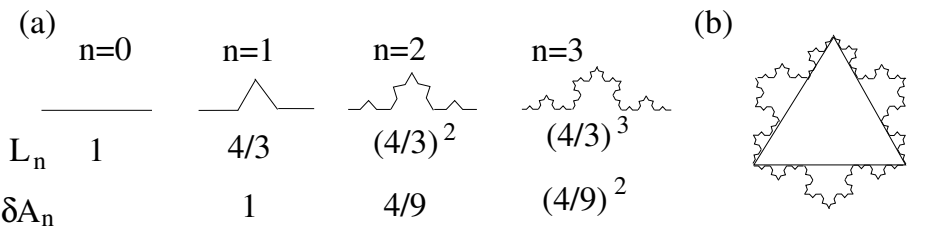


Figure 10.4: (a)Iterative construction of the Koch curve. An interval of length 1 is divided into 3 equal parts. The middle segment is replaced by an equilateral triangle of side length 1/3 without the base. This procedure is repeated for each segment. The length of the curve (L_n) and the area (A_n) under it for the first few generations are given. (b) Similar construction with a triangle. After infinite iteration we get a closed loop of infinite perimeter but of finite area. This is equivalent to S^1 where "1" refers to the topological dimension.

Problem 10.4.4: What is the configuration space of Ising spins on a square lattice? Explore if there is any mapping to real numbers, or any connection with the fractal dimension of the set of numbers, as found for a one dimensional chain of spins.

10.4.2 Koch curve: $d_{\rm t} = 1, d_{\rm f} > 1$

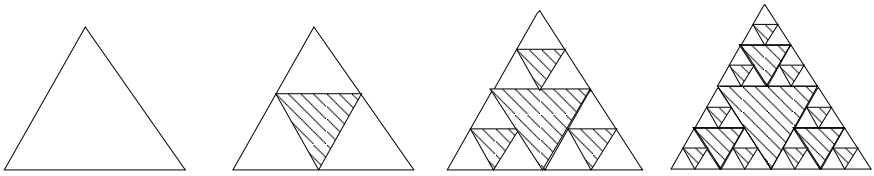
A Koch curve is defined in Fig. 10.4. Instead of deleting the middle 1/3 as in the Cantor set, we add an extra piece increasing the length of the line. This curve has the following properties:

1. A point disconnects it. Therefore it has a *topological dimension* $d_{\rm t} = 1$.
2. The generation-wise lengths are $1, \frac{4}{3}, \left(\frac{4}{3}\right)^2, \ldots \left(\frac{4}{3}\right)^n, \ldots$ so that the length $L_n \to \infty$ as $n \to \infty$. However the area under the curve is $1 + \frac{4}{9} + \left(\frac{4}{9}\right)^2 + \ldots = \frac{9}{5}$.
3. For a stick length $l_n = 1/3^n$, the number of sticks is $C_n = 4^n$. For a scale factor $b = 1/3$, the ratio of the two numbers $C_{n+1}/C_n = 4$. The fractal dimension is $d_{\rm f} = \lim_{n\to\infty} \frac{\ln(C_n/C_{n-1})}{\ln(1/b)} = \frac{\ln 4}{\ln 3}$

A practical procedure is to cover the curve with a square grid of unit length $l = 1$ and count the number of boxes C_l occupied by the curve. Then change the grid size by a scale factor b and count C_{bl}. One may then use the slope of the log-log plot with the definition of Eq. (10.10) to determine $d_{\rm f}$.

If we take $\epsilon = 1/3^n$ as the length of the measuring stick with $\epsilon \to 0$ as $n \to \infty$, then the measured length $(4/3)^n$ is dependent on the scale via $n = -\ln\epsilon/\ln 3$. We may generalize this result. The length measured at scale ϵ behaves as

$$L(\epsilon) \stackrel{\epsilon\to 0}{=} L_0 \epsilon^{-\alpha}, \quad \text{with } \alpha = d_{\rm f} - d_{\rm t}. \tag{10.15}$$



n	0	1	2	3
l_n	1	1/2	1/4	1/8
C_n	1	3	9	27

Figure 10.5: (a)Iterative construction of the Sierpinski gasket. One way is to view it as an aggregation of triangles so that the side length increases by a factor of 2, with the inner one missing (shaded triangles). Another view is to punch holes and remove the inner 1/3 of every triangle. The length for counting (l_n) and the number (C_n) of triangles are noted underneath.

For those cases where the two dimensions match (as in R), the length is independent of the scale. In such cases, one may talk of the length of the curve, and such curves are called *rectifiable curve*.

We see a curve of topological dimension 1 but of a fractal dimension between 1 and 2. Koch curve is also an example of a continuous but nowhere differentiable curve. As a closed curve, Fig. 10.4b, we get a continuous curve enclosing a finite area, though of infinite length. Fig. 10.4b is therefore topologically equivalent to S^1, where we now recognize the superscript as the topological dimension of the boundary.

Problem 10.4.5: Construct a space filling curve, i,e, a curve of topological dimension 1 but of fractal dimension 2. An example is the Peano curve. See Ref. [5].

10.4.3 Sierpinski Gasket: $d_t = 1, d_f > 1$

The construction of the Sierpinski gasket is shown in Fig. 10.5. This fractal can be disconnected by isolated points and is therefore topologically one-dimensional.

A fractal that can be disconnected by a finite set of points is called a finitely ramified fractal. A finitely ramified fractal is therefore has a topological dimension 1. The fact that we need 3 copies at a scale factor of 2 tells us that the fractal dimension of the Sierpinski gasket is $d_f = \frac{\ln 3}{\ln 2}$.

An intuitive way of arguing that its fractal dimension is less than 2 is to note that there are holes at every scale, and therefore holes will be present no matter at what resolution we look at, unlike a compact object. This of course requires $n \to \infty$.

Problem 10.4.6: Show that the Sierpinski gasket is topologically equivalent to S^1 but it is not rectifiable (i.e., its length $\to \infty$).

Problem 10.4.7: Sierpinski Carpet: Take a square of side length 1. Divide each side into three pieces of length $1/3$ and remove the inner square of area $1/9$. Repeat this process ad infinitum.

(a) Show that the total area is zero.

(b) Show that the carpet consists of points with a ternary expansion $(\sum_{n=1}^{\infty} \frac{a_n}{3^n}, \sum_{n=1}^{\infty} \frac{b_n}{3^n})$ where $(a_n, b_n) \in \mathcal{S}$, $\mathcal{S} = \{0,1,2\} \times \{0,1,2\} - \{(1,1)\}$. The topological space is $\mathcal{S}^{Z_+}$.

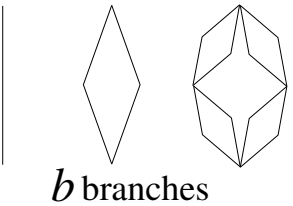
(c) Identify the Cantor sets along the diagonal and the medians.

(c) Show that the fractal dimension is $d_{\rm f} = \frac{\ln 8}{\ln 3}$.

(d) Show that it is an infinitely ramified fractal, but with topological dimension $= 1$. Note that the topological dimension by definition is an integer. For the carpet, it has to be less than 2, and it is not 0. Hence it is 1. Construct a direct proof of this.

(e) Show that the Sierpinski Gasket can be homeomorphically embedded in the Sierpinski carpet. Prove a more general statement: "any Jordan curve[7] can be homeomorphically embedded in the Sierpinski carpet." This was proved by Sierpinski in 1916.

Problem 10.4.8: Hierarchical lattices



A hierarchical lattice is built by successive replacement of a bond by a motif. In the example the motif consists of a diamond like object with b branches and $2b$ bonds. This is also called a diamond hierarchical lattice. Show that the fractal dimension is $d_{\rm f} = \frac{\ln 2b}{\ln 2}$.

The hierarchical construction of such lattices helps in easy implementation of renormalization group transformations or scaling [7].

Problem 10.4.9: *Julia set:* For an iterated map $z_{n+1} = f(z_n)$, an arbitrary point z would flow to a stable fixed point given by $z^* = f(z^*)$. In the complex z plane, there may however be a set of points which do not flow to any of these stable fixed points. An obvious example is the unstable fixed point. This set of special points which do not flow to any of the fixed points on iterations is called the Julia set of the map [10]. As an example consider $z_{n+1} = z_n^2 + c$, where c is a complex number. Choose any complex number c and find the Julia set.
HINT: Write a computer program to do the iteration and plot the special points in the complex plane.

[7] A Jordan curve is a planar simple closed curve homeomorphic to a circle. Simple here means nonintersecting.

Take $c = 0$. (i) Show that any $z, |z| < 1$, flows to zero, while $z, |z| > 1$ flows to infinity. (ii) The map has three fixed points $0, 1, \infty$, of which $z = 1$ is unstable. (Do it for real z.) (iii) Check that the points on the unit circle $|z| = 1$ do not flow to 0 or ∞. Therefore, the Julia set for $c = 0$ is the unit circle with dimension $= 1$.

For $c \neq 0$, show that, for small $|c|$, the Hausdorff dimension of the Julia set is given by (Ref. [11])

$$d_H = 1 + \frac{|c|^2}{4 \ln 2} + ...$$

10.4.4 Paths in Quantum mechanics: $d_t = 1, d_f = 2$

We now argue that the trajectories of a nonrelativistic quantum particle is a fractal obeying Eq. (10.15). The traditional Brownian motion also belongs to the same class.

From the uncertainty principle, fluctuations in position (δx) and momentum (δp) are related by $\delta x \, \delta p \sim \hbar$ and $\delta p = m \delta x / \delta t$, ($t$ being time), it follows that [12]

$$(\delta x)^2 \sim \delta t. \tag{10.16}$$

Take these $\delta x, \delta t$ as the scales for space and time to measure the length of a trajectory from (x_1, t_1) to (x_2, t_2). The time interval $T = t_2 - t_1$ consists of $N = T/\delta t$ pieces so that the length is

$$L = N \, \delta x = \frac{T}{\delta t} \delta x = \frac{T}{\delta x}, \tag{10.17}$$

where Eq. (10.16) has been used. Therefore, $L \to \infty$ as $\delta t \to 0$. We conclude, by comparing with Eq. (10.15), that the fractal dimension of the trajectory is $d_f = 2$, though a path, by definition, has a topological dimension $d_t = 1$.

Problem 10.4.10: For the cases where the propagator in the path integral approach can be calculated exactly, it is observed that only the classical path contributes. By definition, a *classical path* has $d_f = d_t = 1$. Show that this happens because of the interference of the nearby *quantum paths.*

10.5 Dimensions related to physical problems

To go beyond topology and geometry, we need to study some physical problem on a fractal. These could be of several types as mentioned in Sec. 10.2. Let's consider those cases one by one [13].

10.5.1 Spectral dimension

A physical way to explore a space is to use a probe that would in principle involve the whole space. In Euclidean space, a few such probes are (1) diffusion processes or random walks, (2) elastic waves or lattice vibration and (3) any quantum mechanics problem. The common link among the three types is via the Laplacian in the Euclidean space as follows.

1. The diffusion process is described by the differential equation for any diffusing field ψ,
$$\frac{\partial \psi}{\partial t} = D\nabla^2\psi, \tag{10.18}$$
where D is the diffusion constant.

2. The free particle Schrödinger equation is described by
$$i\hbar\frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\psi, \tag{10.19}$$
where m is the mass of the particle and $\hbar$ is the Planck constant divided by 2π.

3. The wave equation, e.g., describing sound waves is
$$\frac{\partial^2 \psi}{\partial t^2} = c^2\nabla^2\psi. \tag{10.20}$$

The diffusion equation also occurs in heat transport and defining the Laplacian for any space is often called the heat-kernel problem. For some of the fractals, it is easier to consider the lattice vibration problem or the scalar version the resistor problem in the zero frequency limit. The dimensionality of the space can then be defined by the appropriate generalization of the Euclidean results. This we do below. The dimension we obtain in this way is called the *spectral dimension* d_s.

Diffusion, random walk

For a diffusion problem in $\mathbb{R}^{d}$[8] the probability distribution of the end to end vector $\mathbf{R}$ is a Gaussian
$$P(\mathbf{R}) = \frac{1}{(2\pi t)^{d/2}}\ \exp\left(-\frac{dR^2}{2t}\right) \equiv \frac{1}{(2\pi t)^{d/2}}\ f(R/\sqrt{t}). \tag{10.21}$$
In this limit short scale details are not expected to be important. A general form allowing a dimensionality dependence is to write it as
$$P(\mathbf{R}) \sim t^{-d_s/2}\ f(R/\sqrt{t^{1/d_w}}), \tag{10.22}$$
defining a walk dimension d_w and a spectral dimension d_s. For $\mathbb{R}^n$, $d_s = n, d_w = 2$. Since probability is normalized, where the integration over $\mathbf{R}$ involves the fractal dimension, a change of variable gives
$$\int d^{d_{\rm f}}R\ P(\mathbf{R}) \sim t^{d_{\rm f}/d_w - d_s/2}.$$

[8]For a random walk on a lattice, the Gaussian distribution of Eq. (10.21) is valid for lengths.

Since the power of t should be zero, it follows that

$$d_s = 2\frac{d_{\rm f}}{d_w}. \tag{10.23}$$

Only for $d_w = 2$, the Hausdorff and the spectral dimensions match. This turns out to be the case for hierarchical lattices of Prob. 3.6.

Based on Eq. (10.22), the spectral dimension can be defined from the return to origin probability $P(0,t)$ as

$$d_s = -2 \lim_{t\to\infty} \frac{\ln P(0,t)}{\ln t}. \tag{10.24}$$

As for the paths in quantum mechanics (Sec. 10.4.4), d_w is the fractal dimension of the walk. It may also be seen as the scaling relation between space rescaling and time rescaling as determined by the dynamics. In such contexts, d_w, denoted by z, is called the dynamic exponent. For diffusion like processes $z = 2$. The dynamic exponent is an important characteristic quantity near various phase transitions.

Density of states

For a quantum mechanical problem with energy dispersion relation $\omega = Ck^p$ for small wave vector $k = |\mathbf{k}|$ in d dimensions, the density of states for $\omega \to 0$ is $\rho(\omega) \sim \omega^{-1+d/p}$, with the integrated density of states as $I(\omega) \sim \omega^{-d/p}$. For the electronic problem with a quadratic dispersion relation, $p = 2$ and one gets $\rho(\omega) \sim \omega^{-1+d/2}, I(\omega) \sim \omega^{d/2}$. For phonons, $p = 1$ and so, $\rho_{\rm phonon}(\omega) \sim \omega^{d-1}$, which gives the Debye law for specific heat. The low energy excitations involve $k \to 0$ (long wavelength) for which small scale details do not matter. In other words, the large scale dynamical features of the fractal or the object are probed by the long wavelength propagation.[9] As a generalization, d in the above densities may be replaced by d_s which is called the spectral dimension.

It is shown in Appendix B that d_s for the Sierpinski Gasket is different from its fractal dimension by explicitly calculating the spectral dimension.

10.6 Which *d*?

We now go back to the questions asked at the beginning of this chapter, Sec. 10.2.

10.6.1 Thermodynamic equation of state

The thermodynamic relation alluded to at the beginning actually comes from the density of states. In the grand canonical ensemble, the grand potential

[9] One may recall that the planar (two dimensional) structure of graphite was predicted from the T^2 dependence of the low temperature lattice specific heat.

(the equivalent of free energy) is $-PV$, and for a noninteracting gas, it can be written as

$$PV = k_B T \int d\epsilon\ \rho(\epsilon)\ \ln \mathcal{Z}(\epsilon, \beta, \mu), \tag{10.25}$$

where ϵ is the energy, $\rho(\epsilon)$ the density of states, $\beta = 1/k_B T$ the inverse temperature, μ the chemical potential, and $\mathcal{Z}$ is the single state grand canonical partition function,

$$\mathcal{Z} = \sum_{n=0}^{\infty} g_n \exp[-\beta(\epsilon - \mu)n], \tag{10.26}$$

where

$$g_n = \begin{cases} 1, & \text{for bosons,} \\ (n!)^{-1}, & \text{for classical particles,} \\ \left[\begin{array}{ll} 1, & \text{if } n = 0, 1 \\ 0, & \text{if } n > 1 \end{array}\right. & \text{for fermions,} \end{cases} \tag{10.27}$$

though the explicit form is not required for our argument.

The average particle number and the average energy are then given by

$$\langle N \rangle = \int d\epsilon\ \rho(\epsilon)\ \overline{n}(\epsilon, \beta, \mu), \quad \text{and} \quad U = \int d\epsilon\ \rho(\epsilon)\ \epsilon\ \overline{n}(\epsilon, \beta, \mu), \tag{10.28}$$

where

$$\overline{n}(\epsilon, \beta, \mu) = k_B T \frac{\partial \ln \mathcal{Z}}{\partial \epsilon}. \tag{10.29}$$

We now take the definition of spectral dimension to write $\rho(\epsilon) = A\epsilon^{-1+d_s/2}$ for free particles of mass m with a dispersion relation $\epsilon = p^2/2m$, with A some constant. The density of states may be divergent but is always integrable. With these information in hand, an integration by parts of the integral for U in Eq. (10.28), in conjunction with the (negative) grand potential in Eq. (10.25) gives us the required relation

$$PV = \frac{2}{d_s} U. \tag{10.30}$$

The thermodynamic fundamental relation for an ideal gas involves the spectral dimension of the space.

10.6.2 Phase transitions

The notion of lower critical dimension arose from studies of symmetry breaking in various dimensions. See Refs. [8,9] for an introduction to critical phenomena.

Two different cases are to be considered, namely a discrete or a continuous symmetry breaking. To be concrete, it is better to consider a particular case. A typical Hamiltonian is $H = -J \sum_{<ij>} \mathbf{S}_i \cdot \mathbf{S}_j$ where $\mathbf{S}_i$ is a unit n-component spin vector at site i of say a hypercubic lattice. For the Ising model $S_i = \pm 1$ is just a discrete variable with $n = 1$. Continuous cases correspond to $n \geq 2$; the

planar xy model has $n = 2$, the three dimensional spin $n = 3$ is the Heisenberg ferromagnet, etc. The Hamiltonian allows an ordered ground state where all spins are parallel. Therefore at zero temperature ($T = 0$) we get a ferromagnetic state that breaks the rotational invariance of H. Note that H remains invariant under a rotation of the spins by any amount if performed on all the spins (global symmetry). For the Ising case, the symmetry is discrete with $S_i \to -S_i$. As we raise temperature, thermal fluctuations tend to destroy the perfect alignment of spins due to entropic reasons. Therefore the question arises whether the broken symmetry state persists at nonzero temperatures. It is known that at very high temperatures entropy wins yielding a paramagnetic phase. In case the ordered state persists, then there has to be a special temperature $T = T_c$ (called the Curie point) at which a phase transition takes place. If we look at this particular H in various dimensions, then $T_c \neq 0$ only for $d > d_l$, where d_l is called the lower critical dimension. For the Ising case, $d_l = 1$ while for $n \geq 2$, $d_l = 2$. For a fractal, to which d are we referring?

Discrete case: Ising model

The Landau-Peierls argument for the Ising Hamiltonian is a generic way of determining the lower critical dimension for discrete symmetry breaking. Let us start with an ordered state, say all up spins at $T = 0$, and isolate a domain of opposite spins (down spins). By symmetry, the two states have the same energy, and, therefore, the cost of flipping the spins is in the creation of the boundary. Furthermore, same number of down spins can be enclosed by boundaries of different shapes. The boundary therefore has an entropy associated with it. Since the boundaries define the topological dimension of the system, we may write the change in free energy as

$$\Delta F = \sigma L^{d_{\rm t}-1} - T s_0 L^{d_{\rm t}-1}, \tag{10.31}$$

where σ is the energy cost of creating a unit "area" boundary and s_0 is the associated entropy. E.g., for the Ising case on a square lattice, $\sigma = 2J$. The free energy expression in Eq. (10.31) is valid for $d_{\rm t} > 1$ for which such flipped domains may not destroy the ordered state if $T < \sigma/s_0$. Thus a ferromagnetic state with broken symmetry can exist at nonzero temperatures. If $d_{\rm t} = 1$, the cost in energy is independent of the size but the flipped block can be placed anywhere on the lattice giving an entropy $\propto \ln N$ where N is the size of the system. For large N, entropy dominates, and so the system goes over to a paramagnetic state at any nonzero T. *The lower critical dimension is therefore one, and it is the topological dimension that matters.* This means the Ising model on a Sierpinski Gasket does not show a ferromagnetic state.

The above simple picture is not complete, because we have seen that the topological and the fractal dimensions are not sufficient to characterize all fractals. For example, the Ising model does not show any phase transition for the Sierpinski Gasket, a finitely ramified fractal, but does show a transition on the Sierpinski carpet, an infinitely ramified fractal, even though both fractals have

topological dimension 1. With many parameters, the lower critical dimension loses its significance. In any case, one may safely say that, for discrete symmetry, the condition of topological dimension ≤ 1 is necessary but not sufficient for *no* symmetry breaking transition [14].

Continuous case: crystal, xy, Heisenberg

Let us first consider the case of crystal, and ask the question whether the crystalline state, a state of continuous symmetry breaking, can survive in presence of lattice vibrations due to thermal fluctuations. Let $\mathbf{u}(\mathbf{r})$ be the small displacement of the particle at site $\mathbf{r}$ of the crystal. The thermally averaged correlation function $C(r) = \langle \mathbf{u}(0) \cdot \mathbf{u}(\mathbf{r}) \rangle$ tells us if the crystal state can be defined. In case this correlation does not decay to zero for large separation r, the long range order of the state gets destroyed. The independent vibration modes are the Fourier modes, $\mathbf{u}(\mathbf{k})$, with a dispersion relation $\omega_k = v|k|$, at least for small k, in Euclidean spaces. As independent oscillators, by equipartition theorem, $\langle u(k)^2 \rangle \sim k_B T/\omega^2$ with ω in the range $(0, \omega_{\max})$. The real space correlation then involves an integral over all the modes,

$$C(r) \sim \int \rho(\omega) \, \frac{1}{\omega^2} \, d\omega.$$

With $\rho(\omega) \sim \omega^{d_s - 1}$, the correlation seems to diverge for $d_s \leq 2$. The divergence comes from the low frequency part which corresponds to vibrations spanning large distances, thus exploring the real space. It is therefore the *spectral dimension* that matters. The symmetry is restored by thermal fluctuations if $d_s \leq 2$. Note that $d_s = 2$ is excluded for the ordered state to exist.

Similar arguments can be used for the xy or the Heisenberg magnets. It is still important to know why the arguments for the discrete symmetry cannot be applied here. As the order parameter space is continuous and connected[10], there is no well defined domain wall separating the states. Even if we start with a sharp wall, the variations near the wall can be smoothened-out to make a thicker wall. The thicker the wall, the less costly it is, invalidating the Landau-Peierls argument for discrete symmetry.

10.6.3 Bound states in quantum mechanics

Let us consider a particle in an attractive short range potential well. We know from explicit solutions that though a bound state is guaranteed in low dimensions, it is not so in higher dimensions. What is the borderline dimension?

In a path integral approach, the trajectories spending a large fraction of time in the well contribute to the propagator for a bound state. These paths also involve excursions in the classically forbidden region. But once it is out of the well, it must come back for a bound state. The overall distance spanned

[10] See Chapters 1, 2 and 9 of this book.

(generally measured by the root mean squared distance) in the classically forbidden region determines the width, ξ, of the bound state wavefunction. The larger the width, the smaller is the bound state energy. By uncertainty principle, the energy, E, is given by $|E| \sim 1/\xi^2$, with $E \to 0$ as $\xi \to \infty$. The question therefore is tantamount to asking whether ξ can be infinite even when the well is not vanishing.

Suppose, we want the propagator or the Green function for the particle from the center of the well (origin) to origin, $K(0T|00)$ in time $T \to \infty$. Let $K_b(0t_2|0t_1)$ and $K_c(0t_2|0t_1)$ be the propagators for paths inside the well and in the classically forbidden region from time t_1 to t_2, and v be the tunneling coefficient. Then, treating time as a discrete variable (for simplicity)

$$K(0T|00) = K_b(0T|00) + \sum_{t_1,t_2} K_b(0T|0t_2) v K_c(0t_2|0t_1) v K_b(0t_1|00) + ... \quad (10.32)$$

Generating functions can be introduced as

$$G(z) = \sum_t K(0t|00) z^t, G_b(z) = \sum_t K_b(0t|00) z^t, G_c(z) = \sum_t K_c(0t|00) z^t,$$

so that Eq. (10.32) for $T \to \infty$ can be written as a geometric series

$$\begin{aligned} G(z) &= G_b(z) + G_b(z) v^2 G_c(z) G_b(z) \\ &\quad + G_b(z) v^2 G_c(z) G_b(z) v^2 G_c(z) G_b(z) + ... \\ &= \frac{G_b(z)}{1 - v^2 G_c(z) G_b(z)}. \end{aligned} \quad (10.33)$$

The singularity coming from the denominator of Eq. (10.33) determines the quantum bound state energy, while the singularities of G_b and G_c determine the energies of the classical bound state and the unbound state.

The important quantity here is then the return probability, that a particle going from the origin comes back to origin in time t. The probability of returning to origin in time t is given by $K_c \sim t^{-d_s/2}$, as we saw in Sec. 10.5.1. $G_c(z)$ is then determined by d_s. A direct analysis of the singularities of $G(z)$ shows that a bound state with excursions in the classically forbidden region may not exist if $d_s > 2$. We refer to Ref. [15] for details. In short, the existence of a bound state is determined by the spectral dimension of the space.

10.7 Beyond geometry: engineering and anomalous dimensions

It is possible to go beyond geometric figures, and use the ideas of the previous sections in a broader context. Any function $f(x)$ is to be called scale invariant if $f(x) = b^\eta \, f(bx)$, under a scale transformation $x \to bx$. If this is true for any

b, we may choose $b = 1/x$ to get $f(x) = x^{-\eta} f(1)$, a pure power law. Most often scale invariance and power laws are used synonymously.[11]

10.7.1 Engineering dimension

In order to distinguish a pure power law from other types, let us consider a few special cases like,

$$f_2(x) = \frac{1}{(x+a)^c}, \quad f_3(x) = e^{-x/a}, \quad \text{and} \quad f_4(x) = \frac{e^{-x/a}}{x^c}. \quad (x > 0) \quad (10.34)$$

None of these functions show scale invariance in the true sense, but can be written as

$$f_j(x) = x^{-c}\, F_j\left(\frac{x}{a}\right), \quad (j = 2, 3, 4), \quad (10.35)$$

with $c = 0$ for f_3, and F_j another function. Such forms are called scaling forms and can be arrived at by a dimensional analysis. Taking x, a as lengths, if $f_i(x)$ has a dimension of L^{-c}, L being the dimension of length, then the prefactor x^{-c} takes care of the dimension of the function, with F_j taking care of the additional dependence on x and on a. Since $F_j(z)$ has to be dimensionless, its argument can only be x/a. This leads to the form of Eq. (10.35). The exponent $-c$ that comes from dimensional analysis is called the *engineering dimension* of f_j (see Eq. (10.3)).

If we scale all lengths by a factor b, $x \to bx, a \to ba$, then (keeping the a-dependence explicitly in the arguments)

$$f_{2,4}(bx, ba) = b^{-c}\, f_{2,4}(x, a), \quad \text{and } f_3(x, a) = f_3(bx, ba), \quad (10.36)$$

where the power of b in the prefactor just reflects the power one expects from dimensional analysis, its *engineering dimension*. By choosing $b = 1/x$, we recover the forms in Eq. (10.35).

10.7.2 Anomalous dimension

In situations where a refers to a small scale of the problem while x is a large length, e.g. a may be the microscopic range of interaction or lattice spacing while x may be a macroscopic distance, then $x/a \to \infty$ can be achieved by making $x \gg a$, or even by taking $a \to 0$. In this situation, naively a may be set to 0, with F_i approaching a constant. Here we see, $f_{2,4}(x \gg a) \sim x^{-c}$, while $f_3(x) \approx 0$.

There could be situations where $f(x) = x^{-c} F(x/a)$ is the correct form with the engineering dimension but $F(z) \overset{z\to\infty}{\to} (x/a)^{-\eta}$, then for $x/a \gg 1$, $f(x) \sim a^{\eta} x^{-c-\eta}$. Most importantly, even in the limit $x \gg a$, a cannot naively

[11] In contrast to the examples discussed earlier which had a discrete scale invariance (only particular values of b are allowed), this is a case of continuous scale invariance. This distinction is important.

be set to 0. The problem is often stated in a dramatic way by setting $a = 1$ to write $f(x) \sim x^{-c-\eta}$ creating an *illusion of violation* of the standard dimensional analysis. Consequently, this additional exponent η is called the *anomalous dimension* of f. It is more natural to call $c + \eta$ the scaling dimension.

There is actually no violation of dimensional analysis as can be seen by scaling both x and a because

$$f(bx, ba) = b^{-c} f(x, a). \tag{10.37}$$

If we just scale x, the large length scale, without scaling the intrinsic lengths like a, we get

$$f(bx, a) = b^{-c-\eta}\ f(x, a), \tag{10.38}$$

which, by choosing $b = 1/x$, says, $f(x, a) = \text{const}\ x^{-c-\eta}$ Note that scaling just a gives

$$f(x, ba) = b^{\eta} f(x, a). \tag{10.39}$$

Eq. (10.39) gives $f(bx, ba) = b^{\eta} f(bx, a)$, which can be combined with Eq. (10.37), to write $f(bx, a)$ in the form of Eq. (10.38). This suggests that the scaling behaviour of f for large x, i.e., how the function changes as the variable x is changed by a scale factor, can be determined by combining dimensional analysis (engineering dimension) with the changes expected as the short distance scale is changed.[12]

10.7.3 Renormalization group flow equations

One way to generate the anomalous scaling behaviour is to obtain the renormalization group (RG) flow equations. Let us treat a as a continuous variable and take $b = 1 + \delta l$ so that $ba = a + \delta a$, with $\delta a = a\ \delta l$. Then Eq. (10.39), by Taylor expansion, can be written as

$$a\ \frac{\partial f}{\partial a} = \eta\ f. \tag{10.40}$$

If we define a dimensionless quantity $\hat{f} = a^c\ f$, then by direct differentiation with respect a, and using Eq. (10.40), we obtain

$$a\ \frac{\partial \hat{f}}{\partial a} = c\hat{f} + \eta\ \hat{f}. \tag{10.41}$$

For $\eta = 0$, the above equation is the expected equation with the engineering dimension c (compare with Eq. (10.3)). The extra η-dependent term gives the anomalous contribution. Such equations that describe the change in the function as a microscopic cut-off like variable (here a) is changed, are called renormalization group flow equations. In general, in an RG flow equation, η would be dependent on the parameters of the problem, and only in special situations (called fixed points), η becomes a constant. Under those conditions, i.e., at the fixed points, a proper power law is obtained. Proper scale invariance is observed at these fixed points.

[12] In many practical situations, a plays the role of short distance cut-off.

Examples of flow equations

As an example. let there be two variables v_1, v_2 with engineering dimensions c_1, c_2 respectively. With an arbitrary length L, the dimensionless parameters are $u_1 = v_1 L^{-c_1}$, and $u_2 = v_2 L^{-c_2}$. The L dependence can be written in a form analogous to Eq. (10.2)

$$L\frac{\partial u_1}{\partial L} = -c_1 u_1, \quad \text{and} \quad L\frac{\partial u_2}{\partial L} = -c_2 u_2, \tag{10.42}$$

where v_1, v_2 are held constant. Now if it so happens that, due to interactions or nonlinearities, the actual L dependence takes a form

$$L\frac{\partial u_1}{\partial L} = -c_1 u_1 + b_1 u_1^2 + \mathrm{O}(u_1^3), \quad \text{and} \quad L\frac{\partial u_2}{\partial L} = -c_2 u_2 + b_{12} u_1 u_2 + ..., \tag{10.43}$$

then u_1 attains a scale independent value at the fixed point $u_1 = 0$ and $u_1^* = c_1/b_1$. These are fixed points because $\partial u_1/\partial L = 0$.

Around $u_1 = 0$, it is the engineering dimension that matters even for u_2, but at the nontrivial fixed point $u_1^* = c_1/b_1$, it seems that u_2 acquires a new dimension $-\hat{c}_2 = -c_2 + \eta$, where $\eta = b_{12} u_1^* = b_{12} c_1/b_1$. This η is the anomalous dimension of u_2. The idea of renormalization group (RG) is to obtain equations like Eq. (10.43) to study deviations from trivial behaviours. "Trivial" here, of course, means results obtained by dimensional analysis. [8, 9]

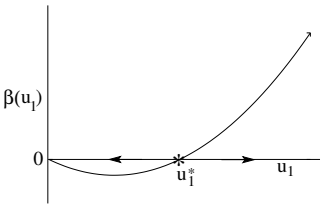


Figure 10.6: RG flow of u_1. The flow is described the β-function as given by Eq. (10.43). The zero of the β-function, u_1^* gives a critical point in this case as it is an unstable point. The flows on the two sides of u_1^* are indicated by the arrows.

Length scales from RG flow equations

To elaborate on the renormalization group behaviour, we define the RG flow equation as $L\partial u_1/\partial L = \beta(u_1)$. For concreteness, take $c_1, b_1 > 0$ with $u_1^* > 0$. The β-function is shown in Fig. 10.6, with u_1^* as a stable fixed point, a zero of the β-function. Any $u_1 < u_1^*$ flows to zero, while $u_1 > u_1^*$ flows to infinity. These can be checked by a direct integration of the flow equation for u_1. Therefore, $u_1 = u_1^*$ is a critical point separating the two phases described by $u_1 = 0$ and $u_1 = \infty$.

The growth of u_1 away from the fixed point is also an important characterization of the function. By linearizing around $u_1 = u_1^*$, with $\delta u = u_1 - u_1^*$,

$$L\frac{d\,\delta u}{dL} = \left(\frac{d\beta}{du_1}\right)_{u_1 = u_1^*} \delta u, \quad \text{or,} \quad \delta u = |u_{10} - u_1^*|\ (L/L_0)^{1/\nu}, \tag{10.44}$$

for the initial condition $u_1 = u_{10}$ at $L = L_0$, and $\nu = \left[(d\beta/du_1)_{u_1=u_1^*}\right]^{-1}$. We see that δu reaches a preassigned value Δ at a length ξ, where $\xi \sim |u_{10} - u_1^*|^{-\nu}$. In other words, ξ diverges as u_{10} approaches u_1^*. The existence of a diverging length is the hallmark of a critical point. It is at this critical point, the nontrivial fixed point in this example, that u_2 was found to acquire an anomalous dimension.

Problem 10.7.1: Consider a particle in three dimensions in a central potential of the form (i) $V(r) = -A/(r + a)$, (ii) $V(r) = -\frac{e^{-\alpha r}}{r}$, each of which reduces to the attractive Coulomb potential for $a, \alpha \to 0$. Discuss qualitatively the nature of the spectrum by comparing with the Hydrogen atom spectrum.

HINT: Which part of the Coulomb potential is changed by a or α? Consequently which part of the Hydrogen atom bound state spectrum ($E_n \sim -n^{-2}$, E_n being the energy of the nth state), will show a drastic change, small n or large n? It may help to compare the size of the nth state with the length scales from a or α.

10.7.4 Example: localization by disorder - scaling of conductance

Let us consider the conductance $g(L)$ of a metallic sample in the shape of a cube of side length L. For small sizes, the conductance of the sample is determined by the conductivity, σ_0. A macroscopic sample is obtained by successive rescaling L to $2L$ and so on. Now, the conductance is due to the propagating electrons. In a pure metal (say a crystalline sample), the electrons are completely delocalized as, e.g., described by the Bloch waves. In contrast, strong disorder, like impurities in the system, destroys the translational symmetry, and can, instead, produce localized states for the electrons. If these are localized over a length ξ, then one may observe some conductance for lengths $< \xi$ but not for $L \gg \xi$. A question of importance is whether a macroscopic sample remains metallic under disorder or there is a critical strength of disorder beyond which a metal becomes an insulator. A simpleminded RG approach helps in answering this question.

For a good conductor, we have $g(L) = \sigma_0 L^{d-1}/L \sim L^{d-2}$ for a d-dimensional hypercube, because the conductance is proportional to the geometric factor L^{d-1}/L. Defining $t = 1/(2\pi g)$, the scaling of t can be expressed, in analogy with Eq. (10.3) and Eq. (10.42), as [13]

$$\frac{\partial \ln t(L)}{\partial \ln L} = \epsilon \equiv (2 - d), \quad \text{or,} \quad L\frac{\partial t(L)}{\partial L} \equiv \beta(t) = \epsilon t, \quad (\text{small} \quad t). \tag{10.45}$$

This is the metallic regime.[14]

[13] It is made dimensionless by the the universal constant $e^2/\hbar$, where e is the electronic charge.

[14] This definition of the beta function follows the convention in statistical physics. The definition used in the original paper [16] involves the log derivative.

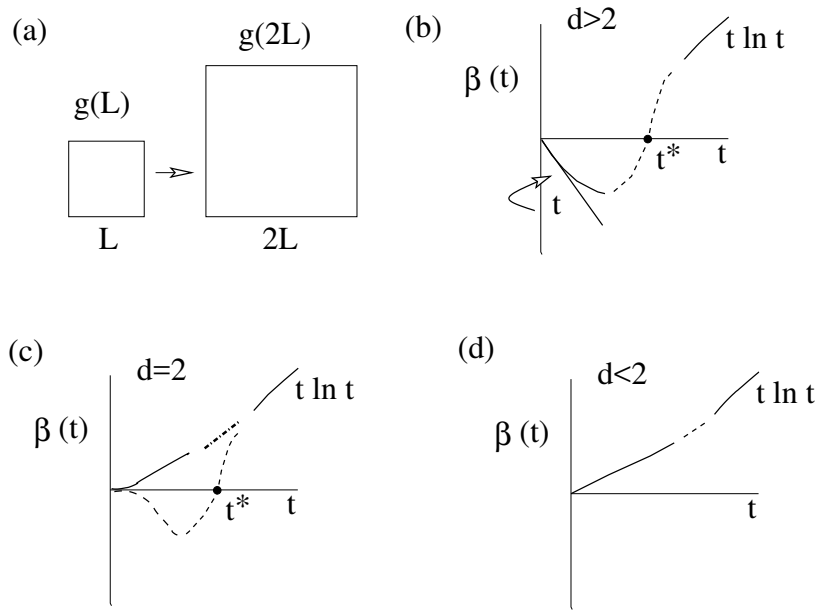


Figure 10.7: (a) Rescaling of a hypercubic box of side length L to $2L$. The conductance (> 0) goes from $g(L)$ to $g(2L)$. (b) The RG beta function $\beta(t)$, for $d > 2$, goes from $\beta(t) \sim -t$ for small t to positive values $\sim t \ln t$, for large t. The dotted line is a possible extrapolation which necessarily goes through a zero, $\beta(t^*) = 0$. Here, t^* is the RG fixed point. (c) For $d = 2$, the β-function has zero slope at origin. Therefore, there are two possibilities, one with a zero and one without any. (d) For $d < 2$, the slope at origin is positive, and there is unlikely to be any zero. The dashed lines are possible extrapolations

On the other hand, if the states are localized in the strong disorder limit, i.e., all states are localized with a localization length ξ, then the conductance can be expressed as $g(L) = g_0 \exp(-L/\xi)$. In the limit of large L, $g \to 0$, and so we expect

$$L\frac{\partial g(L)}{\partial L} = g \ \ln g, \quad (g \to 0), \quad \text{or,} \quad L\frac{\partial t(L)}{\partial L} = t \ \ln t, \quad (t \to \infty). \tag{10.46}$$

By assuming that $\beta(t)$ depends only on t (one parameter scaling, as for u_1 in Eq. (10.43), we combine Eqs. 10.45 and 10.46 as

$$\beta(t) = \begin{cases} \epsilon\, t, & \text{for } t \to 0, \\ t \ln t, & \text{for } t \to \infty. \end{cases} \tag{10.47}$$

We see that $\beta(t)$ for $d = 3$ goes from a negative value for small $t(> 0)$ to a positive value, as shown schematically in Fig. 10.7. This means that there is a fixed point, $\beta(t^*) = 0$, where the parameter t does not change with scale ("scale

invariant"). It is straightforward to see that for any initial value $t_0 < t^*$, the flow equation on integration to large L takes $t \to 0$. Therefore on large scales the system behaves like a metal. In contrast for $t_0 > t^*$ the flow goes to $t \to \infty$, an insulator. The difference in behaviour on the two sides of t^* is ensured by $\beta'(t^*) \equiv d\beta(t)/dt|_{t^*} > 0$. The unstable fixed point therefore represents a metal-insulator transition.

The power of the RG flow equation can be seen in several ways. If there is a fixed point with $\beta'(t^*) > 0$,i.e., an unstable fixed point, it represents a transition point. Since no such fixed point exists for $d < 2$, the flow always goes to the insulator region. In other words, all states will be localized for $d < 2$. The linearized form around the fixed point is

$$L\frac{d\Delta t}{dL} \approx \beta'(t^*)\,\Delta t, \quad \Delta t \equiv t - t^*. \tag{10.48}$$

For small $\Delta_0 \equiv |t_0 - t^*|$, we may now find the length scale ξ at which Δt reaches a predetermined value $\tilde{\Delta}$. On integration, Eq. 10.48 gives (omitting the subscript of t_0)

$$\xi \sim |t - t^*|^{-\nu}, \quad \text{where} \quad \nu = \frac{1}{\beta'(t^*)} \sim \frac{1}{|\epsilon|}. \tag{10.49}$$

In fact, Eq.10.49 is a very general prediction from RG for any critical point.

Problem 10.7.2: The disorder problem involves a Hamiltonian with random matrix elements. Depending on the symmetry, the disorder problem can be classified in several groups with distinct β-functions. For each of the following β-functions, determine the fixed point and the corresponding value of ν, and discuss the behaviour in two dimensions (i.e., for $\epsilon = 0$).

1. For a class, called the orthogonal symmetry class, $\beta(t) = \epsilon t + 2t^2 + ...$ Show that $t^* \approx |\epsilon|/2$ and $\nu = 1/|\epsilon|, \epsilon < 0$. In two dimensions, the flow is towards the insulator side, i.e., to a state where all states are localized.

2. For a class, called the unitary symmetry class, $\beta(t) = \epsilon t + 2t^3 + ...$ Show that $t^* \approx \sqrt{|\epsilon|/2}, \nu = 1/(2|\epsilon|), \epsilon < 0$. The two dimensional behaviour is same as the previous one, i.e., all states are localized.

3. There is a class called the symplectic class for which $\beta(t) = \epsilon t - t^2 + ...$ Show that at $d = 2$, the behaviour is different from the above two, because now the flow is towards the metal side.

For more details see F. Evers and A. D. Mirlin, Rev.Mod.Phys. **80**, 1355 (2008).

10.8 Multifractality

For completeness we mention another idea, namely multifractality. For a detailed discussion, see Ref [17].

In the quantum mechanics context, the difference between the wavefunctions of a bound state and an unbound state can be expressed in terms of the finite size behaviour. For a bound particle the size L of the box is not important so long it is much larger than the width of the wavefunction. For an unbound state $\psi \sim L^{-d/2}$. These can be combined to write the behaviour of the finite size effect of the moments of the wavefunction as

$$\int d^d r\ |\psi(r)|^{2q} \sim \begin{cases} L^0, & \text{(bound)}, \\ L^{-d(q-1)}, & \text{(unbound)}, \\ L^{-\tau_q}, & \text{(critical)}, \end{cases} \tag{10.50}$$

where we introduced a new class called "critical" wave function for which τ_q is not linear in q (note that the normalization condition requires $\tau_1 = 0$). For the bound state of width ξ, the insensitivity of the boundary, for $L \gg \xi$, is expressed by the power law L^0. For extended states, the moments are completely determined by the size of the box, with the exponent $d(q-1)$ following from dimensional analysis. In contrast, for the critical case, we find that for every moment a new length scale τ_p is required so that a critical wave function requires an infinite number of length scales to describe it. It is generally written as $\tau_q = d(q-1) + \Delta_q$, where Δ_q is the anomalous dimension. In other words, moments of ψ explore different aspects of how the wave function is spread out in space.

For a large system, we may also revert to the box counting method of Sec. 10.4 as follows. For a wave function $\psi(\mathbf{x})$, the probability $p(x) = |\psi(x)|^2$. So in the box method, instead of counting elements, we put an weight as

$$P_q = \sum_i \left(\int_{i\text{th box}} d^d x\ |\psi(\mathbf{x})|^2 \right)^q = \sum_i \left(\int d^d x p(x) \right)^q,$$

where the summation is over all the boxes of size b covering the sample.[15] There are $n = (L/b)^d$ number of boxes. Necessarily, $P_1 = 1, P_0 = n$. In analogy with Eq. (10.50), we define

$$P_q \sim \left(\frac{L}{b} \right)^{-\tau_q}. \tag{10.51}$$

In such a situation, the quantity of interest is the fractal dimension of the set of points where $p = |\psi|^2 \sim L^{-\alpha}$. Let the fractal dimension be given by $f(\alpha)$, i.e., the measure of the set of point with $|\psi|^2 \sim L^{-\alpha}$ is $L^{f(\alpha)}$. This fractal dimension depends continuously on α, justifying the name of "multifractal". We just state here that the multifractal spectrum $f(\alpha)$ is related to τ_q by a Legendre transformation as

$$f(\alpha) = q\alpha - \tau_q, \quad \text{where} \quad \alpha = \partial \tau_q / \partial q.$$

[15] In cases involving disorder, as in the localization problem, a sample averaged quantity $[P_q]_{\text{dis}}$ is to be calculated, where $[...]_{\text{dis}}$ denotes an averaging over samples of disorder. For convenience, we drop the averaging symbol.

The wavefunction at the localization transition mentioned in the Sec. 10.7.4 is an example of a multifractal wave function.

A tractable example of multifractal-like behaviour would be a power law decay of a wave function $\psi(r)$ which for large r behaves as $\sim r^{-u}$, where $1/2 < u < 1$ In this situation, since $\int dr|\psi|^2$ is convergent, the wave function is normalizable, and therefore represents a bound state. The unusual nature of the bound state can be seen from the behaviour of the qth moment given by $\int dr\ r^q\ |\psi|^2$ which is divergent for $q > 2u - 1$. Note that q is not necessarily an integer. A bound state with energy $E < 0$ gives a length scale, $\xi \sim 1/\sqrt{|E|}$ which one may associate with the scale beyond which the wave function decays exponentially as $\exp(-r/\xi)$. In a sense this gives the width of the wave function and for most cases, like the square well potential or short range potentials, this scale is enough. However the situation we are considering corresponds to the case where for $\xi \to \infty$, the wave function goes over to the power law decay so that for E very close to zero, there will be an intermediate range where the power law form is visible; for example a form like $\psi(r) \sim r^{-u}\ \exp(-r/\xi)$. A length scale to characterize the wavefunction in this intermediate range can be obtained from the moments as $\int dr\ r^q\ |\psi|^2 \sim \int^{\xi} dr\ r^{q-2u} \sim \xi^{q-2u+1}$, for $\xi \to \infty$. In this limit the normalization constant is a number independent of ξ. We therefore find $l_q \sim \xi^{\tau_q}$ where $\tau_q = 1 - (2u-1)/q$, with an anomalous exponent $D_q = -(2u-1)/q$. This is not a carefully crafted example but occurs at the unbinding transition of a quantum particle in a potential $V(r) + a/r^2$ in three dimensions where $V(r)$ is a short range attractive potential. By tuning the short range potential, a zero energy bound state can be formed whose wave function decays in the power law fashion just mentioned.

10.9 Conclusion

In this chapter, we explored various definitions of dimensions, like the topological, the Hausdorff, the box or Minkowski, and the spectral dimensions by embedding the set in Euclidean space. When the Hausdorff and the box dimensions are the same, it is called the fractal dimension, and the set is called a fractal if this fractal dimension is different from the topological dimension. For Euclidean spaces, all these definitions give the same number, which, by construction, is a positive integer. However we have explicitly constructed various subsets of the Euclidean space whose dimensions are not integers. Further generalizations are made to study power law behaviour of various physical quantities in terms of renormalization group flow equations. The flow equations show the emergence of anomalous dimensions at certain special fixed points as opposed to engineering dimension determined by dimensional analysis. We also discussed how various physical properties are determined by various dimensions of the space.

Appendix A: Entropy and fractal dimension

We here establish the relation between entropy per spin of a chain of spins and the fractal dimension of the configuration space when mapped on to real numbers. The discussion is for a semi-infinite chain of spins which can be labeled by integers, $1, 2, ...$

Let us consider a chain of spins, each taking two values, 0, and 1. If all configurations are equally likely to occur, then the number of configurations for N spins is $C_N = 2^N$. The entropy per spin is therefore $s = k_B(\ln C_N)/N = k_B \ln 2$. In general, the entropy per spin can be written as a derivative

$$s = k_B \ \frac{\ln C_{N+1} - \ln C_N}{(N+1) - N} = k_B \ \left.\frac{\partial \ln C_N}{\partial N}\right|_{N\to\infty}, \tag{10.52}$$

where the last term is in the continuum limit.

Now we convert the strings of 0 and 1, $(s_n = 0, 1 | n = 1, ..., N)$, to a real number by using base $m = 3$ as

$$x = \sum_{n=1}^{N} \frac{2s_n}{m^n}. \tag{10.53}$$

As discussed in Sec. 10.4.1, for $N \to \infty$, x forms the Cantor set. We chose 3 because of our familiarity with the Cantor set, but $m = 3$ is nothing special.

The fractal dimension of the subset of real numbers generated by all the spin configurations is given by the box dimension Eq. (10.11). As we go from N to $N+1$ in the number of spins, we add a higher order term $2s_{N+1}/m^{N+1}$ in x. This is equivalent to changing the scale of the box size from m^{-N} to $m^{-(N+1)}$, the latter being the finer scale. In other words, the "box"size has been changed by a scale factor $1/m$. The denominator of Eq. (10.11), with $b = 1/m$, becomes $\ln m^{N+1} - \ln m^N = [(N+1) - N] \ln m$. The fractal dimension of the set is therefore

$$\boxed{d_f = \frac{1}{\ln m} \ \frac{\ln C_{N+1} - \ln C_N}{(N+1) - N} = \frac{1}{\ln m} \ \frac{s}{k_B},}$$

where we used Eq. (10.53). This establishes the connection between the entropy per spin and the fractal dimension of the set of real numbers equivalent to the configuration space.

For the Ising case, with $m = 3$, we see $s \propto \ln 2/\ln 3 = \log_3 2$.

One may wonder, why we chose base 3 ($m = 3$). In general, for any choice of $m \geq 2$, the real numbers

$$x = \sum_{n=1}^{N} \frac{(m-1)s_n}{m^n}, \tag{10.54}$$

form a subset of $[0, 1]$. As the above result shows, we could have chosen any m to get $s \propto \ln 2/\ln m$. The extra factor $\ln m$ could easily be absorbed in k_B

which is equivalent to changing the base of the logarithm. Our choice of $m = 3$ is motivated by the fact that it is the smallest integer for uniqueness of the mapping, and our familiarity with the Cantor set.

Let us clarify the problem of mapping by taking the simplest situation of base 2 ($m = 2$). For any string we define $x = \sum_{i=1}^{\infty} s_n/2^n$, so that $x \in [0, 1]$. Note that we get the whole interval. However, two strings $s_1 s_2 ... s_n 011111....$ gives the same value of x as $s_1 s_2 ... s_n 1000000....$, like in traditional decimal system where $1.0 = 0.9999999....$ Therefore with base 2, we get a mapping from the binary strings to the real numbers $\in [0, 1]$, but it is not unique; this is a many to one mapping. In contrast, with base $m \geq 3$, we get a unique one-to-one, onto and invertible mapping via Eq. (10.54). Nevertheless, as Prob 4.3 shows, the entropy can be related to the dimension of the space of points generated out of the strings with any base m. This includes $m = 2$ with $\log_2 2 = 1$, the dimension of the interval $[0, 1]$!

In fields, like computing and telecommunication, $\log_2 2 = 1$ is used as a unit (called Shannon) of information content of one bit (same as entropy) [18].

The connection between entropy and fractal dimension for more general situations of spin chains are given as problems (Prob 4.1,4.2,4.3). Whether this connection can be extended to more general systems or general lattices remain to be seen.

Appendix B: Complex dimension: continuous and discrete Scaling

This appendix is technical in nature and may be skipped without loss of continuity [3, 19].

The self similarity discussed so far is of geometric nature. This may be true for any property of a system. If a function $f(x)$ satisfies a relation $f(x) = b^{\mu} f(bx)$ then $f(x)$ is said to be scale invariant as b may be viewed as a scale factor for the variable. If b is arbitrary, then we may choose $bx = 1$ to obtain $f(x) \sim x^{-\mu}$, a power law dependence on x. Thus power laws are synonymous to scale invariance - something one sees near critical points. Since b is arbitrary, such a scale invariance is called a continuous scale invariance.

The scale invariance we saw for the geometric fractals are not continuous but discrete. Instead of choosing powers of 3, suppose we choose some other $b, 3^n < 1/b < 3^{n+1}$ as the scale factor for the Cantor set. For such a scale factor $x = 1/b$, the number of pieces would remain the same, changing only when x matches with the correct scaling factor. It is then possible to write

$$C_x = C x^{d_f} F\left(\frac{\ln x}{\ln 3}\right), \tag{10.55}$$

where $F(z)$ is a periodic function of periodicity 1. A Fourier expansion gives

$$F(z) = \sum_n (a_n e^{i2\pi nz} + a_n^* e^{-i2\pi nz}), \tag{10.56}$$

where $*$ denotes complex conjugation. On substitution in Eq. (10.55), one gets a simpler power law form with

$$d'_{\rm f} = d_{\rm f} + i\frac{2\pi n}{\ln 3}, \quad n \in \mathbb{Z}, \tag{10.57}$$

a tower of complex dimensions. Restricting to the first mode, the oscillatory behaviour is of the form

$$C_x \sim x^{d_{\rm f}} \cos\left(2\pi \frac{\ln x}{\ln 3}\right). \tag{10.58}$$

Such an oscillatory behaviour for an arbitrary scale factor is a distinct signature of a discrete scale invariance. Whether these will have any important perceptible effect ultimately depends on the amplitudes a_n. In many situations, $|a_n|$ turns out to be extremely small compared to the nonoscillatory terms.

A notable example of a continuous scale invariance breaking into a discrete one is the Efimov effect in three body quantum mechanics or its classical analog in three stranded DNA.

B.1 Cantor string

Consider the complement of the Cantor set in the closed interval $[0, 1]$. This is the set of disjoint lengths (open intervals) which add up to a length 1, but still it is the whole line segment minus the set of points belonging to the Cantor set. A bounded open set of $\mathbb{R}$ is a fractal string, and the particular one we are discussing is the Cantor string. A poetic name is a one-dimensional drum with fractal boundary. A relevant question is "Does one hear the shape of a drum?"

The fractal sting is described by the set of lengths l_j, $j = 1, \infty$. For the Cantor string, these are $\frac{1}{3}, \frac{1}{9}, \frac{1}{9}, ...$, keeping track of the multiplicities (i.e., degeneracies), length l_j occurring m_j times. Let us define a zeta function

$$\zeta(s) = \sum_{j=1}^{\infty} m_j l_j^s, \tag{10.59}$$

where s is a complex number so that the series is convergent. For $s = 1$, $\zeta(1) = 1$, the length of the string. Since $l_j < 1$, the series is definitely convergent for large positive real s. The minimum real value of s for which it is convergent happens to be the fractal dimension of the boundary. On analytic continuation, one may define $\zeta(s)$ over the complex s-plane with singularities which are the complex dimensions of the boundary set. With an abuse of definition, the fractal dimension of the boundary is also called the dimension of the string.

For the Cantor string, $l_j = \frac{1}{3^j}$ with degeneracy $m_j = 2^{j-1}$, so that

$$\zeta(s) = \sum_{j=1} \frac{1}{3^s} \left(\frac{2}{3^s}\right)^{j-1} = \frac{3^{-s}}{1 - (2 \times 3^{-s})}, \tag{10.60}$$

which has poles at $3^s = 2$, with s given by Eq. (10.57).

Why should such a zeta function be useful? This becomes clear if we look upon Eq. (10.59) as a transformation for the weights w_j. Since $l_j < 1$ and $l_j \to 0$ for $j \to \infty$, one may, in a very nonrigorous way, write the function as an integral

$$\zeta(s) = \int_0^\infty x^{-s} w(x) dx, \tag{10.61}$$

identifying the zeta function as the Mellin transformation of the weight function. A quantity of interest is the number of intervals of size less than x, $N(x) \sim \int_0^x w(x)dx$ for which the zeta functions are useful.

Appendix C: Spectral dimension for the Sierpinski Gasket

We show how to calculate the spectral dimension in a particular case, viz., a scalar phonon problem on a Sierpinski gasket [20].

A scalar phonon problem on the Sierpinski gasket involves springs along the bonds of with equal masses at the sites but the restoring forces are added added disregarding the vectorial nature of the forces. This is equivalent to a resistor problem of finding the equivalent resistance between two sites if all the bonds are occupied by 1 Ohm resistors. By Kirchoff's law, all the voltages are linearly added.

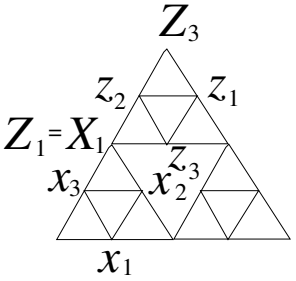


Figure 10.8: Notations for displacements or voltages in the scalar phonon problem. The capital letters will survive on elimination of the small letter variables from the equations.

With the notation $\lambda = m\omega^2/K$ where m is the mass at the sites and the spring constant on the bonds, the equation of motion for say x_1 is

$$(4 - \lambda)x_1 - x_2 - x_3 = X_2 + X_3, \tag{10.62}$$

while the equations for x_2, x_3 are obtained by appropriate permutations of the variables. Similarly for z_i's. For the capital variables,

$$(4 - \lambda)X_1 = x_2 + x_3 + z_2 + z_3, \tag{10.63}$$

and so on. It is now straightforward to eliminate x_i's and z_i's to obtain an equation involving only the capital variables, as

$$(4 = \lambda')X_1 = X_2 + X_3 + Z_2 + Z_3, \quad \lambda' = \lambda(5 - \lambda). \tag{10.64}$$

so that for small frequencies, $\lambda' = 5\lambda$.

Because of the change in the number of degrees of freedom under this scale change by $b = 2$, the density of states should change as

$$\rho_{L/b}(\omega) = b^{-d_{\rm f}} \rho_L(\omega), \tag{10.65}$$

while the frequency itself may scale as

$$\omega(L) = b^{-\kappa} \omega(L/b). \tag{10.66}$$

Since the number of states remain invariant, i.e., $\rho_{L/b}(\omega')d\omega' = \rho_L(\omega)d\omega$, we have (using Eqs. (10.65,10.66))

$$\rho_{L/b}(\omega) = b^{-\kappa} \rho_L(\omega b^{-\kappa}) = b^{-d_{\rm f}} \rho_L(\omega). \tag{10.67}$$

If we choose $\omega b^{-\kappa} = 1$, then $\rho_L(\omega) = \omega^{d_{\rm f}/\kappa - 1}$. The spectral dimension is therefore $d_s = d_{\rm f}/\kappa$.

From $\lambda' \approx 5\lambda$, we have $z^{2\kappa} = 5$, or, $\kappa = \frac{\ln 5}{2 \ln 2} \neq 1$. Combining all, we find the spectral dimension of the Sierpinski Gasket to be

$$d_s = \left(\frac{\ln 3}{\ln 2}\right) \Big/ \left(\frac{\ln 5}{2 \ln 2}\right) = 2\, \frac{\ln 3}{\ln 5}, \tag{10.68}$$

using the fractal dimension, $d_{\rm f} = \ln 3 / \ln 2$.

References

[1] Benoit B. Mandelbrot, *The Fractal Geometry of Nature* (W. H. Freeman and Company, 1982).

[2] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*, (Wiley, 3rd Edition, 2014).

[3] M. Lapidus and M. van Frankenhuijsen, *Fractal Geometry, Complex Dimensions and Zeta Functions Geometry and Spectra of Fractal Strings*, (Springer, New York, 2013).

[4] Michel Hans, Am. J. Phys. **51**, 694 (1983).

[5] For a list of fractals with fractal dimensions, see https://en.wikipedia.org/wiki/List_of_fractals_by_Hausdorff_dimension

[6] See Sec 9.14 of the online notes, J.Preskill, http://www.theory.caltech.edu/∼preskill/ph219/topological.pdf

[7] Robert B. Griffiths and Miron Kaufman, Phys. Rev. B **26**, 5022 (1982).

[8] S. M. Bhattacharjee, "Critical Phenomena: An Introduction from a modern perspective" in "*Field theoretic methods in condensed matter physics*" (TRiPS, Hindusthan Publising Agency, Pune, 2001). (cond-mat/0011011).

[9] J. Cardy, *Scaling and Renormalization in Statistical Physics*, (Cambridge Univ. Press, London, 1996).

[10] John Milnor, *Dynamics in One Complex Variable*, 3 ed, Hindusthan Book Agency, Pune, India (2012) (Indian edition). For a shorter version, https://arxiv.org/pdf/math/9201272v1.pdf.

[11] D. Ruelle, Ergod. Th. & Dynam. Sys., **2**, 99 (1982) (Appendix B).

[12] H. Kröger, Phys. Rept **323**, 81 (2000).

[13] J-P. Bouchaud and A. Georges, Phys. Rept. **195**, 127 (1990).

[14] Y. Gefen, B. B. Mandelbrot, and A. Aharony, Phys. Rev. Letts **45**, 855 (1980); J. Phys A: Math. Gen. **17**, 435 (1984); ibid **17**, 1277 (1984).

[15] S. Mukherji and S. M. Bhattacharjee, Phys. Rev. E **63**, 051103 (2001).

[16] E. Abrahams, P. W. Anderson, D. C. Licciardello, and T. V. Ramakrishnan, Phys. Rev. Lett. **42**, 673 (1979)

[17] M. Janssen, Int. J. Mod. Phys. B 8, 943 (1994).

[18] For a discussion on definitions of entropy, see, e.g., S. M. Bhattacharjee, "Entropy and perpetual computers", Physics Teacher. **45**, 12 (2003) (cond-mat/0310332).

[19] D. Sornette, Phys. Rept. **297**, 239 (1998).

[20] R. Rammal and G. Toulose, J. Physique Letts. **44**, L-13 (1985).

11

Quantum Geometry and Topology

R. Shankar

Attempts to understand the phenomenon of the robustness of the values of the Hall conductivity in quantum Hall systems led to the idea of characterizing the ground state of many electron systems using topological invariants. The discovery of the so called geometric phase in quantum systems led to the exploration of the quantum geometry of many electron ground states.

Quantum geometry is a general concept applicable to any quantum system. The physical states of any quantum system is the space of rays of a Hilbert space. The inner product induces a natural geometry in this space characterized by the so called quantum geometric tensor. These general features of quantum systems can be discussed in the context of one particle quantum mechanics. In these lectures, we do so in the context of a such a class of models relevant to condensed matter physics, namely, tight binding models.

We begin by introducing the quantum geometry in general. We then apply the concepts to tight binding models and discuss the quantum metric and the so called Pancharatnam curvature associated with the spectral bands. Finally we present some physical manifestations of these geometric constructs.

11.1 Introduction

Geometry is the mathematical description of shapes. It has been used in physics from the time physics became quantitative, eg. characterizing the shapes of the planetary orbits. Topology may loosely be thought of as the mathematics of connectivity. It has not been used as much as geometry for quantifying physical

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_11

systems. However, in the past few decades, after the discovery of the quantum Hall effect, there has been continuous research activity applying ideas of both geometry and topology to understand and characterize the phases (ground states at zero temperature) of many electron systems. While several of the ideas and concepts used are intrinsically "many-body", many of them can be understood and illustrated at the level of single-particle quantum mechanics. This set of lectures attempts to do this.

We begin by giving an example of what we mean by "geometry and topology" in quantum mechanics. Consequences of symmetry are heavily exploited to analyze quantum systems. Symmetry is a geometric property since it is a characteristic of the shape of the wavefunctions. A widely used consequence of symmetry is selection rules. For example, consider a rotationally invariant atomic system for which

$$[\mathbf{J}, H] = 0, \tag{11.1}$$

where H is the Hamiltonian and $\mathbf{J}$ is the angular momentum operator. We then have simultaneous eigenstates of the Hamiltonian, $\mathbf{J}\cdot\mathbf{J}$ and J^z,

$$\begin{aligned} H|n,j,m\rangle &= E_{nj}\,|n,j,m\rangle, & \text{(11.2a)}\\ \mathbf{J}\cdot\mathbf{J}|n,j,m\rangle &= \hbar^2 j(j+1)\,|n,j,m\rangle, & \text{(11.2b)}\\ J^z|n,j,m\rangle &= \hbar m\,|n,j,m\rangle. & \text{(11.2c)} \end{aligned}$$

These lead to selection rules as to what transitions are possible and what are prohibited when the system interacts with an electromagnetic field. A direct consequence of Eq.(11.2a) is that *matrix elements of the Hamiltonian between states of different j, m vanish.*

Next, we describe a consequence of quantum topology, the so called superselection rules. We are taught that there are two qualitatively different values for angular momentum, integer valued and half-odd integer valued. The particles with integer valued j are bosons and particles with half-odd integer valued j are fermions. The statistics of the particles is relevant only when we are considering two-particle (or more) wavefunctions. However, there is a consequence of the fact that these two are "qualitatively" different even in single particle quantum mechanics.

Consider a state which is a superposition of a $j = \frac{1}{2}$ and a $j = 1$ state,

$$|\psi\rangle = \alpha\left|\frac{1}{2},\frac{1}{2}\right\rangle + \beta|1,1\rangle. \tag{11.3}$$

Now let us perform a 2π rotation about the z-axis on $|\psi\rangle$,

$$U(2\pi,z)|\psi\rangle = -\alpha\left|\frac{1}{2},\frac{1}{2}\right\rangle + \beta|1,1\rangle. \tag{11.4}$$

The minus sign with the half-odd integer spin is exactly the "qualitative" difference between integer and half-odd integer spins. The consequence is that a 2π rotation on $|\psi\rangle$ changes it,

$$U(2\pi,z)|\psi\rangle \neq e^{i\phi}|\psi\rangle. \tag{11.5}$$

However, it is a *physical* requirement that 2π rotations leave the physical state unchanged. Thus we are led to the conclusion that *physical states cannot be superpositions of integer and half-odd integer angular momentum states.* This is a super-selection rule which is a consequence of quantum topology.

Geometry and Topology of what manifold? What is the relation of of that space to the physics of the system? Hopefully, the following sections will shed some light on these questions.

11.2 The space of physical states

11.2.1 Rays in Hilbert space

Quantum theory represents physical states by rays in a Hilbert space. What we mean by a "ray" in Hilbert space is the following. Consider a vector in Hilbert space, $|\psi\rangle$. We take it to be normalized, $\langle\psi|\psi\rangle = 1$. If $|\psi\rangle$ represents a physical state, then $e^{i\phi}|\psi\rangle$ represents the same physical state. This is the statement that the phase of the wave-function is not physically detectable. Thus the correspondence between normalized vectors in a Hilbert space and physical states is many-to-one.

If we think of an analogous situation in a three dimensional real Euclidean space: the set of normalized vectors are the set of unit vectors $\hat{n}$. Two unit vectors $\hat{n}$ and $-\hat{n}$ represent the same state. So each state is a direction, where we do not distinguish between the forward and backward direction. These are what we call rays.

The set of objects in the Hilbert space which are in one-to-one correspondence with physical states are the so called pure state density matrices or projectors [1],

$$\rho(\psi) \equiv \frac{|\psi\rangle\langle\psi|}{\langle\psi|\psi\rangle}. \tag{11.6}$$

It is easy to see that if we substitute $\lambda|\psi\rangle$ for $|\psi\rangle$ in Eq. (11.6), where λ is any complex number, then $\rho(\psi)$ is unchanged. Thus, the set of pure state density matrices are in one-to-one correspondence with physical states. These are generally defined by,

$$\rho^2 = \rho, \qquad \mathrm{tr}\rho = 1. \tag{11.7}$$

The space of physical states, namely the space of rays of a Hilbert space is called the projective Hilbert space.

11.2.2 Two level systems

To illustrate the above concepts, let us examine the simplest Hilbert space, namely a two level system. We denote an orthonormal basis by, $|n\rangle,\ n = 0, 1$. The general normalized vector can be written in terms of three parameters,

$$|\Omega, \theta, \phi\rangle = e^{i\Omega}\left(\cos\frac{\theta}{2}|0\rangle + \sin\frac{\theta}{2}e^{i\phi}|1\rangle\right). \tag{11.8}$$

The physical state does not depend on Ω. Further, $|\Omega, 0, \phi\rangle$ represents the same physical state for all ϕ. Similarly, $|\Omega, \pi, \phi\rangle$ also represents the same physical state for all ϕ. These identifications are the properties of the polar coordinates on the sphere. $(\theta, \phi) = (0, \phi)$ all represent the north pole and (π, ϕ) all represent the south pole. Thus the space of physical states of a two level system is a 2-sphere called the Bloch sphere.

This becomes very clear when we write down the projectors,

$$\rho(\theta, \phi) \equiv |\Omega, \theta, \phi\rangle\langle\Omega, \theta, \phi| \tag{11.9}$$

$$= \frac{1}{2}(I + \hat{n} \cdot \boldsymbol{\tau}) \tag{11.10}$$

where I is the identity matrix, $\boldsymbol{\tau}$ are the three Pauli spin matrices and $\hat{n}$ is the unit vector,

$$\hat{n} \equiv \sin\theta\cos\phi\hat{x} + \sin\theta\sin\phi\hat{y} + \cos\theta\hat{z} \tag{11.11}$$

Eq. (11.9) explicitly shows that the physical states of a two level system are in one-to-one correspondence with the set of unit vectors, namely points on a sphere.

11.2.3 *N*-level systems

We denote an orthonormal basis by, $|n\rangle,\ n = 0, \ldots, N-1$. The general vector can be written as

$$|z\rangle = \sum_{n=0}^{N-1} z_n |n\rangle \tag{11.12}$$

where z_n are N complex numbers ($2N$ real parameters). We denote the projector corresponding to $|z\rangle$ by,

$$\rho(z) \equiv \frac{|z\rangle\langle z|}{\langle z|z\rangle} \tag{11.13}$$

It follows that $\rho(z) = \rho(\lambda z)$, where λ is any complex number.

The space of physical states of an N level system are hence parameterized by $2N-2$ real parameters. They define a manifold called the complex projective space, denoted by $\mathbb{C}P_{N-1}$.

11.3 Quantum Geometry

11.3.1 The inner product and Bargmann invariants

Quantum theory gives a physical interpretation to the inner product of two states. Inner product $\langle\psi|\chi\rangle$ is defined to be the probability amplitude of finding the system to be in the state $|\psi\rangle$, given that it is in $|\chi\rangle$. The probability amplitude is not measurable. The measure-able quantity is the modulus square

of the probability amplitude, which is the probability of the event. This can be written in terms of the projectors,

$$|\langle\psi|\chi\rangle|^2 = \mathrm{tr}\left(\rho(\psi)\rho(\chi)\right). \tag{11.14}$$

It is clear that any measure-able quantity has to be expressible in terms of the projectors. We define a sequence of such quantities called Bargmann invariants [2].

$$B^{(N)}(\psi_1, \psi_2, \ldots, \psi_N)) \equiv \mathrm{tr}\left(\prod_{n=1}^{N} \rho(\psi_n)\right). \tag{11.15}$$

$B^{(2)}(\psi, \chi)$ is exactly the quantity in Eq. (11.14). We can write $B^{(3)}$ explicitly as,

$$B^{(3)}(\psi_1, \psi_2, \psi_3) = \langle\psi_1|\psi_2\rangle\langle\psi_2|\psi_3\rangle\langle\psi_3|\psi_1\rangle. \tag{11.16}$$

By construction, $B^{(3)}$ is measure-able. It is a complex number with a magnitude and a phase. Much of these notes will attempt to understand the physical meaning of the phase, in the context of the dynamics of a particle in a periodic potential.

11.3.2 Distance and Geometric Phase

The second Bargmann invariant has been used to define a distance between two states ψ_1 and ψ_2 [2],

$$d(\psi_1, \psi_2) = \sqrt{1 - \left(B^{(2)}(\psi_1, \psi_2)\right)^{\frac{\lambda}{2}}}, \tag{11.17}$$

where λ is a real number ≥ 1. The maximum distance between any two states is 1. This occurs when the two states are orthogonal to each other. The minimum distance of 0 occurs only between a state and itself. For $\lambda \geq 1$ the distance formula satisfies the triangle inequality,

$$d(\psi_1, \psi_2) + d(\psi_2, \psi_3) \geq d(\psi_3, \psi_1). \tag{11.18}$$

We denote the phase of the Bargmann invariants by,

$$B^{(N)}(\psi_1, \ldots \psi_N) \equiv e^{i\Omega^{(N)}(\psi_1, \ldots, \psi_N)} \left|B^{(N)}(\psi_1, \ldots \psi_N)\right|. \tag{11.19}$$

It is easy to check that the following additive law is valid,

$$\Omega^{(4)}(\psi_1, \psi_2, \psi_3, \psi_4) = \Omega^{(3)}(\psi_1, \psi_2, \psi_3) + \Omega^{(3)}(\psi_1, \psi_3, \psi_4). \tag{11.20}$$

It then follows that the phase of any Bargmann invariant can be constructed as a sum of 3-vertex Bargmann invariants. An $N-$vertex invariant defines an N-sided polygon in the space of rays, depicted for $N = 4$ in Fig. 11.1. This polygon can always be triangulated. The phase of the N-vertex invariant is the sum of the phases of the 3-vertex invariants of the triangles.

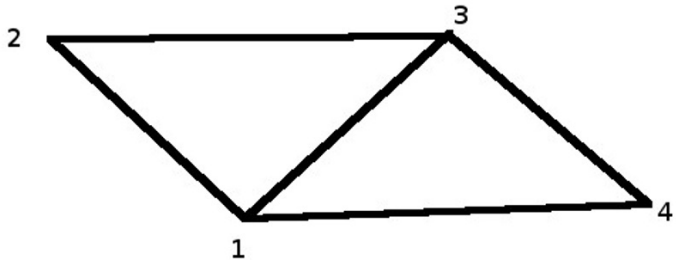


Figure 11.1: A depiction of the 4 states ψ_1, ψ_2, ψ_3 and ψ_4 as discussed in Eq. (11.20).

Mukunda and Simon had identified the phase of the Bargmann invariant with the so called geometric phase (also called the Berry phase or the Pancharatnam-Berry phase) of quantum mechanics. We refer the reader to the original paper [1] for a detailed and general discussion. Below, we will discuss the relation in a more restricted context.

11.3.3 The quantum geometric tensor

The space of physical states is very large and we are rarely interested in the whole space. In most cases, we are only interested in a small corner of it dictated by the physics of the problem. Eg., typically in condensed matter systems, we are interested in the ground state and low lying excitations. While these states are not known a-priori, very often we can guess that, to a good approximation, they lie in some sub-space. We then project our problem into this subspace and work there.

So in this section we concentrate on a subspace of the space of physical states of (real) dimension N_S. We parameterize the subspace by setting up a local coordinate system. We denote the local coordinates by $\xi_a, a = 1, ..., N_S$, where ξ_a are real numbers. The projectors representing the physical states are denoted by $\rho(\xi)$.

We will first examine the distance between nearby points and define a metric, $g_{ab}(\xi)$, on this subspace. We put the square of the distance between $\rho(\xi + d\xi)$ and $\rho(\xi)$ equal to $g_{ab}(\xi)d\xi^a d\xi^b$. We have,

$$d^2(\xi + d\xi, \xi) = 1 - \left(\mathrm{tr}\left(\rho(\xi + d\xi)\rho(\xi)\right)\right)^{\frac{\lambda}{2}} . \tag{11.21}$$

The fact that $\rho(\xi)$ are projectors implies

$$\mathrm{tr}\rho^2 = \mathrm{tr}\rho = 1 \;\Rightarrow\; \mathrm{tr}\left(\rho\partial_a\rho\right) = 0 \;\Rightarrow\; \mathrm{tr}\left(\partial_a\rho\partial_b\rho\right) = -\mathrm{tr}\left(\rho\partial_a\partial_b\rho\right), \tag{11.22}$$

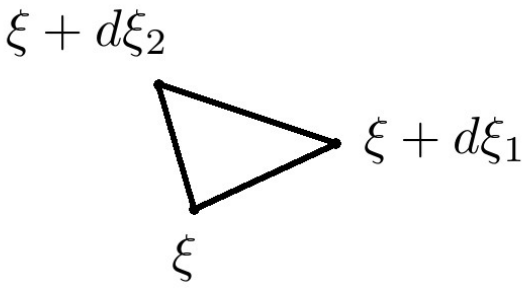


Figure 11.2: A depiction of the three states close to each other as discussed in Eq. (11.25).

where we use the notation $\partial_a \equiv \frac{\partial}{\partial \xi^a}$. Expanding the RHS of Eq. (11.21) in a Taylor series and using the properties in Eq. (11.22),

$$d^2(\xi + d\xi, \xi) = \frac{\lambda}{4}\, \mathrm{tr}\left(\partial_a \rho(\xi) \partial_b \rho(\xi)\right) d\xi^a d\xi^b. \tag{11.23}$$

Thus the distance formula in Eq. (11.21) implies a metric on the subspace,

$$g_{ab}(\xi) = \frac{\lambda}{4}\, \mathrm{tr}\left(\partial_a \rho(\xi) \partial_b \rho(\xi)\right). \tag{11.24}$$

Next let us examine the phase of the 3-vertex Bargmann invariant when the three vertices are very close to each other,

$$\Omega(\xi, \xi + d\xi_1, \xi + d\xi_2) = \mathrm{Im}\ \ln\left(\mathrm{tr}\left(\rho(\xi)\rho(\xi + d\xi_1)\rho(\xi + d\xi_2)\right)\right). \tag{11.25}$$

Expanding the RHS of Eq. (11.25) in a Taylor series and using the properties in Eq. (11.22),

$$\Omega(\xi, \xi + d\xi_1, \xi + d\xi_2) = \frac{1}{2i}\mathrm{tr}\left(\rho(\xi)\left[\partial_a \rho(\xi), \partial_b \rho(\xi)\right]\right) d\xi_1^a d\xi_2^b. \tag{11.26}$$

Before going further, let us discuss the RHS of Eq. (11.26) for the case of $N_S = 3$, i.e., when $a, b = 1, 2, 3$. We can re-write it as,

$$\begin{aligned}
\Omega(\boldsymbol{\xi}, \boldsymbol{\xi} + d\boldsymbol{\xi}_1, \boldsymbol{\xi} + d\boldsymbol{\xi}_2) &= \frac{1}{2i}\mathrm{tr}\left(\rho(\boldsymbol{\xi})\left[\partial_a \rho(\boldsymbol{\xi}), \partial_b \rho(\boldsymbol{\xi})\right]\right) \frac{1}{2}\left(d\xi_1^a d\xi_2^b - d\xi_1^b d\xi_2^a\right) \\
&= \frac{1}{2i}\mathrm{tr}\left(\rho(\boldsymbol{\xi})\left[\partial_a \rho(\boldsymbol{\xi}), \partial_b \rho(\boldsymbol{\xi})\right]\right) \frac{1}{2}\epsilon_{abc}\left(\mathbf{d\boldsymbol{\xi}}_1 \times d\boldsymbol{\xi}_2\right)^c,
\end{aligned}$$

to express it in terms of vectors as

$$\Omega(\boldsymbol{\xi}, \boldsymbol{\xi} + d\boldsymbol{\xi}_1, \boldsymbol{\xi} + d\boldsymbol{\xi}_2) = \frac{1}{2}\mathcal{B}(\boldsymbol{\xi}) \cdot \left(d\boldsymbol{\xi}_1 \times d\boldsymbol{\xi}_2\right), \tag{11.27a}$$

$$\mathcal{B}_a(\boldsymbol{\xi}) \equiv \frac{1}{2i}\epsilon_{abc}\mathrm{tr}\left(\rho(\boldsymbol{\xi})\left[\partial_b \rho(\boldsymbol{\xi}), \partial_c \rho(\boldsymbol{\xi})\right]\right). \tag{11.27b}$$

Eq. (11.27a) implies that the phase can be interpreted as the flux of a vector field, $\mathcal{B}$ defined in Eq. (11.27b), passing through the small triangle depicted in Fig. 11.2. Note that $\frac{1}{2}d\boldsymbol{\xi}_1 \times d\boldsymbol{\xi}_2$ is the area element of the small triangle.

This interpretation, suitably generalized, holds for all dimensions. [1]We define an anti-symmetric tensor,

$$\mathcal{F}_{ab}(\boldsymbol{\xi}) \equiv \frac{1}{2i}\mathrm{tr}\left(\rho(\boldsymbol{\xi})\left[\partial_a\rho(\boldsymbol{\xi}), \partial_b\rho(\boldsymbol{\xi})\right]\right). \tag{11.28}$$

We parametrize the points on the small triangle by two parameters, $\boldsymbol{\xi}(s,t)$ such that $\boldsymbol{\xi}(s+ds,t) = \boldsymbol{\xi} + d\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}(s,t+dt) = \boldsymbol{\xi} + d\boldsymbol{\xi}_2$. Thus $d\xi_1^a = \partial_s\xi^a ds$ and $d\xi_2^a = \partial_t\xi^a dt$. Eq. (11.26) can now be written as,

$$\begin{aligned}\Omega(\xi, \xi+d\xi_1, \xi+d\xi_2) &= \mathcal{F}_{ab}(\xi)\frac{1}{2}\left(\partial_s\xi^a\partial_t\xi^b - \partial_s\xi^b\partial_t\xi^a\right) dsdt &\text{(11.29a)}\\ &= \mathcal{F}_{ab}(\xi)\; d\xi^a \wedge d\xi^b. &\text{(11.29b)}\end{aligned}$$

Thus the phase of the 3-vertex Bargmann invariant can be interpreted as the integral of a two-form over a surface with the small triangle as the boundary.

As mentioned earlier, an N-vertex Bargmann invariant defines an N-sided polygon in the ray space. It can always be triangulated and its phase is the sum of the phases of these triangles. Since any closed curve in the ray space is a limit of polygons, we can use the above results and associate a phase with every closed curve in the ray space,

$$\Omega(\partial\Sigma) = \int_\Sigma \mathcal{F}_{ab}(\xi) d\xi^a \wedge d\xi^b, \tag{11.30}$$

where we have denoted the closed curve by $\partial\Sigma$ and a surface whose boundary is $\partial\Sigma$ by Σ.

The triangulation can be done in many ways, i.e., there are many surfaces with a given closed curve as its boundary. However, by construction, the phase of the Bargmann invariant is independent of the triangulation. This implies that $\mathcal{F}$ must be a closed 2-form ($d\mathcal{F} = 0$),

$$\partial_a\mathcal{F}_{bc} + \partial_b\mathcal{F}_{ca} + \partial_c\mathcal{F}_{ab} = 0. \tag{11.31}$$

It can be explicitly verified from Eq. (11.28) that this is indeed true. The fact that $\mathcal{F}$ is closed implies that it can be locally expressed as $d\mathcal{A}$, where $\mathcal{A}$ is a one-form,

$$\mathcal{F}_{ab} = \partial_a\mathcal{A}_b - \partial_b\mathcal{A}_a. \tag{11.32}$$

In terms of $\mathcal{A}$, the phase of the Bargmann invariant on $\partial\Sigma$ can be written as a line integral along the curve thus manifestly showing that it does not depend on what surface, Σ, is chosen, provided its boundary is $\partial\Sigma$. Therefore,

$$\Omega(\partial\Sigma) = \oint_{\partial\Sigma} \mathcal{A}_a d\xi^a. \tag{11.33}$$

[1]See Chapter 1 (Sec. 1.3) and Chapter 5 (Sec. 5.2) for differential forms, and Chapter 4 for triangulation.

The reader can explicitly verify that the following choice of $\mathcal{A}$ reproduces $\mathcal{F}$ defined in Eq. (11.28),

$$\mathcal{A}_a = \frac{1}{i}\langle\xi|\partial_a|\xi\rangle. \tag{11.34}$$

This is exactly the definition of the so called Pancharatnam-Berry (PB) connection. $\mathcal{F}$ is then the PB curvature. Thus the phase of the Bargmann invariant is the geometric phase of quantum mechanics.

A closed 2-form that we (physicists) are very familiar with is the dual to the magnetic field ($\boldsymbol{\nabla}\cdot\mathbf{B}=0$). So this analogy is made quite often in the literature.

The quantum geometric tensor: Definition

We have defined two second rank tensors, namely, the metric, g_{ab}, which is a symmetric tensor, and the PB curvature, $\mathcal{F}_{ab}$, which is anti-symmetric. The sum of the two is often called the quantum geometric tensor, [2]

$$Q_{ab} \equiv g_{ab} + \mathcal{F}_{ab}. \tag{11.35}$$

Note that the PB connection and the curvature are not the same as the affine connection and the Riemann curvature that can be constructed from the metric. The Christoffel symbols constructed from the metric define the parallel transport of vectors in the ray space, whereas the PB connection defines the parallel transport of wave-functions.

11.3.4 Examples

We will now work out some explicit examples of simple sub-spaces, the two level system described earlier and the set of coherent states of a one-dimensional particle.

Two level systems

As discussed earlier, the ray space for a two level system is the Bloch sphere and the projectors are given by Eq. (11.9),

$$\rho(\theta,\phi) = \frac{1}{2}\left(I + \hat{n}\cdot\boldsymbol{\tau}\right). \tag{11.36}$$

Let us work out the components of the quantum metric using Eq. (11.24). Using the notation, $\xi_1=\theta,\ \xi_2=\phi$,

$$g_{ab}(\theta,\phi) = \frac{\lambda}{8}\partial_a\hat{n}\cdot\partial_b\hat{n}. \tag{11.37}$$

[2] See Chapters 5 and 6.

The three independent components can be worked out to get,

$$g_{\theta\theta} = \frac{\lambda}{8}, \quad g_{\theta\phi} = 0, \quad g_{\phi\phi} = \frac{\lambda}{8}\sin^2\theta. \tag{11.38}$$

This is proportional to the standard metric on a sphere induced by a Euclidean metric in the three dimensional space that it is imbedded in.

The PB curvature tensor has only one independent non-zero component. From Eqs. (11.28) and (11.36) we have,

$$\mathcal{F}(\theta,\phi)_{\theta\phi} = \frac{1}{2}\hat{n}\cdot(\partial_\theta\hat{n}\times\partial_\phi\hat{n}) \tag{11.39a}$$

$$= \frac{1}{2}\sin\theta. \tag{11.39b}$$

The geometric phase of a curve bounding a region Σ on the sphere is therefore,

$$\Omega = \frac{1}{2}\int_\Sigma \sin\theta\ d\theta d\phi. \tag{11.40}$$

This is (half) the area of, or (half) the solid angle subtended by Σ. In the analogy with the magnetic field, it would correspond to the magnetic field of a monopole of strength 1/2 placed at the center of the Bloch sphere.

Coherent states

Coherent states are an extremely useful subset of the states of a quantum mechanical particle. They are in one-to-one correspondence with points in the classical phase space. They are minimum uncertainty wave-packets peaked around a position and momentum.

Restricting ourselves to one dimension for simplicity, we denote the position and momentum operators by,

$$[\hat{x},\hat{p}] = i\hbar. \tag{11.41}$$

The coherent state vectors are defined as,

$$|x,p\rangle \equiv e^{\frac{i}{\hbar}(x\hat{p}-p\hat{x})}|0\rangle, \tag{11.42}$$

where $|0\rangle$ is a real gaussian wave packet, peaked about $x = 0$. It is easily verified that,

$$\langle x,p|\hat{x}|x,p\rangle = x, \qquad \langle x,p|\hat{p}|x,p\rangle = p. \tag{11.43}$$

The projectors provide a resolution of the identity,

$$\int \frac{dxdp}{2\pi\hbar}\ |x,p\rangle\langle x,p| = I. \tag{11.44}$$

The 2-vertex Bargmann invariant is computed to be,

$$|\langle x_1,p_1|x_2,p_2\rangle|^2 = e^{-\frac{1}{\hbar}\left(\frac{1}{m\omega}(p_1-p_2)^2+m\omega(x_1-x_2)^2\right)}. \tag{11.45}$$

This leads to a metric on the phase space,

$$g_{pp} = \frac{\lambda}{2m\omega\hbar}, \quad g_{px} = 0, \quad g_{xx} = \frac{\lambda m\omega}{2\hbar}, \tag{11.46}$$

where λ is a part of the the definition of the distance, and $m\omega$ is a part of the definition of the coherent states and is the parameter that determines how the minimum uncertainty is distributed between the position and the momenta.

The phase of the 3-vertex Bargmann invariant can also be explicitly computed to get,

$$\Omega\left((x_1,p_1),(x_2,p_2),(x_3,p_3)\right) = \frac{1}{2\hbar}\left((x_1-x_2)(p_2-p_3)-(x_2-x_3)(p_1-p_2)\right). \tag{11.47}$$

The RHS is the area of the triangle in phase space, $(x_1,p_1),(x_2,p_2),(x_3,p_3)$, in units of $\hbar$. In terms of the analogy to the magnetic field, we have a constant magnetic field perpendicular to the phase space plane. The PB curvature has only one independent non-zero component,

$$\mathcal{F}_{pp} = 0 = \mathcal{F}_{xx}, \ \mathcal{F}_{px} = \frac{1}{\hbar} = -\mathcal{F}_{xp}.$$

This is called the symplectic 2-form in classical mechanics where it describes the Poisson bracket structure.

11.4 Periodic systems

11.4.1 Tight-binding models

The application of the ideas of quantum geometry discussed in section 11.3 were, to this authors knowledge, first applied to the system of electrons in a periodic potential by Mazari and Vanderbilt [9]. In this section we will examine the quantum geometry of a class of such systems, N_B-band tight binding models.

Consider a d-dimensional Bravais lattice with sites given by

$$\mathbf{R}_I = \sum_{i=1}^{d} I_i \mathbf{e}_i, \tag{11.48}$$

where $\mathbf{e}_i$ are the d basis vectors. We will be mainly interested in $d = 2, 3$. Each unit cell has N_B orbitals centered at $\mathbf{R}_{I\alpha} = \mathbf{R}_I + \mathbf{r}_\alpha$, $\alpha = 1, \ldots, N_B$. We denote the wave functions of the orbitals by $|\mathbf{R}_I, \alpha\rangle$ and assume they form an orthonormal basis,

$$\langle \mathbf{R}_I, \alpha | \mathbf{R}_J, \beta\rangle = \delta_{\alpha\beta} \prod_{i=1}^{d} \delta_{I_i J_i}, \qquad \text{and} \qquad \sum_{\mathbf{R}_I, \alpha} |\mathbf{R}_I \alpha\rangle\langle \mathbf{R}_I \alpha| = I. \tag{11.49}$$

The Hamiltonian is given by,

$$\hat{h} = \sum_{I\alpha, J\beta} |\mathbf{R}_I \alpha\rangle h_{\alpha\beta}(\mathbf{R}_I - \mathbf{R}_J)\langle \mathbf{R}_J \beta|. \tag{11.50}$$

Translation invariance and the Brillouin zone

The Hamiltonian in Eq. (11.50) is invariant under lattice translations, $\hat{T}_i|\mathbf{R}_I, \alpha\rangle = |\mathbf{R}_I + \mathbf{e}_i, \alpha\rangle$. Thus we can have simultaneous eigenstates of $\hat{h}$ and $\hat{T}_i$ (Bloch's theorem). We choose these to be,

$$|\mathbf{k}, \alpha\rangle = \sum_I e^{i\mathbf{k}\cdot\mathbf{R}_{I\alpha}}|\mathbf{R}_I, \alpha\rangle, \tag{11.51}$$

where $\mathbf{k}$ are the pseudo-momenta taking values in the Brillouin zone. Namely from the above equation, the state $|\mathbf{k}, \alpha\rangle$ and the state $|\mathbf{k}+\mathbf{G}_i, \alpha\rangle$ are the same if,

$$\mathbf{G}_i \cdot \mathbf{e}_j = 2\pi\delta_{ij}. \tag{11.52}$$

Therefore, $\mathbf{k}$ and $\mathbf{k}+\mathbf{G}$, represent the same state.

The Hamiltonian in the $|\mathbf{k}, \alpha\rangle$ basis is,

$$\hat{h} = \sum_{\mathbf{k}\in BZ} \sum_{\alpha,\beta} |\mathbf{k}\alpha\rangle h_{\alpha,\beta}(\mathbf{k})\langle\mathbf{k}\beta|, \tag{11.53}$$

where BZ stands for the Brillouin zone which is a d-torus defined by the identifications,

$$\mathbf{k} \sim \mathbf{k} + \mathbf{G}_i, \qquad i = 1, \ldots, d. \tag{11.54}$$

11.4.2 Spectral bands

The eigenstates of $\hat{h}$ are of the form,

$$|n\mathbf{k}\rangle = u^n_\alpha(\mathbf{k})|\mathbf{k}\alpha\rangle, \tag{11.55}$$

where,

$$h_{\alpha,\beta}(\mathbf{k})u^n_\beta(\mathbf{k}) = \epsilon^n(\mathbf{k})u^n_\alpha(\mathbf{k}). \tag{11.56}$$

The spectrum of $\hat{h}$ is therefore,

$$\hat{h}|n\mathbf{k}\rangle = \epsilon^n(\mathbf{k})|n\mathbf{k}\rangle. \tag{11.57}$$

Thus, at each point in the Brillouin zone, $\mathbf{k}$, we have an N_B level system. The space of physical states for this system is $\mathbb{C}P_{N_B 1}$. The hopping matrix $h(k)$ has N_B energy eigenvalues labelled by n. Generically these eigenvalues are non-degenerate but of course not necessarily so. These are the energy bands.

The surface of energies

Each energy band, $\epsilon^n(\mathbf{k})$, is a function on the Brillouin zone. We can visualize it as a d-dimensional surface in the $d+1$-dimensional space consisting of the Brillouin zone and an extra orthogonal dimension representing the single particle energy (the E-k space). The resulting space is $T^d \times R^1$, where T^d is the d-dimensional torus and R^1 the real line. All the energy eigenvalues of the tight-binding Hamiltonian form N_B surfaces in this space, one corresponding to each band. Wherever two of the eigenvalues are degenerate, the surfaces of the corresponding bands touch.

The surface of states

The eigenstate of $\hat{h}$, corresponding to each energy band, defines a map from the Brillouin zone to the space of physical states, $\mathbb{C}P_{N_B-1}$,

$$\mathbf{k} \to \rho_n(k) \equiv |n\mathbf{k}\rangle\langle n\mathbf{k}|. \tag{11.58}$$

Thus each band of states can be visualized as a surface in the space of rays.

11.4.3 Examples

Let us examine some examples of two band models, $N_B = 2$. In this case, at each $\mathbf{k}$, we have a two level system. The space of physical states is the Bloch sphere.

1-d hopping

Consider spinless particles hopping on a one-dimensional lattice, with alternating hopping matrix elements. This system is a two band system in one dimension ($d = 1,\ N_B = 2$). We denote the orbitals by $|I\alpha\rangle,\ \alpha = 1, 2$. The Hamiltonian is

$$\hat{h} = \sum_I \ (|I,1\rangle\langle I,2| + h.c) + t' \left(|I,1\rangle\langle I+1,2| + h.c\right). \tag{11.59}$$

The Brillouin zone is a circle, which we take to be $-\pi \le k \le \pi$. In the k basis, as in Eq. (11.53),

$$\hat{h} = \int_{-\pi}^{\pi} \frac{dk}{2\pi}\ |k\alpha\rangle h_{\alpha\beta}(k)\langle k\beta|. \tag{11.60}$$

The Fourier transform of the hopping matrix can be written as,

$$\begin{aligned} h(k) &= \boldsymbol{\alpha}\cdot\mathbf{\Pi}(\mathbf{k}), && (11.61)\\ \alpha^x &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, && (11.62)\\ \alpha^y &= \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, && (11.63)\\ \Pi^x(\mathbf{k}) &= (1 + t'\cos k), && (11.64)\\ \Pi^y(\mathbf{k}) &= -t'\sin k. && (11.65) \end{aligned}$$

It is simple to solve for the spectrum. The two energy eigenvalues are,

$$\epsilon^{\pm}(k) = \pm\sqrt{(1-t')^2 + 4t'\cos^2\frac{k}{2}} \equiv \pm\epsilon(k). \tag{11.66}$$

The density matrices of the corresponding two eigenstates are

$$\rho_{\pm}(k) = \frac{1}{2}\left(I \pm \frac{1}{\epsilon}h(k)\right). \tag{11.67}$$

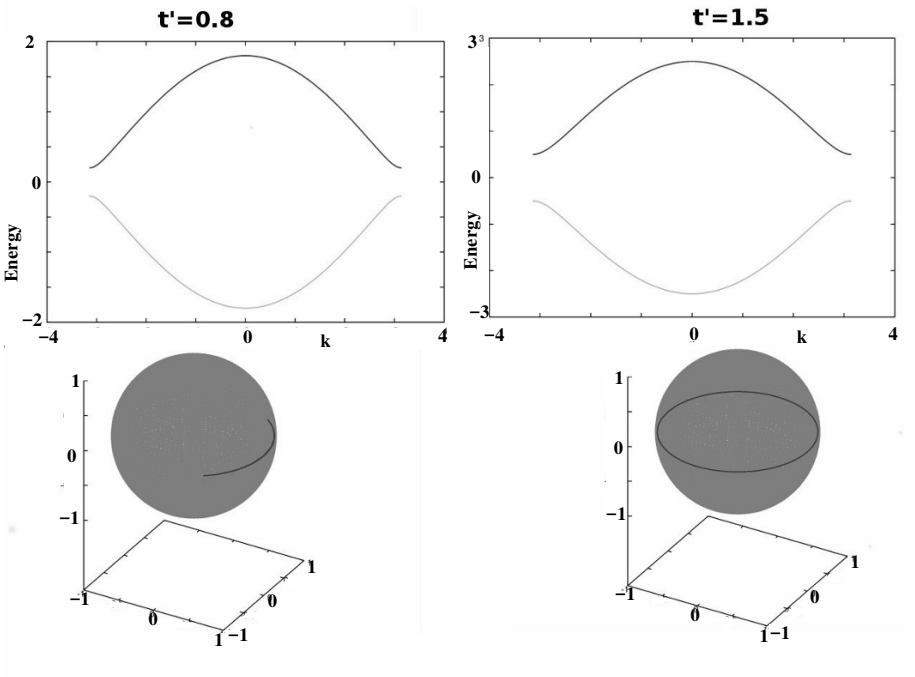


Figure 11.3: The two energy surfaces (curves) of the 1-d model are plotted on top. The positive energy band in dark and the negative energy band in light. The values of the parameter t' are indicated on the top. The surface (curve) of states of the positive energy band on the Bloch sphere (grey) is plotted in dark color at the bottom. The curve of the negative energy band will be the antipodal points.

To visualize the mapping of the BZ to the Bloch sphere, we use the parameterization in Eq. (11.9). The two polar angles of the points on the Bloch sphere corresponding to the two eigenstates are,

$$\theta_{\pm}(k) = \frac{\pi}{2}, \tag{11.68}$$

$$\phi_{\pm}(k) = \sin^{-1}\left(\mp\frac{\sin k}{\epsilon(k)}\right). \tag{11.69}$$

We see that the map lies on the equator. Figure 11.3 plots the energy surfaces and the surface of states for $t' = 0.8$ and 1.5. Only the positive energy band eigenstates are shown. For $t' \leq 1$, the curve is an arc of the equator. Examining the eigenstates, we can see that in this parameter range, two values of k map on to the same value of ϕ. Thus the map is not one to one. In the case of $t' \geq 1$, we have a one to one map and the curve winds around the equator once.

Honeycomb lattice

Next, consider the honeycomb lattice shown in Fig. 11.4. The sites of the lattice are,

$$\mathbf{R}_{I\alpha} = I_1\mathbf{e}_1 + I_2\mathbf{e}_2 + \mathbf{r}_\alpha, \tag{11.70}$$

$$\mathbf{e}_1 = \hat{x} \qquad \mathbf{e}_2 = -\frac{1}{2}\hat{x} + \frac{\sqrt{3}}{2}\hat{y}, \tag{11.71}$$

$$\mathbf{r}_{A(B)} = +(-)\frac{\sqrt{3}}{8}\hat{z}. \tag{11.72}$$

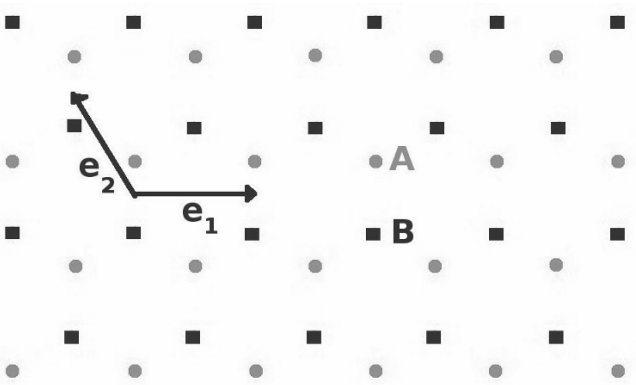


Figure 11.4: The honeycomb lattice. The basis vectors of the triangular Bravais lattice are $\mathbf{e}_1$ and $\mathbf{e}_2$. The two sub-lattices are formed by the square and the round symbols.

We study the Hamiltonian corresponding to a staggered onsite potential and nearest neighbour hopping,

$$\begin{aligned}\hat{h} &= \sum_I \Delta\left(|\mathbf{R}_{IA}\rangle\langle\mathbf{R}_{IA}| - \mathbf{R}_{IA}\rangle\langle\mathbf{R}_{IA}|\right) \\ &\quad + \sum_{\langle I\alpha, J\beta\rangle} \left(|\mathbf{R}_{I\alpha}\rangle\langle\mathbf{R}_{J\beta}| + |\mathbf{R}_{J\beta}\rangle\langle\mathbf{R}_{I\alpha}|\right),\end{aligned} \tag{11.73}$$

where the symbol $\langle I\alpha, J\beta\rangle$ denotes that the sum is only over nearest neighbours. The Brillouin zone is

$$\mathbf{k} = \frac{1}{2\pi}\left(k_1\mathbf{G}_1 + k_2\mathbf{G}_2\right), -\pi \leq k_1, k_2 \leq \pi,$$

where the reciprocal lattice vectors $\mathbf{G}_i$ are defined in Eq. (11.52). The Fourier transform of the hopping matrix in Eq. (11.73) is,

$$\begin{aligned} h(\mathbf{k}) &= \boldsymbol{\alpha}\cdot\boldsymbol{\Pi}(\mathbf{k}) + \beta\Delta, && (11.74)\\ \beta &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, && (11.75)\\ \Pi^x(\mathbf{k}) &= (1+\cos k_1 + \cos k_2)\,, && (11.76)\\ \Pi^y(\mathbf{k}) &= (\sin k_1 - \sin k_2)\,. && (11.77) \end{aligned}$$

The α matrices are defined in Eqs. (11.62, 11.63). The two energy eigenvalues are,

$$\epsilon^{\pm}(\mathbf{k}) = \pm\sqrt{\boldsymbol{\Pi}(\mathbf{k})\cdot\boldsymbol{\Pi}(\mathbf{k}) + \Delta^2} \equiv \pm\epsilon(\mathbf{k}). \tag{11.78}$$

The projectors on the two corresponding eigenstates are,

$$\rho_{\pm}(\mathbf{k}) = \frac{1}{2}\left(I \pm \frac{1}{\epsilon(k)}h(\mathbf{k})\right). \tag{11.79}$$

In the parameterization in Eq. (11.9), we have

$$\begin{aligned} \theta(\mathbf{k}) &= \cos^{-1}\left(\frac{\Delta}{\epsilon(\mathbf{k})}\right), && (11.80)\\ \phi(\mathbf{k}) &= \tan^{-1}\left(\frac{\Pi^y(\mathbf{k})}{\Pi^x(\mathbf{k})}\right). && (11.81) \end{aligned}$$

Fig. (11.5) shows the energy surfaces for three values of Δ the gap between the two bands decreases with Δ and touch at the so called Dirac points at $\Delta = 0$. The lower panel shows the positive energy eigenstates on a regular 100×100 grid on the Brillouin zone plotted on the Bloch sphere. At large Δ all the states map on to a small region around the north pole. The region expands with decreasing Δ. For small Δ most of the points map into a small band around the equator. At $\Delta = 0$, it can be analytically seen that all the states collapse into the equator.

The Haldane model

The Haldane model [3] is a model that was invented by Haldane to show that the quantum Hall effect could be induced by the band structure even in absence of an external magnetic field. It is defined on the honeycomb lattice. The Hamiltonian includes next nearest neighbour hopping,

$$\begin{aligned} \hat{h} &= \sum_I \Delta\left(|\mathbf{R}_{IA}\rangle\langle\mathbf{R}_{IA}| - \mathbf{R}_{IA}\rangle\langle\mathbf{R}_{IA}|\right) \\ &\quad + \sum_{\langle I\alpha, J\beta\rangle} \chi_\alpha\left(|\mathbf{R}_{I\alpha}\rangle\langle\mathbf{R}_{J\beta}| + |\mathbf{R}_{J\beta}\rangle\langle\mathbf{R}_{I\alpha}|\right) \\ &\quad + \frac{i\Delta'}{3\sqrt{3}} \sum_{\langle\langle I\alpha, J\beta\rangle\rangle} \left(|\mathbf{R}_{I\alpha}\rangle\langle\mathbf{R}_{J\beta}| - |\mathbf{R}_{J\beta}\rangle\langle\mathbf{R}_{I\alpha}|\right). \end{aligned} \tag{11.82}$$

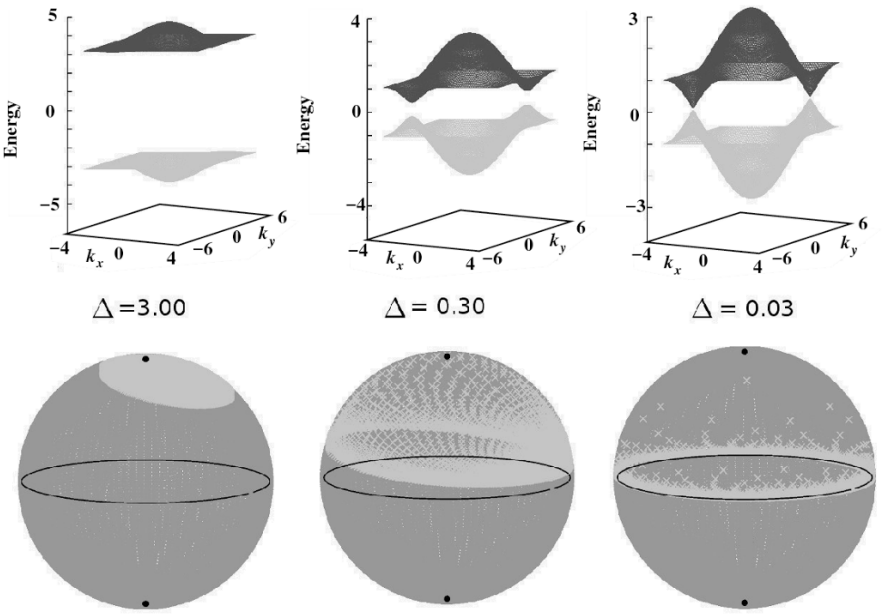


Figure 11.5: The energy surfaces and the surface of states on the Bloch sphere for the honeycomb lattice model at three values of Δ. The light grey crosses on the dark grey Bloch sphere are the eigenstates of the positive energy band computed on a regular 100×100 grid on the Brillouin zone.

where the symbol $\langle\langle \ldots \rangle\rangle$ denotes that the sum is over the second nearest neighbours. $\chi_A = 1$ and $\chi_B = -1$. The Fourier transform of the hopping matrix in Eq. (11.82) is,

$$h(\mathbf{k}) = \boldsymbol{\alpha} \cdot \mathbf{\Pi}(\mathbf{k}) + \beta \tilde{\Delta}(\mathbf{k}), \tag{11.83}$$

$$\tilde{\Delta}(\mathbf{k}) = \Delta - \frac{2\Delta'}{3\sqrt{3}} \left(\sin k_1 + \sin k_2 + \sin k_3\right), \tag{11.84}$$

where $k_3 \equiv -k_1 - k_2$ and all the other quantities are as defined in Eqs. (11.62,11.63,11.75-11.77).

The energy surfaces and the surface of states of the positive energy band is shown in Fig. 11.6. The plots are at $\Delta = 0.3$ and $\Delta' = 0.02,\ 0.2$ and 0.5. At $\Delta' = 0$, the model reduces to the honeycomb lattice model discussed in the previous section. As Δ' increases from zero, the gap between the two bands decreases at one point. The states continue to map on to the northern hemisphere alone. At $\Delta' = \Delta$, the gap closes, namely the two energy surfaces touch at a point. As we will discuss in detail later, the map to the Bloch sphere is

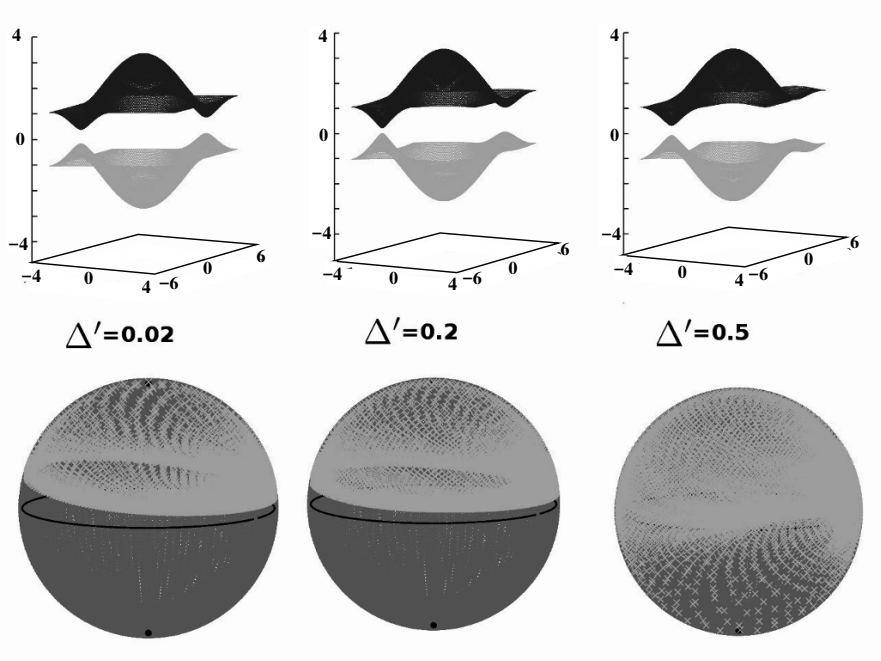


Figure 11.6: The energy surfaces and the surface of states on the Bloch sphere for the Haldane model with $\Delta = 0.3$ and three values of Δ'. The light grey crosses on the dark grey Bloch sphere are the eigenstates of the positive energy band computed on a regular 100×100 grid on the Brillouin zone. The axes in the top figures are same as in Fig. 11.5.

not well defined at this point. At $\Delta' > \Delta$ the gap opens up again but now the map covers the entire Bloch sphere.

11.4.4 Quantum geometry of the spectral bands

To summarize our discussion of tight binding models so far, the spectrum breaks up into bands. The energy eigenvalues of each band defines a surface in the energy-(quasi) momentum space. The eigenstates define mapping of the Brillouin zone to the space of physical states of an N_B-level system, namely $\mathbb{C}P_{N_B-1}$. Thus each band corresponds to a surface in $\mathbb{C}P_{N_B-1}$. We call this the image of the Brillouin zone.

The quantum geometry discussed in section 11.2 can be applied to characterize the "shape" of this surface. Recall, that in section 11.2 we were considering a subspace of the space of physical states parameterized by some parameters, ξ_a. The surface in the physical space of states defined by the each band is a such a subspace parameterized by the quasi-momenta, $\mathbf{k}$. Thus, as in Eq.

(11.24) we can define a metric for the n^{th} band,

$$g^n_{ij}(\mathbf{k}) \equiv \frac{\lambda}{4}\text{tr}\partial_i\rho_n(\mathbf{k})\partial_j\rho_n(\mathbf{k}). \tag{11.85}$$

This metric gives the distance between the eigenstates of the n^{th} band at $\mathbf{k}$ and $\mathbf{k}+d\mathbf{k}$ to be,

$$ds^2(\mathbf{k},\mathbf{k}+d\mathbf{k}) = g^n_{ij}(\mathbf{k})dk^i dk^j. \tag{11.86}$$

We will discuss the physical manifestation of the metric in the next section.

We define the Pancharatnam-Berry curvature (PBC) for each band as in Eq. (11.28),

$$\mathcal{F}^n_{ij}(\mathbf{k}) = \frac{1}{2i}\text{tr}\left(\rho_n(\mathbf{k})\left[\partial_i\rho_n(\mathbf{k}),\partial_j\rho_n(\mathbf{k})\right]\right). \tag{11.87}$$

The integral of $\mathcal{F}^n$ over any surface, Σ, in the Brillouin zone is the geometric phase picked up when the eigenstate of the n^{th} band is transported around the boundary of Σ.

The Chern invariant

The integral of the PBC of a band over any closed two dimensional surface in the Brillouin zone is a topological invariant, in the sense that it is insensitive to smooth changes in the surface. Thus for two dimensional systems, the integral of the the PBC over the full Brillouin zone is a topological invariant that characterizes the band. This invariant, called the Chern invariant or the Chern number, was identified with the Hall conductivity of the system in the path breaking paper by Thouless, Kohmoto, Nightingale and den Nijs (TKNN) [4].

To see why the integral of the PBC over a closed two dimensional surface is insensitive to smooth changes of the surface, recall that as discussed in section 11.3, the integral of the PBC over a surface is the phase of the Bargman invariant constructed along the boundary of that surface. Therefore, it's value is the same for any other surface with the same boundary. Note that since it is a phase, this equality is up to an additive factor of $2n\pi$, where n is an integer. More precisely, if we have two surfaces, Σ' and Σ'' with the same boundary, C, then we must have,

$$\int_{\Sigma'}\mathcal{F}_{ij}dk^i\wedge dk^j = \int_{\Sigma''}\mathcal{F}_{ij}dk^i\wedge dk^j \;+\; 2n\pi. \tag{11.88}$$

The union of Σ' and Σ'' is a closed surface which we call Σ, as schematically depicted in figure 11.7. Now consider shrinking C to a point. In the limit, we have $\Sigma' = \Sigma$ and the integral over Σ'' in Eq. (11.88) vanishes. We then have,

$$\int_{\Sigma}\mathcal{F}_{ij}dk^i\wedge dk^j = 2n\pi. \tag{11.89}$$

Thus, the value of the PBC integrated over any closed surface in the Brillouin zone has to be an integral multiple of 2π. Note that this argument does not

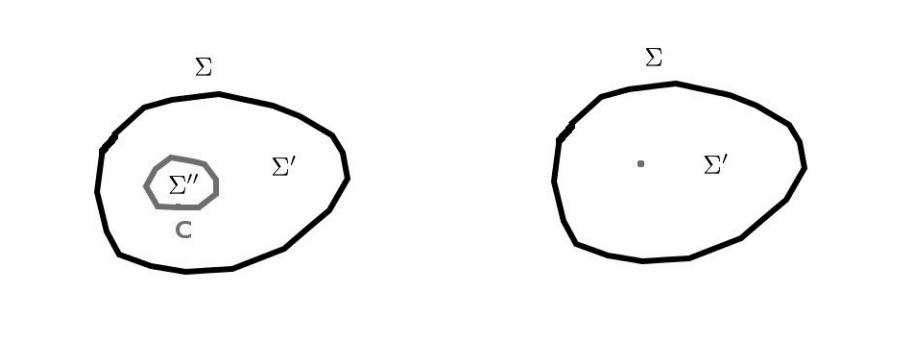


Figure 11.7: The two surfaces Σ' and Σ'' sharing the boundary C are shown in the left. Their union is the closed surface Σ. C has been shrunk to a point in the sketch on the right.

imply that n is non-zero. What it implies that if the integral does not vanish, then it has to be an integral multiple of 2π. Further, since the integral depends only on C and not on Σ', in the limit of C vanishing we expect it to be invariant under smooth changes of Σ.

It can be explicitly verified that there are systems where the Chern invariant is non-zero. With a little patience, the expression for the PBC of the Haldane model can be analytically computed using Eq. (11.87). Figure 11.8 shows the result of the calculation for $\Delta = 0.3$ and $\Delta' = 0.0, 0.2$ and 0.4. At $\Delta' = 0$, the PBC is anti-symmetric under reflections and the PBC integrates to zero. It is peaked around two points on the Brillouin zone, normally referred to as the K and K' points. As Δ' increases, the reflection symmetry is broken, the positive peak becomes higher and narrower and the negative peak becomes lower but broader, as shown for $\Delta' = 0.2$ in Fig. 11.8. For $\Delta' < \Delta$, the PBC continues to integrate to zero. When $\Delta' > \Delta$, the high and narrow positive peak becomes a high and narrow negative peak, as shown in Fig. 11.8 for $\Delta' = 0.4$. The Chern invariant in this regime is -2π.

11.4.5 Dirac points and topological transitions

We have seen in the previous section that the Haldane model exhibits a topological transition at $\Delta' = \Delta$, in the sense that the Chern invariant changes from zero at $\Delta' < \Delta$ to -2π at $\Delta' > \Delta$. In this section, we will examine the model more closely in the vicinity of the transition. We concentrate on the $\Delta > 0$ regime. The other regime is similar.

As can be seen from figure 11.8, the action occurs near the $\mathbf{K}$ and $\mathbf{K}'$ points. These are the points with $(k_1, k_2) = \pm\left(\frac{2\pi}{3}, \frac{2\pi}{3}\right)$. At these points, we have $\mathbf{\Pi}(\mathbf{k}) = 0$.

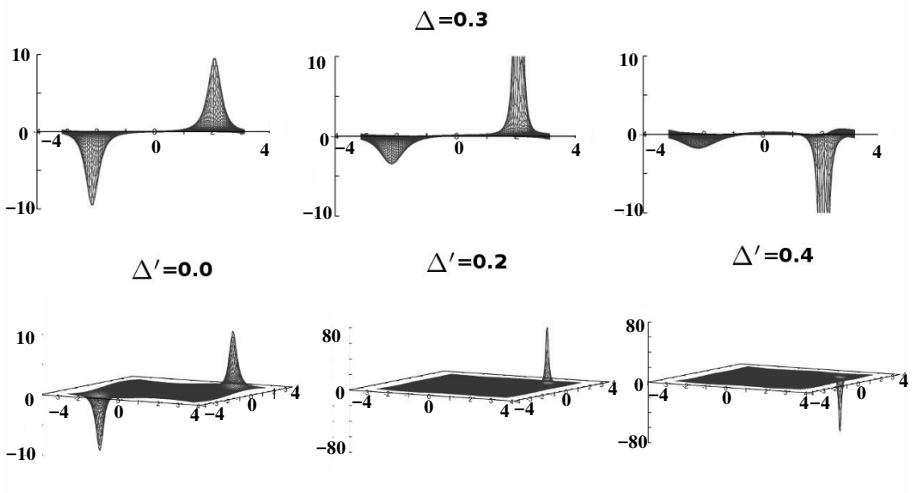


Figure 11.8: The PBC of the Haldane model, at $\Delta = 0.3$ and the indicated values of Δ', plotted as a function of k_1 and k_2. Note that the scale of the z-axis is much larger in the two bottom right plots.

When $\Delta > \Delta'$ the states of the positive energy band at $\mathbf{K}$ and $\mathbf{K}'$ map on to the north pole of the Bloch sphere and the states of the negative energy band at these points map on to the south pole. When $\Delta < \Delta'$, the situation is the same at $\mathbf{K}'$. However at $\mathbf{K}$, the situation is reversed since the sign of δ changes. The state of the positive energy band now maps to the south pole and the state of the negative energy band maps to the north pole. At $\Delta = \Delta'$, we have $h(\mathbf{K}) = 0$. The two eigenvalues are degenerate and any state is the eigenvector. Thus, there is an ambiguity is which state should be assigned to the positive energy band and which one to the negative energy band. This ambiguity manifests as a singularity in the PBC at $\mathbf{K}$. Consequently, the Chern invariant is ill defined at the transition point $\Delta = \Delta'$.

To see this explicitly, we compute the PBC in the vicinity of $\mathbf{K}$. we put $\mathbf{k} = \mathbf{K} + \mathbf{q}$ and expand to linear order in $\mathbf{q}$. We have,

$$\Pi_x(\mathbf{K}+\mathbf{q}) = -\frac{\sqrt{3}}{2}(q_1+q_2), \tag{11.90}$$

$$\Pi_y(\mathbf{K}+\mathbf{q}) = -\frac{1}{2}(q_1-q_2), \tag{11.91}$$

$$\tilde{\Delta}(\mathbf{K}+\mathbf{q}) = \Delta - \Delta', \tag{11.92}$$

where $q_{1(2)} = \mathbf{e}_{1(2)} \cdot \mathbf{q}$. So we choose a coordinate system, $\hat{x} \equiv -(\mathbf{e}_1 + \mathbf{e}_2)$, $\hat{y} = -\frac{1}{\sqrt{3}}(\mathbf{e}_1 - \mathbf{e}_2)$ and write,

$$h_K(\mathbf{q}) = \frac{\sqrt{3}}{2}\boldsymbol{\alpha} \cdot \mathbf{q} + \beta\delta, \tag{11.93}$$

where $\delta \equiv \Delta - \Delta'$. The PBC is easy to compute using Eq. (11.87). The answer is

$$\mathcal{F}_{ij}(\mathbf{K}+\mathbf{q}) = \frac{3}{2}\frac{\delta}{\left(\frac{3}{4}q^2+\delta^2\right)^{\frac{3}{2}}}\epsilon_{ij}. \tag{11.94}$$

The RHS of the above equation is ill defined when $\delta, q^2 \to 0$. If the $\delta \to 0$ limit is taken first, it is zero. On the other hand, if the $q^2 \to 0$ limit is taken first, it diverges.

Thus, at the transition point, the energy bands touch at the so called Dirac points. This degeneracy implies an ambiguity of the mapping of the Dirac points to the Bloch sphere. Consequently, the PBC and hence the Chern number are ill-defined at the transition point. This is why the Chern number can change discontinuously in the transition.

11.5 Physical manifestation

As mentioned earlier, the identification of the Chern number with the Hall conductance of two dimensional systems [4] initiated the interest in exploring the quantum geometry of quantum Hall systems. The PBC was shown to physically manifest itself as an anomalous velocity [5–7], which results in a component of the current perpendicular to the applied electric field, the Hall current. The role of the quantum metric in fractional quantum Hall systems has been explored recently [8].

The Haldane model [3] showed that the band structure could induce a PBC leading to a quantum Hall state even in absence of a magnetic field. These systems are called Chern insulators. The quantum metric was related to the so called localization tensor leading to a characterization of the ground states of insulators [10].

In this section, we will discuss these physical effects, namely the relation of the PBC to the Hall current and the metric to the localization tensor.

11.5.1 Dynamics constrained to a Band

The band gaps in most materials tend to be rather high ($\sim eV$) compared to typical temperatures ($\sim meV$) and the energy that external fields can impart to the electrons. In such cases, to a good approximation, the dynamics is constrained to remain in a band. In other words, to a good approximation, we can project to the subspace of states of the form,

$$|\psi_n\rangle = \int_{BZ} \psi_n(\mathbf{k})|n\mathbf{k}\rangle. \tag{11.95}$$

The projection operator into this band is,

$$\hat{P}_n = \int_{BZ} \rho_n(\mathbf{k}). \tag{11.96}$$

The semiclassical dynamics of wave packets in such a system was in the presence of slowly varying external fields, a scalar potential, $V(\mathbf{R})$, and a magnetic field, $F_{ij}(\mathbf{R})$, was analyses by Sundaram and Niu [6]. They derived the following equations the central position, $\mathbf{R}$, and momenta $\hbar\mathbf{k}$ of the wavepackets,

$$\frac{dR_i}{dt} = \frac{1}{\hbar}\frac{\partial\epsilon^n}{\partial k_i} + \mathcal{F}_{ij}(\mathbf{k})\frac{dk_j}{dt}, \tag{11.97}$$

$$\hbar\frac{dk_i}{dt} = -e\frac{\partial V}{\partial R_i} + eF_{ij}\frac{dR_j}{dt}. \tag{11.98}$$

These equations, beautifully symmetric between $\mathbf{R}$ and $\mathbf{k}$, motivate the interpretation of the PBC as a magnetic field in momentum space. The second term in Eq. (11.97) is called the anomalous velocity. The Lorentz force gives a velocity dependent acceleration to the particle, analogously the anomalous velocity is an acceleration dependent velocity.

If we turn off the external magnetic field, i.e. set $F_{ij} = 0$, and substitute Eq. (11.98) in 11.97), we get

$$\frac{dR_i}{dt} = \frac{1}{\hbar}\frac{\partial\epsilon^n}{\partial k_i} - \frac{e}{\hbar}\mathcal{F}_{ij}(\mathbf{k})E_j. \tag{11.99}$$

Thus we see that the PBC provides a component of the velocity that is perpendicular to the electric field $\mathbf{E} = -\boldsymbol{\nabla}V$.

The analogy can be extended. In the presence of a magnetic field, the components of the velocities do not commute. The velocity operators are,

$$\hat{\mathbf{v}} = \frac{1}{m}\left(\hat{\mathbf{p}} - e\mathbf{A}(\hat{\mathbf{R}})\right). \tag{11.100}$$

This implies,

$$[\hat{v}_i, \hat{v}_j] = \frac{i\hbar e}{m^2}F_{ij}. \tag{11.101}$$

The Heisenberg equation of motion for the velocity gives the Lorentz force,

$$\frac{d\hat{v}_i}{dt} = \frac{1}{i\hbar}\left[\hat{v}_i, \frac{m}{2}\hat{\mathbf{v}}\cdot\hat{\mathbf{v}}\right],$$

$$\Rightarrow m\frac{d\hat{v}_i}{dt} = eF_{ij}\hat{v}_j. \tag{11.102}$$

In the next two subsections, we will discuss the analogy in the momentum space. We will first show that the component of the position operators, when projected to a band, have anomalous commutators. These lead to the anomalous velocity.

The anomalous commutators

We first analyze the position operator, projected to the n^{th} band. The position operator is,

$$\hat{\mathbf{R}} = \sum_{I\alpha} |I\alpha\rangle\mathbf{R}_{I\alpha}\langle I\alpha|. \tag{11.103}$$

We compute the matrix elements of the components of $\hat{\mathbf{R}}_{I\alpha}$ between two states in the n^{th} band,

$$\langle\chi_n|\hat{R}_i|\psi_n\rangle = \int_{k',k\in BZ}\sum_{I\alpha}\chi_n^*(\mathbf{k}')\langle n\mathbf{k}'|I\alpha\rangle R_{I\alpha i}\langle I\alpha|n\mathbf{k}\rangle\psi_n(\mathbf{k}) \quad (11.104)$$

We now use the facts,

$$\langle I\alpha|n\mathbf{k}\rangle = u_\alpha^n(\mathbf{k})e^{i\mathbf{k}\cdot\mathbf{R}_{I\alpha}}, \quad (11.105)$$

$$\sum_I e^{i(\mathbf{k}-\mathbf{k}')\cdot\mathbf{R}_{I\alpha}} = (2\pi)^d\delta^d(\mathbf{k}-\mathbf{k}'), \quad (11.106)$$

to obtain,

$$\begin{aligned}\langle\chi_n|\hat{R}_i|\psi_n\rangle &= \int_{k',k\in BZ}\chi_n^*(\mathbf{k}')u_\alpha^{n\,*}(\mathbf{k}')\left(-i\frac{\partial}{\partial k_i}\delta^d(\mathbf{k}-\mathbf{k}')\right)u_\alpha^n(\mathbf{k})\psi_n(\mathbf{k})\\ &= \int_{k\in BZ}\chi_n^*(\mathbf{k})\left(i\frac{\partial}{\partial k_i}-\mathcal{A}_i^n(\mathbf{k})\right)\psi_n(\mathbf{k}), \quad (11.107)\\ \mathcal{A}_i^n(\mathbf{k}) &\equiv \frac{1}{i}u^{n\dagger}(\mathbf{k})\frac{\partial u^n(\mathbf{k})}{\partial k_i}. \quad (11.108)\end{aligned}$$

Thus, when projected to the n^{th} band, the position operator is represented by,

$$\hat{\mathbf{R}}_n = i\boldsymbol{\nabla}_k - \boldsymbol{\mathcal{A}}(\mathbf{k}). \quad (11.109)$$

It then follows that the components of the position operator do not commute and that the commutator is exactly the PBC, viz.,

$$\left[\hat{R}_{ni},\hat{R}_{nj}\right] = i\mathcal{F}_{ij}^n(\mathbf{k}). \quad (11.110)$$

In the full Hilbert space, the components of the position operator of course commute. The anomalous commutator is a consequence of the projection into a band. This is not uncommon. Eg., consider the two 3×3 matrices,

$$A = \begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix}, \text{ and } B = \begin{pmatrix}0&0&1\\1&0&0\\0&1&0\end{pmatrix}. \quad (11.111)$$

It is easy to check that $[A,B]=0$. In fact $B=A^2$. If these matrices are projected to the subspace consisting of the top two components,

$$A' = \begin{pmatrix}0&1\\0&0\end{pmatrix}, \text{ and } B' = \begin{pmatrix}0&0\\1&0\end{pmatrix}. \quad (11.112)$$

We then have,

$$[A',B'] = \begin{pmatrix}1&0\\0&-1\end{pmatrix} \neq 0. \quad (11.113)$$

The anomalous velocity

We consider the system in the presence of slowly varying external potential $V(\mathbf{R}_{I\alpha})$. The Hamiltonian, projected into the n^{th} band is,

$$\hat{h}_n = \hat{\epsilon}_n + e\hat{V}_n, \tag{11.114}$$

$$\hat{\epsilon}_n \equiv \int_{k\in BZ} |nk\rangle \epsilon^n(k) \langle n\mathbf{k}|, \tag{11.115}$$

$$\hat{V}_n \equiv V(\hat{\mathbf{R}}_n). \tag{11.116}$$

The Heisenberg equation for the time evolution of the position operator is,

$$\frac{d}{dt}\hat{R}_{ni} = \frac{1}{i\hbar}\left[\hat{R}_{ni}, \hat{h}_n\right]. \tag{11.117}$$

It is easy to show that,

$$i\hbar\left[\hat{R}_{ni}, \hat{\epsilon}_n\right] = \int_{k\in BZ} |nk\rangle \frac{1}{\hbar}\frac{\partial \epsilon^n(\mathbf{k})}{\partial k_i} \epsilon^n(\mathbf{k}) \langle n\mathbf{k}| \equiv \hat{v}_{ni}. \tag{11.118}$$

It is not possible to give a general expression for the commutator of $\hat{R}_{ni}$ with $V(\hat{\mathbf{R}}_n)$ without knowing the functional form of V. However, in the classical limit we can write,

$$\left[\hat{R}_{ni}, V(\hat{\mathbf{R}}_n)\right] = \frac{\partial V}{\partial R_j}\left[\hat{R}_{ni}, \hat{R}_{nj}\right] + \ldots, \tag{11.119}$$

where the $\ldots$ in Eq. 11.119 denote terms higher order in $\hbar$ when we work in terms of the momenta,$\mathbf{p} = \hbar\mathbf{k}$, instead of the wave vectors. Eg., $\hat{R}_{ni} = i\hbar\frac{\partial}{\partial p_i} - \hbar\mathcal{A}_(\mathbf{p})$, $\mathcal{A}_i(\mathbf{p}) \equiv -iu^{n\dagger}\frac{\partial}{\partial p_i}u^n$. Thus, in the semi-classical limit we have the Heisenberg equation of motion,

$$\frac{d}{dt}\hat{R}_{ni} = \hat{v}_i + e\mathcal{F}^n_{ij}E_j. \tag{11.120}$$

The anomalous commutators give rise to the anomalous velocity.

The localization tensor

Consider the expectation value of the anti-commutator of the position operators in a band eigenstate,

$$\begin{aligned}\langle n\mathbf{k}|\left\{\hat{R}_{ni}, \hat{R}_{nj}\right\}|n\mathbf{k}\rangle &= \left(\left(i\frac{\partial}{\partial k_i} - \mathcal{A}^n_i\right)u^n\right)^\dagger \left(i\frac{\partial}{\partial k_i} - \mathcal{A}^n_i\right)u^n + i \leftrightarrow j \\ &= \partial_i u^{n\dagger}\partial_j u^n - \mathcal{A}^n_i\mathcal{A}^n_j + i \leftrightarrow j. \end{aligned}\tag{11.121}$$

With some algebra, the RHS can be shown to be equal to

$$\partial_i u^{n\dagger}\partial_j u^n - \mathcal{A}^n_i\mathcal{A}^n_j + i \leftrightarrow j = \mathrm{tr}\partial_i\rho_n(\mathbf{k})\,\partial_j\rho_n(\mathbf{k}). \tag{11.122}$$

Thus we have the result,

$$g_{ij}(\mathbf{k}) = \frac{\lambda}{4}\langle n\mathbf{k}| \left\{ \hat{R}_{ni}, \hat{R}_{nj} \right\} |n\mathbf{k}\rangle. \tag{11.123}$$

When the RHS is integrated over the Brillouin zone, it is known as the localization tensor, which is a measure of the spatial spread and shape of the wavefunctions in the band.

11.6 Summary

To summarize our discussion, we started by discussing quantum systems in general. The space of physical states for any quantum system is the space of rays in a Hilbert space, namely a projective Hilbert space. The pure state density matrices are in one-to-one correspondence with the physical states and hence are a good representation of them.

The inner product of the Hilbert space defines a natural geometry on the space of physical states. This geometry is nicely described in terms of the so called Bargmann invariants which are traces of products of the pure state density matrices. A distance between any two physical states and a geometric phase for every closed curve in the space of physical states are defined in terms of the Bargmann invariants. The distance between two infinitesimally separated states defines a metric in the space of physical states and the geometric phase of an infinitesimal loop dines a curvature, the Panchathnam-Berry curvature.

We then applied these ideas to tight binding models. The discrete translation symmetry of these models implies that the quasi-momenta that define the Brillouin zone are conserved. The spectrum of these models breaks up into spectral bands. Each band is defined by a set of energy eigenvalues and eigenvectors at every point on the Brillouin zone. Thus there is an energy surface and a surface of states for every band. The quantum metric and PBC of the surface of states induces a metric and PBC on the Brillouin zone. The PBC, integrated over the Brillouin zone is a topological invariant, the Chern invariant.

A semiclassical analysis of dynamics restricted to a band shows that the PBC can be interpreted as a "magnetic field" in the quasi-momentum space. It induces an acceleration dependent anomalous component of the velocity, analogous to the Lorentz force which induces a velocity dependent component of the acceleration. The anomalous component of the the velocity results in a Hall current for many particle systems that is proportional to the PBC. This leads to a semi-classical understanding of the classic result of TKNN [4] which identified the Hall conductivity of quantum Hall systems with the Chern invariant.

In the presence of a magnetic field, the components of the velocity operators do not commute and the commutator is proportional to the field tensor. Analogously, the position operators projected to a band do not commute and their commutator is proportional to the PBC of that band. Further, the anticommutator of the projected position operators is proportional to the quantum

metric. The integral of the metric over the Brillouin zone, is the second moment of the position operator and is called the localization tensor, which plays an important role in characterizing insulators [10].

Acknowledgement

I am grateful to my colleague, S.R. Hassan, IMSc., for motivating me to get interested in the quantum geometry of condensed matter systems and many long discussions on the topic.

References

[1] N. Mukunda and R. Simon, Annals of Physics **228**, 205-268 (1993).

[2] V. Bargmann, J. Math. Phys. **5**, 862 (1964).

[3] F.D.M Haldane, Phys. Rev. Lett. **61**, 2015 (1988).

[4] D.J. Thouless, M. Kohmoto, M.P. Nightingale and M. den Nijs, Phys. Rev. Lett. **49**, 405 (1982).

[5] R. Karplus and J.M. Luttinger, Phys. Rev. **95**, 1194 (1954).

[6] G. Sundaram and Q. Niu, Phys. Rev. **B 59**, 14915 (1999).

[7] F.D.M. Haldane, Phys. Rev. Lett. 93, 206602 (2004).

[8] F.D.M. Haldane, Phys. Rev. Lett. **107**, 116801 (2011).

[9] N. Mazari and D. Vanderbilt, Phys. Rev. **B 26**, 12847 (1997).

[10] R. Resta, Eur. Phys. J. **B 79**, 121 (2011).

12

Topology, geometry and quantum interference in condensed matter physics

Alexander Abanov

The methods of quantum field theory are widely used in condensed matter physics. In particular, the concept of an effective action was proven useful when studying low temperature and long distance behavior of condensed matter systems. Often the degrees of freedom which appear due to spontaneous symmetry breaking or an emergent gauge symmetry, have non-trivial topology. In those cases, the terms in the effective action describing low energy degrees of freedom can be metric independent (topological). We consider a few examples of topological terms of different types and discuss some of their consequences. We will also discuss the origin of these terms and calculate effective actions for several fermionic models. In this approach, topological terms appear as phases of fermionic determinants and represent quantum anomalies of fermionic models. In addition to the wide use of topological terms in high energy physics, they appeared to be useful in studies of charge and spin density waves, Quantum Hall Effect, spin chains, frustrated magnets, topological insulators and superconductors, and some models of high-temperature superconductivity.

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_12

12.1 Introductory remarks

12.1.1 Theory of Everything in condensed matter physics

In condensed matter physics, we believe that we know the "Theory of Everything" – the fundamental equations potentially describing all observable phenomena in condensed matter physics. Essentially, those equations are Schrödinger equations for electrons and nuclei, together with the Maxwell equations describing electromagnetic interactions. [1] However, there is a long way from knowing fundamental equations and being able to actually describe collective behavior of 10^{20} or so nuclei and electrons forming condensed matter systems, like, liquids, solids, superfluids, superconductors, quantum Hall systems, etc. Having "more" particles makes macroscopic systems behave very differently from collections of just a few particles. New qualitative features appear when one goes from microscopic to macroscopic systems. [2] We refer to new phenomena appearing at macroscopic scales as to "emergent phenomena".

The goal of condensed matter physics is not finding fundamental laws but rather finding their consequences. In particular, we are interested in finding efficient ways to describe emergent macroscopic phenomena. While it is very hard to derive macroscopic phenomena by solving fundamental equations we have few guiding principles that allow us to write effective descriptions of those phenomena. Such principles include the use of symmetries and associated conservation laws, mechanisms of spontaneous symmetry breaking, the concept of quasiparticles etc. Early examples of effective descriptions include thermodynamics and hydrodynamics.

In these lectures, we focus on the "topological properties" of condensed matter systems and their descriptions. Topological properties, in general, are the properties robust with respect to continuous deformations. They are emergent properties and it is important to understand them in the context of condensed matter physics as they might be the most stable properties insensitive to deformations and perturbations always present in realistic materials. The main focus of these lectures will be on topological properties related to quantum physics.

12.1.2 Spontaneous symmetry breaking and an emergent topology

If there were no separation of scales in Nature, the task of theoretical physicists would be formidable. Fortunately, in many cases one can "integrate out" fast degrees of freedom and effectively describe properties of microscopic systems at low temperatures, low frequencies, and large distances using relatively simple continuous field theory descriptions. This happens due to the presence of exact or approximate symmetries in the underlying microscopic system. More precisely, it is due to a phenomenon of spontaneous symmetry breaking.

Suppose that the exact Hamiltonian of some condensed matter system has some continuous symmetry given by Lie group G. A good system to keep

in mind as an example is an isotropic ferro- or antiferromagnet with an $SU(2)$ symmetry with respect to global rotations of all spins. Then it is possible that at some values of parameters of the Hamiltonian,[1] the ground state of the system breaks the symmetry up to some subgroup H of G. If this happens, we say that the symmetry of the Hamiltonian is spontaneously broken by its ground state. One can characterize this ground state by some element n of a coset space G/H. In our example of the magnet we take $H = SO(2) = U(1)$ and $G/H = SU(2)/U(1) = S^2$. The element of a coset space in this case is a point of two-dimensional sphere S^2 which labels the direction of magnetization of our system and subgroup H is just the group of all rotations around the direction of magnetization which is obviously a symmetry of the Hamiltonian and of the ground state of the system. In the presence of spontaneous breaking of a continuous symmetry the ground state is infinitely degenerate, since any $n \in G/H$ gives the ground state with the same energy. Indeed, any two states characterized by $n_1, n_2 \in G/H$ have the same energy since they can be transformed into one another by some element $g \in G$, which is an exact symmetry transformation of the Hamiltonian.

Now consider another state of the quantum system which is locally in the vicinity of spatial point x, and is very close to the ground state of the system labeled by some $n(x) \in G/H$. We assume that $n(x)$ is not constant in space but changes very slowly with a typical wavenumber k. In such a case, $n(x)$ is called an order parameter of the system. We denote the energy of this state per unit volume measured from the energy of the ground state $\epsilon(k)$. The limit of small $k \to 0$ corresponds to the order parameter which is constant in space $n(x) = n_0$ and, therefore, $\epsilon(k) \to 0$ as $k \to 0$. We obtain that when continuous symmetry is spontaneously broken, the ground state of the system is not isolated but there are always excited states whose energies are infinitesimally close to the ground state energy. These heuristic arguments can be made more rigorous and lead to the Goldstone theorem[2]. The theorem states that in quantum field theory with spontaneously broken continuous symmetry there are massless particles which energy $\epsilon(k) \to 0$ as $k \to 0$.

If one is interested in low energy physics one necessarily should take these massless modes (or Goldstone bosons) into account. Moreover, the nature of these massless modes is dictated essentially by the symmetry (and its breaking) of the system, and one expects, therefore, that the correct low energy description should depend only on symmetries of the system but not on its every microscopic detail.

[1] We consider here the case of zero temperature for simplicity.

[2] For a full formulation of the Goldstone theorem for relativistic field theory as well as for its proof see e.g., Ref. [3]. We avoid it here because we are generally interested in a wider range of systems, e.g., without Lorentz invariance. There are still some analogs of Goldstone theorem there. For example in the case of a ferromagnet there are still massless particles – magnons. However, the number of independent massless particles is not correctly given by the Goldstone theorem for relativistic systems.

A natural variable describing the dynamics of Goldstone modes is the order parameter itself. For example, in the case of relativistically invariant system described by the order parameter $\mathbf{n} \in S^2$, we immediately write

$$S_{\rm NLSM} = \int d^{d+1}x \, \frac{1}{2g} (\partial_\mu \mathbf{n})^2 + (\text{other terms}). \tag{12.1}$$

We have written here the most obvious term of an effective action which is both Lorentz invariant and $SU(2)$ invariant (with respect to rotations of a unit, three-component vector $\mathbf{n}^2 = 1$). Here g is a coupling constant which should be obtained from a detailed microscopic theory. The "other terms" are the terms which are higher order in gradients and (possibly) topological terms. The model Eq. (12.1) is referred to as a "non-linear σ-model"[3]. In different spatial dimensions, higher gradient terms of non-linear σ-models might be relevant. We are not discussing those terms as well as the issue of renormalizability of σ-models, concentrating instead on the allowed topological terms. Therefore, we will keep only the kinetic term $\frac{1}{2g}(\partial_\mu \mathbf{n})^2$ in the gradient expansion of an effective Lagrangian as well as all allowed topological terms.

Before proceeding to our main subject – topological terms, let us make two important remarks. Firstly, very often (especially in condensed matter systems) the symmetries of the Hamiltonian are approximate and there are terms in the Hamiltonian which explicitly but weakly break the symmetry. This does not invalidate the speculations of this section. The difference will be that would-be-Goldstone particles acquire small mass. The weaker the explicit symmetry breaking of the Hamiltonian the smaller is the mass of "Goldstone" particles. One can proceed with the derivation of the non-linear σ-model which will contain weak symmetry breaking terms (such as easy-axis anisotropy for magnets). This model will have non-trivial dynamics at energies bigger than the smallest of masses.

Second remark is that there are other mechanisms in addition to the spontaneous symmetry breaking which result in low energy excitations. One of the most important mechanisms is realized when local (or gauge) symmetry is present. Then, gauge invariance plus locality demands the presence of massless particles (e.g, photons) in the system. The low energy theories in this case are gauge theories. Similar to an explicit symmetry breaking, in case of Goldstone particles there are mechanisms which generate masses for gauge bosons. These are, e.g., Higgs mechanism and confinement of gauge fields. We will have some examples of topological terms made out of gauge fields in these lectures although our main focus will be on non-linear sigma models[4]. We also do not consider here cases with massless fermionic degrees of freedom we concentrate exclusively on bosonic effective theories.

[3]The origin of the term is in effective theories of weak interactions [4,5]. Non-linear comes from the non-linear realization of symmetries in this model. E.g., constraint $\mathbf{n}^2 = 1$ is non-linear. Sigma (σ) is a historic notation for the "order parameter" in theories of weak interactions.

[4]Non-linear sigma models and gauge theories have a lot in common [6].

12.1.3 Additional reading

The focus of these lectures is on the effect of topological terms in the action on physical properties of condensed matter systems. We are not discussing here classification of topological defects in textures in ordered media. The latter is a well developed subject (see the classical review [7]). For reader's convenience we collected a few exercises on topological textures and related examples in Appendix A.

The subject of topological terms or broader "topological phases of matter" is huge, and we do not do it justice in these lectures. In particular, I do not try to give a complete bibliography in these lectures. Instead, with a few exceptions I refer not to the original papers but to textbooks or reviews.

The subject of homotopy classification of topological defects and textures in ordered media is not discussed in this lectures. However, it is a necessary prerequisite to understanding topological terms discussed here. I recommend a classic reference [7]. To make this text more self-contained I also collected few exercises on that topic in Appendix A, groups in Appendix C.

I would recommend the following textbooks close in spirit to the point of view presented here [6, 8, 9]. Some of the technical details of fermionic determinant calculations can be found in [10, 11]. Topological terms are intimately related to geometric or Berry phases [12] and to quantum anomalies in field theories [13].

In these lectures, I avoid using any advanced topological and geometrical tools. However, I highly recommend studying all necessary mathematics seriously. See, for example, other chapters of this book. In addition, there are many beautiful books that give good introduction to the subject for physicists. See, for example Refs. [14–18].

These notes are based on the lectures given by the author at "SERC School on Topology and Condensed Matter Physics" in Kolkata, India in December 2015.

12.2 Motivating example: a particle on a ring.

12.2.1 Classical particle on a ring: Action, Lagrangian, and Hamiltonian

As a simple motivating example let us consider a particle on a ring. Classically, the motion can be described by the principle of least action. A classical action S of a particle can be taken as

$$S[\phi] = \int dt\, L(\phi, \dot{\phi}), \tag{12.2}$$

$$L = \frac{M}{2}\dot{\phi}^2 + A\dot{\phi}, \tag{12.3}$$

should be minimal (locally) on classical trajectories. Here, the angle $\phi(t)$ is chosen to be a generalized coordinate of the particle on a ring, M is a moment of inertia of a particle (or mass for a unit ring), A is some constant.

Euler-Langrange equations of motion are given in terms of Lagrangian L by $\frac{d}{dt}\frac{\partial L}{\partial \dot{\phi}} - \frac{\partial L}{\partial \phi} = 0$, or explicitly

$$M\ddot{\phi} = 0. \tag{12.4}$$

A particle with a given an initial velocity moves with a constant angular velocity along the ring. Notice, that the last term of Eq. (12.3) does not have any effect on the motion of the particle. Indeed, this term is a total time derivative and can not affect the principle of least action [19].

Given an initial position of a particle on a ring at $t = t_1$ and a final position at $t = t_2$ there are infinitely many solutions of Eq. (12.4). They can be labeled by the integer number of times the particle goes around the ring to reach its final position. This happens because of the *nontrivial topology* of the ring – one should identify $\phi = \phi + 2\pi$ as labeling the same point on the ring. This is not very important classically as we can safely think of the angle ϕ taking all real values from $-\infty$ to $+\infty$. Given initial position $\phi(t_1) = \phi_1$ and initial velocity $\dot{\phi}(t_1) = \omega_1$ one can unambiguously determine the position of the particle $\phi(t)$ at all future times using Eq. (12.4).

Let us now introduce the momentum conjugated to ϕ as

$$p = \frac{\partial L}{\partial \dot{\phi}} = M\dot{\phi} + A, \tag{12.5}$$

and the Hamiltonian as

$$H = p\dot{\phi} - L = \frac{1}{2M}(p - A)^2. \tag{12.6}$$

Corresponding Hamilton equations of motion

$$\dot{\phi} = \frac{1}{M}(p - A), \tag{12.7}$$

$$\dot{p} = 0, \tag{12.8}$$

are equivalent to Eq. (12.4).

Notice that although the parameter A explicitly enters the Hamiltonian formalism, it only changes the definition of generalized momentum $M\dot{\phi} + A$ instead of more conventional $M\dot{\phi}$. It does not change the solution of equations of motion and can be removed by a simple canonical transformation $p \to p + A$. We will see below that this changes for a *quantum* particle.

12.2.2 Quantum particle on a ring: Hamiltonian and spectrum

Let us now consider a quantum particle on a ring. We replace classical Poisson's bracket $\{p, \phi\} = 1$ by quantum commutator $[p, \phi] = -i\hbar$ and use ϕ-representation, i.e., we describe our states by wave functions on a ring $\psi(\phi)$. In

the following, we will put $\hbar = 1$. In this representation, we can use $p = -i\partial_\phi$ and rewrite Eq. (12.6) as a quantum Hamiltonian

$$H = \frac{1}{2M}\left(-i\partial_\phi - A\right)^2. \tag{12.9}$$

The eigenstates and eigenvalues of this Hamiltonian are given by solutions of stationary Schrödinger equation $H\psi = E\psi$. We impose *periodic boundary conditions* requiring $\psi(\phi + 2\pi) = \psi(\phi)$, i.e., the wave function is required to be a single-valued function on the ring. The eigenfunctions and eigenvalues of Eq. (12.9) are given by

$$\psi_m = e^{im\phi}, \tag{12.10}$$

$$E_m = \frac{1}{2M}(m - A)^2, \tag{12.11}$$

where $m = 0, \pm 1, \pm 2, \ldots$ is any integer number — the quantized eigenvalue of the momentum operator $p = -i\partial_\phi$. We notice that although the classical model is not sensitive to parameter A, the quantum one is, because of the quantization of p. The parameter A can be interpreted as a vector potential of the magnetic flux penetrating the ring. This vector potential is not observable in classical mechanics but affects the quantum spectrum because of multiple-connectedness of the ring (there are many non-equivalent ways to propagate from the point 1 to the point 2 on a ring). More precisely our parameter A should be identified with the vector potential multiplied by $\frac{e}{\hbar c}$. It corresponds to the magnetic flux through the ring $\Phi = A\Phi_0$, where Φ_0 is a flux quantum $\Phi_0 = 2\pi\frac{\hbar c}{e}$.

The A-term of the classical action – *topological term* – can be written as

$$S_{top} = \int_{t_1}^{t_2} dt\, A\dot{\phi} = 2\pi A\frac{\phi_2 - \phi_1}{2\pi} = \theta\frac{\Delta\phi}{2\pi}. \tag{12.12}$$

It depends only on the initial and final values $\phi_{1,2} = \phi(t_{1,2})$, and changes by $\theta = 2\pi A$ every time the particle goes a full circle around the ring in counterclockwise direction. The conventional notation θ for a coefficient in front of this term gave the name *topological theta-term* for this type of topological terms.

The spectrum Eq. (12.11) is shown in Figure 12.1 for three values of flux through the ring $\theta = 0, \pi, \pi/2$ ($A = \Phi/\Phi_0 = 0, 1, 1/2$).

Several comments are in order. (i) An integer flux A-integer or θ - multiple of 2π does not affect the spectrum. (ii) There is an additional symmetry (parity) of the spectrum at θ multiples of π (integer or half-integer flux). (iii) For half-integer flux $\theta = \pi$, the ground state is doubly degenerate $E_0 = E_1$.

Finally, let us try to remove the A term by canonical transformation as in the classical case. We make a gauge transformation $\psi \to e^{iA\phi}\psi$ and obtain $p \to p + A$ and $H = \frac{1}{2M}(-i\partial_\phi)^2$. One might think that we removed the effects of the A term completely. However, this transformation changes the boundary conditions of the problem replacing them by *twisted boundary conditions* $\psi(\phi +$

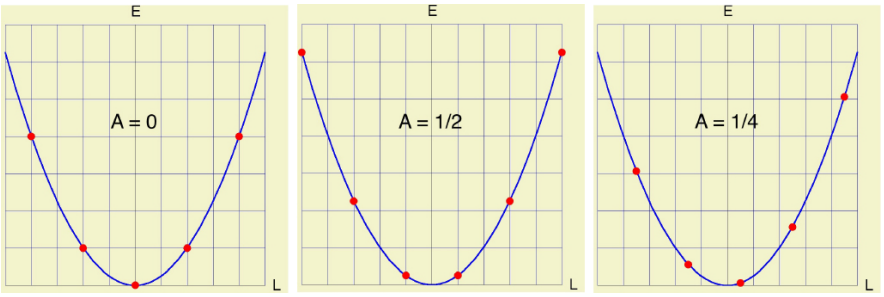


Figure 12.1: The spectrum of the particle on a ring is shown for $A = \theta/2\pi = 0, 1/2, 1/4$ respectively. The classical energy $E(p)$ is represented by a parabola and does not depend on the parameter A.

$2\pi) = e^{-i2\pi A}\psi(\phi)$. The eigenfunctions satisfying twisted boundary conditions are $\psi_m = e^{i(m-A)\phi}$ and produce the same eigenvalues Eq. (12.11). We conclude that it is not possible to remove the effects of topological A-term in quantum mechanics. The parameter A can be formally removed from the Hamiltonian by absorbing it into the boundary conditions. This, however, does not change the spectrum and other physical properties of the system.

12.2.3 Quantum particle on a ring: path integral and Wick's rotation

Quantum mechanics of a particle on a ring described by the classical action Eq. (12.3) can be represented by path integral

$$Z = \int D\phi \, e^{iS[\phi]}, \tag{12.13}$$

where integration is taken over all possible trajectories $\phi(t)$ (with proper boundary values). In this approach the contribution of the topological term to the weight in the path integral is the phase $e^{i\theta\Delta\phi/(2\pi)}$ which is picked up by a particle moving in the presence of the vector potential.

Let us perform *Wick's rotation* replacing the time by an imaginary time $\tau = it$. Then

$$\int dt \, \frac{M\dot{\phi}^2}{2} \quad \rightarrow \quad i \int d\tau \, \frac{M\dot{\phi}^2}{2}, \tag{12.14}$$

$$\int dt \, A\dot{\phi} \quad \rightarrow \quad \int d\tau \, A\dot{\phi}, \tag{12.15}$$

where in the r.h.s dot means the derivative with respect to τ. The path integral Eq. (12.13) is then replaced by a Euclidean path integral

$$Z = \int_{e^{i[\phi(T)-\phi(0)]}=1} \mathcal{D}\phi \, e^{-S[\phi]}, \tag{12.16}$$

where the action

$$S = \int_0^\beta d\tau \left[\frac{M}{2} \dot{\phi}^2 - iA\dot{\phi} \right]. \tag{12.17}$$

We considered the amplitude of the return to the initial point in time β, i.e., $0 < \tau < \beta$. This requires periodic boundary conditions in time $e^{i\phi(0)} = e^{i\phi(\beta)}$.

We notice here that because the A-term is linear in time derivative, it does not change its form under Wick's rotation Eq. (12.15) and therefore, is still imaginary in Euclidean formulation Eq. (12.17). Without imaginary term one could think about e^{-S} as a Boltzmann weight in the classical partition function.

One can satisfy the boundary conditions as $\phi(\beta) - \phi(0) = 2\pi Q$ with any integer Q. We can rewrite the partition function Eq. (12.16) as:

$$Z = \sum_{Q=-\infty}^{+\infty} e^{i\theta Q} \int_{\phi(\beta)-\phi(0)=2\pi Q} \mathcal{D}\phi \, e^{-\int_0^\beta d\tau \, \frac{M}{2}\dot{\phi}^2}. \tag{12.18}$$

We notice here that $\theta = 2\pi n$ – multiple of 2π – is equivalent to $\theta = 0$. Second, we notice that the partition function is split into the sum of path integrals over distinct *topological sectors* characterized by an integer number Q which is called the *winding number*. The contributions of topological sectors to the total partition function are weighed with the complex weights $e^{i\theta Q}$.

For future comparisons, let us write Eq. (12.17) in terms of a unit two-component vector $\mathbf{\Delta} = (\Delta_1, \Delta_2) = (\cos\phi, \sin\phi)$, $\mathbf{\Delta}^2 = 1$.

$$S = \int_0^\beta d\tau \left[\frac{M}{2} \dot{\mathbf{\Delta}}^2 - iA(\Delta_1 \dot{\Delta}_2 - \Delta_2 \dot{\Delta}_1) \right]. \tag{12.19}$$

This is the simplest $(0+1)$-dimensional $O(2)$ non-linear σ-model.

12.2.4 Quantum doublet

Let us consider a particular limit of a very light particle on a circle $M \to 0$ in the presence of half of the flux quantum $A = 1/2$, $\theta = \pi$. With this flux, the ground state of the system is doubly degenerate $E_0 = E_1$ and the rest of the spectrum is separated by the energies $\sim 1/M \to \infty$ from the ground state Eq. (12.11). At large β (low temperatures) we can neglect contributions of all states except for the ground state.

We write the general form of the ground state wave function as $\alpha|+1/2\rangle + \beta|-1/2\rangle$, where $|+1/2\rangle = \psi_0$ and $|-1/2\rangle = \psi_1$. The ground state space (α, β) coincides with the one for a spin 1/2. One might say that Eqs. (12.16-12.17) with $M \to 0$ realize a path integral representation for the quantum spin 1/2. This representation does not have an explicit $SU(2)$ symmetry. We will consider an $SU(2)$-symmetric path integral representation for quantum spins later.

Meanwhile, let us discuss some topological aspects of a plane rotator problem.

12.2.5 Full derivative term and topology

From a mathematical point of view, the motion of a particle on a unit circle with periodic boundary conditions in time is described by a mapping $\phi(\tau) : S^1_\tau \to S^1_\phi$ of a circle formed by compactified time $S^1_\tau = \tau \in [0, \beta]$ into a circle $S^1_\phi = \phi \in [0, 2\pi]$. This mapping can be characterized by an integer *winding number* Q which tells us how many times the image ϕ goes around *target space* S^1_ϕ when variable τ changes from 0 to β.

It can be shown that two such mappings $\phi_1(\tau)$ and $\phi_2(\tau)$ can be continuously deformed into one another, if and only if they have the same winding number. Therefore, all mappings are divided into topological classes enumerated by $Q = 0, \pm 1, \pm 2, \ldots$. Moreover, one can define a group structure on topological classes. First, we define the *product* of two mappings ϕ_1 and ϕ_2 as

$$\phi_2 \cdot \phi_1(\tau) = \begin{cases} \phi_1(2\tau), & \text{for} \quad 0 < \tau < \beta/2\,, \\ \phi_1(\beta) + \phi_2(2\tau - \beta), & \text{for} \quad \beta/2 < \tau < \beta\,. \end{cases}$$

If ϕ_1 belongs to the topological class Q_1 and ϕ_2 to Q_2, their product belongs to the class $Q_1 + Q_2$. One can say that the product operation on mappings induces the structure of Abelian group on the set of topological classes. In this case this group is the group of integer numbers with respect to addition. One can write this fact down symbolically as $\pi_1(S^1) = \mathbb{Z}$, where subscript one denotes that our time is S^1 and S^1 in the argument is our target space. One says that the *first (or fundamental) homotopy group of* S^1 *is the group of integers.*

There is a simple formula giving the topological class $Q \in \mathbb{Z}$ in terms of $\phi(\tau)$

$$Q = \int_0^\beta \frac{d\tau}{2\pi}\, \dot{\phi}\,. \tag{12.20}$$

Let us now assume that we split our partition function into the sum over different topological classes. What are the general restrictions on the possible complex weights which one can introduce in the physical problem. One can deform smoothly any mapping in the class $Q_1 + Q_2$ into two mappings of classes Q_1 and Q_2 which are separated by a long time. Because of the multiplicative property of amplitudes, this means that the weights W_Q associated with topological classes must form a (unitary) representation of the fundamental group of a target space. The only unitary representation of $\mathbb{Z}$ is given by $W_Q = e^{i\theta Q}$ with $0 < \theta < 2\pi$ labelling different representations. In the case of plane rotator, these weights correspond to a phase due to the magnetic flux piercing the one-dimensional ring.

In more general case of, say, particle moving on the *manifold* G (instead of S^1) we have to consider the fundamental group of the target space $\pi_1(G)$, find its unitary representations, and obtain complex weights which could be associated with different topological classes.

12.2.6 Topological terms and quantum interference

As it can be seen from Eq. (12.18) the presence of a topological term in the action ($\theta \neq 0$) results in the interference between topological sectors in the partition function. The Boltzmann weight calculated for a trajectory within a given topological sector Q is additionally weighted with complex phase $e^{i\theta Q}$. This interference can not be removed by Wick's rotation.

12.2.7 General definition of topological terms

We define generally *topological terms* as *metric-independent* terms in the action.

A universal object present in any field theory, is the symmetric *stress-energy tensor* $T_{\mu\nu}$. It can be defined as a variation of the action with respect to the metric $g^{\mu\nu}$. More precisely, an infinitesimal variation of the action can be written as

$$\delta S = \int dx \sqrt{g}\, T_{\mu\nu} \delta g^{\mu\nu}, \tag{12.21}$$

where $\sqrt{g}\, dx$ is an invariant volume of space-time.

It immediately follows from our definition of topological terms that they do not contribute to the stress-energy tensor. If in a field theory all terms are topological we have $T_{\mu\nu} = 0$ for such a theory. These theories are called topological field theories.

A particular general covariant transformation is the rescaling of time. Topological terms do not depend on a time scale. Therefore, the corresponding Lagrangians are linear in time derivatives. They do not transform under Wick's rotation and are always imaginary in Euclidean formulation. They describe quantum interference which is not removable by Wick rotation.

12.2.8 Theta terms and their effects on the quantum problem

Theta terms are topological terms of a particular type. They appear when there exist nontrivial topological textures in space-time. Essentially, these terms are just complex weights of different topological sectors in the path integration. We will go over more details on θ-terms later in the course.

In addition to being imaginary in Euclidean formulation as all other topological terms θ-terms have also some special properties. These properties distinguish them from other types of topological terms. The following is a partial list of the features of topological θ-terms and of their manifestations.

- θ-terms assign complex weights in path integral to space-time textures with integer topological charge Q.
- Realize irreducible 1d-representations of $\pi_D(G)$, where D is the dimension of space-time and G is a target space.
- Quantum interference between topological sectors
- Do not affect equations of motion

- Affect the spectrum of a quantum problem by changing quantization rules
- Periodicity in coupling constant θ. [5]
- θ is not quantized (for $Q \in \mathbb{Z}$)
- For $\theta = 0, \pi$, there is an additional (parity) symmetry
- $\theta = \pi$ – degeneracy of the spectrum. Gapless excitations.
- Equivalent to changes in boundary conditions.
- θ is a new parameter which appears from the ambiguity of quantization of the classical problem for multiply-connected configurational space.

12.2.9 Exercises

Exercise 1: Particle on a ring, path integral

The Euclidean path integral for a particle on a ring with magnetic flux through the ring is given by

$$Z = \int \mathcal{D}\phi \, e^{-\int_0^\beta d\tau \left(\frac{m\dot{\phi}^2}{2} - i\frac{\theta}{2\pi}\dot{\phi} \right)}.$$

Using the decomposition

$$\phi(\tau) = \frac{2\pi}{\beta} Q\tau + \sum_{l\in\mathbb{Z}} \phi_l e^{i\frac{2\pi}{\beta} l\tau},$$

rewrite the partition function as a sum over topological sectors labeled by winding number $Q \in \mathbb{Z}$ and calculate it explicitly. Find the energy spectrum from the obtained expression.

Hint: Use summation formula

$$\sum_{n=-\infty}^{+\infty} e^{-\frac{1}{2}An^2 + iBn} = \sqrt{\frac{2\pi}{A}} \sum_{l=-\infty}^{+\infty} e^{-\frac{1}{2A}(B-2\pi l)^2}.$$

Exercise 2: Spin 1/2 from a particle on a ring

Calculate the partition function of a particle on a ring described in the previous exercise. Find explicit expressions in the limit $m \to 0$, $\theta \to \pi$ but $\theta - \pi \sim m/\beta$. One can interpret the obtained partition function as a partition function of a spin 1/2. What is the physical meaning of the ratio $(\theta - \pi)/m$ in the spin 1/2 interpretation of the result?

Hint: see Sec. 12.2.4.

[5] We assume that configurations are smooth and the space-time manifold is closed (no boundary).

Exercise 3: Metric independence of the topological term

The classical action of a particle on a ring is given by

$$S = \int dt_p \left(\frac{m\dot{\phi}^2}{2} - \frac{\theta}{2\pi}\dot{\phi} \right),$$

where t_p is some "proper" time. Reparametrizing time as $t_p = f(t)$ we have $dt_p = f'dt$, and $dt_p^2 = f'^2 dt^2$ and identify the metric as $g_{00} = f'^2$ and $g^{00} = f'^{-2}$. We also have $\sqrt{g_{00}} = f'$. Rewrite the action in terms of $\phi(t)$ instead of $\phi(t_p)$. Check that it has a proper form if written in terms of the introduced metric. Using the general formula for variation of the action with respect to a metric ($g = \det g_{\mu\nu}$)

$$\delta S = \frac{1}{2} \int dx \sqrt{g}\, T_{\mu\nu} \delta g^{\mu\nu},$$

find the stress-energy tensor for a particle on a ring. Check that T_{00} is, indeed, the energy of the particle.

12.3 Path integral for a single spin

Wess and Zumino introduced an effective Lagrangian to summarize the anomalies in current algebras [20]. E. Witten considered global (topological) aspects of this effective action [21]. Simultaneously, S. P. Novikov studied multi-valued functionals [22]. The corresponding topological terms are referred to as Wess-Zumino-Novikov-Witten terms or more often as just Wess-Zumino terms. In this section, we consider the simplest quantum mechanical (0+1 dimensional) version of such a term which is relevant for path integral formulation of a quantum mechanics of a single spin.

12.3.1 Quantum spin

Let us consider a simple example of how Wess-Zumino effective Lagrangian appears from the "current algebra". To simplify the story we take an example of quantum spin S. This is a quantum mechanical system with an $SU(2)$ spin algebra playing the role of "current algebra" of quantum field theory. We have standard spin commutation relations

$$[S^a, S^b] = i\epsilon^{abc} S^c, \tag{12.22}$$

where a, b, c take values x, y, z. We require that

$$\mathbf{S}^2 = S(S+1), \tag{12.23}$$

where $2S$ is an integer number defining the representation (the value of spin). Let us consider the simplest possible Hamiltonian of a quantum spin in a constant magnetic field

$$H = -\mathbf{h} \cdot \mathbf{S} \tag{12.24}$$

and derive an operator equation of motion

$$\partial_t \mathbf{S} = i[H, \mathbf{S}] = -i[\mathbf{h}\cdot\mathbf{S}, \mathbf{S}] = \mathbf{S}\times\mathbf{h}. \tag{12.25}$$

In the classical limit $S \to \infty$ (or $\hbar \to 0$) it is convenient to write $\mathbf{S} \to S\mathbf{n}$ so that $\mathbf{n}$ is a classical unit vector $\mathbf{n}^2 = 1$ and equation of motion Eq. (12.25) becomes classical equation of motion

$$\partial_t \mathbf{n} = \mathbf{n}\times\mathbf{h}. \tag{12.26}$$

The natural question immediately occurs is what classical action corresponds to this equation of motion. It turns out that writing down this action is not completely trivial problem if one desires for the action to have explicitly $SU(2)$ invariant form. Let us first derive it using non-invariant parameterization in terms of spherical angles $\mathbf{n} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$. We assume that the angle θ is measured from the direction of magnetic field $\mathbf{h} = (0, 0, h)$. The Hamiltonian Eq. (12.24) becomes $H = -hS\cos\theta$ and equations of motion Eq. (12.26) become $\dot\phi = -h$ and $\dot\theta = 0$ – precession around the direction of magnetic field. We obtain these equations as Hamilton's equations, identifying the momentum conjugated to ϕ coordinate as

$$p_\phi = -S(1-\cos\theta). \tag{12.27}$$

Then the classical action of a single spin in magnetic field can be written as

$$S[\mathbf{n}] = -4\pi S W_0 + \int dt\, S\,\mathbf{h}\cdot\mathbf{n}, \tag{12.28}$$

where W_0 is defined using a particular choice of coordinates as

$$W_0 = \frac{1}{4\pi}\int dt\,(1-\cos\theta)\partial_t\phi = \frac{1}{4\pi}\int d\phi\,(1-\cos\theta) = \frac{\Omega}{4\pi}, \tag{12.29}$$

where Ω is a solid angle encompassed by the trajectory of $\mathbf{n}(t)$ during time evolution. The first term in the action Eq. (12.28) has a form of $\int dt\, p_\phi\dot\phi$ and the second is a negative time integral of the Hamiltonian.

Although Eq. (12.29) has a nice geometrical meaning it is written in some particular coordinate system on two-dimensional sphere. It would be nice to have an expression for W_0 which is coordinate independent and explicitly $SU(2)$ invariant (with respect to rotations of $\mathbf{n}$). Such a form, indeed, exists

$$W_0 = \int_0^1 d\rho \int_0^\beta dt\, \frac{1}{8\pi}\epsilon^{\mu\nu}\mathbf{n}\cdot[\partial_\mu\mathbf{n}\times\partial_\nu\mathbf{n}]. \tag{12.30}$$

Here we assume periodic boundary conditions in time $\mathbf{n}(\beta) = \mathbf{n}(0)$, ρ is an auxiliary coordinate $\rho\in[0,1]$. $\mathbf{n}$-field is extended to $\mathbf{n}(t,\rho)$ in such a way that $\mathbf{n}(t,0) = (0,0,1)$ and $\mathbf{n}(t,1) = \mathbf{n}(t)$. Indices μ,ν take values t,ρ.

Wess-Zumino action Eq. (12.30) has a very special property. Although it is defined as an integral over two-dimensional disk parameterized by ρ and t its

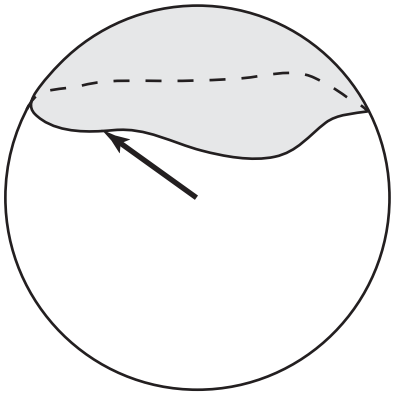


Figure 12.2: The unit vector $\mathbf{n}(\tau)$ draws a closed line on the surface of a sphere with unit radius during its motion in imaginary time. Berry phase is proportional to the solid angle (shaded region) swept by the vector $\mathbf{n}(t)$. One can calculate this solid angle by extending $\mathbf{n}$ into the two-dimensional domain $\mathcal{B}$ as $\mathbf{n}(\rho, t)$ and calculating (12.30).

variation depends only on the values of $\mathbf{n}$ on the boundary of the disk—physical time. Indeed one can check that

$$\begin{aligned} \delta W_0 &= \int_0^1 d\rho \int_0^\beta dt\, \frac{1}{4\pi} \epsilon^{\mu\nu} \mathbf{n} \cdot [\partial_\mu \delta\mathbf{n} \times \partial_\nu \mathbf{n}] \\ &= \int_0^1 d\rho \int_0^\beta dt\, \partial_\mu \left\{ \frac{1}{4\pi} \epsilon^{\mu\nu} \mathbf{n} \cdot [\delta\mathbf{n} \times \partial_\nu \mathbf{n}] \right\} \\ &= \frac{1}{4\pi} \int_0^\beta dt\, \delta\mathbf{n} \cdot [\dot{\mathbf{n}} \times \mathbf{n}]\,, \end{aligned} \tag{12.31}$$

where we used that $\delta\mathbf{n} \cdot [\partial_\mu \delta\mathbf{n} \times \partial_\nu \mathbf{n}] = 0$ because all three vectors $\delta\mathbf{n}$, $\partial_\mu \mathbf{n}$, and $\partial_\nu \mathbf{n}$ lie in the same plane (tangent to the two-dimensional sphere $\mathbf{n}^2 = 1$. Due to this property classical equation of motion does not depend on the arbitrary extension of $\mathbf{n}$ to $\rho \neq 1$.

In quantum physics, however, not only the variation δW_0 but the weight $e^{2\pi i W_0}$ should not depend on unphysical configuration $\mathbf{n}(t, \rho)$ but only on $\mathbf{n}(t, \rho = 1)$. To see that this is indeed so, we consider the configuration $\mathbf{n}(t, \rho)$ as a mapping from two-dimensional disk $(t, \rho) \in \mathcal{B}_+$ into the two-dimensional sphere $\mathbf{n} \in S^2$. Suppose now that we use another extension $\mathbf{n}'(t, \rho)$ and represent it as a mapping of another disk $\mathcal{B}_-$ with the same boundary (physical

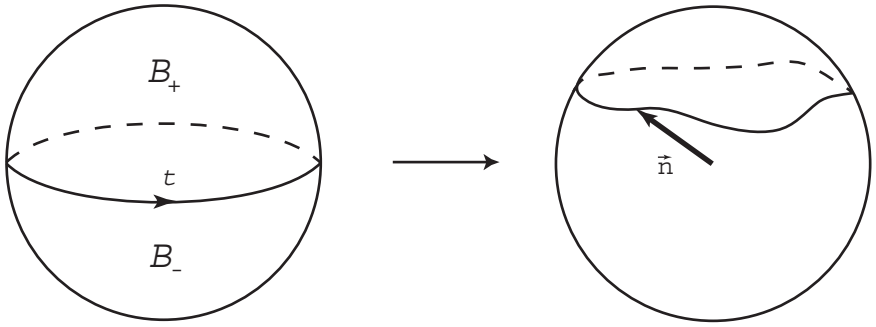


Figure 12.3: Two extensions $\mathbf{n}(t,\rho)$ and $\mathbf{n}'(t,\rho)$ define a mapping $S^2 \to S^2$. The difference $W_0[\mathbf{n}] - W_0[\mathbf{n}']$ gives a winding number of this mapping.

time) into S^2. We have

$$\begin{aligned} W_0[\mathbf{n}] - W_0[\mathbf{n}'] &= \int_{\mathcal{B}_+} d^2x \, \frac{1}{8\pi} \epsilon^{\mu\nu} \mathbf{n} \cdot [\partial_\mu \mathbf{n} \times \partial_\nu \mathbf{n}] \\ &\quad - \int_{\mathcal{B}_-} d^2x \, \frac{1}{8\pi} \epsilon^{\mu\nu} \mathbf{n}' \cdot [\partial_\mu \mathbf{n}' \times \partial_\nu \mathbf{n}'] \\ &= \int_{S^2 = \mathcal{B}_+ \cup \mathcal{B}_-} d^2x \, \frac{1}{8\pi} \epsilon^{\mu\nu} \mathbf{n} \cdot [\partial_\mu \mathbf{n} \times \partial_\nu \mathbf{n}] = k \,, \end{aligned} \tag{12.32}$$

where we changed the orientation of $\mathcal{B}_-$ and considered $\mathcal{B}_\pm$ as an upper (lower) part of some two-dimensional sphere (see Fig.12.3). One can recognize the last integral [15] as a winding number k of the first sphere ($\mathcal{B}_+ \cup \mathcal{B}_-$ around the second $\mathbf{n} \in S^2$). This number is always an integer proving that $e^{2\pi i W_0}$ does not depend on the particular way of an extension $\mathbf{n}(t,\rho)$. We notice here that in general the topological term W_0 can appear in the action only with the coefficient which is a multiple of $2\pi i$. Otherwise, it depends on the unphysical values of $\mathbf{n}(t,\rho)$ and is not defined. Such a term is called[6] "Wess-Zumino term" or "WZ term" by names of Wess and Zumino who discovered a similar term first in the context of four-dimensional quantum field theories [20]. If Wess-Zumino term is present with some coupling constant g so that the weight in partition function is proportional to $e^{2\pi i g W_0}$ we immediately conclude that g must be an integer. This phenomenon is called "topological quantization" of physical constant g and is a very important consequence of Wess-Zumino term.

[6]It is also often called WZW or Wess-Zumino-Witten or even WZWN or Wess-Zumino-Novikov-Witten term to honor E. Witten [21, 23] and S.P. Novikov [22].

To obtain the equations of motion from Eq. (12.28,12.30) we use Eq. (12.31) and introduce Lagrange multiplier λ to enforce constraint $\mathbf{n}^2 = 1$. Then we obtain for the variation of the action

$$\delta_n \left(S[\mathbf{n}] + \lambda(\mathbf{n}^2 - 1) \right) = -4\pi S \frac{1}{4\pi} \left[\dot{\mathbf{n}} \times \mathbf{n} \right] + S\mathbf{h} + 2\lambda \mathbf{n} = 0 \,. \tag{12.33}$$

Vector-multiplying Eq. (12.33) from the right by $\mathbf{n}$ we arrive at Eq. (12.26).

In this simplified treatment we just found some classical action which reproduces the classical limit of operator equations of motion Eq. (12.25). One can proceed more formally starting with commutation relations Eq. (12.22) and quantum Hamiltonian Eq. (12.24) and derive the classical action Eq. (12.28) using, e.g., coherent states method [9].

The purpose of this exercise was to illustrate that the Wess-Zumino term W_0 summarizes at the classical level the commutation relations Eq. (12.22). One can also show that reversely the path integral quantization of Eq. (12.33) produces the commutation relations Eq. (12.22).

12.3.2 Fermionic model

In this section, we use a very simple quantum mechanical example to show how topological terms are generated when one passes from microscopic theory to an effective description. Generally, in condensed matter physics we are dealing with some system of electrons interacting with each other as well as with other degrees of freedom such as a lattice. Let us assume that at some low energy scale we reduced our problem to fermions interacting to a bosonic field. The bosonic field may originate both from the collective behavior of electrons, e.g., magnetization or superconducting order parameter, and from independent degrees of freedom, e.g., from the vibrations of the lattice. For our illustrative example we consider [24]

$$S = \int dt \, \psi^\dagger \left[i\partial_t + m\mathbf{n} \cdot \boldsymbol{\sigma} \right] \psi, \tag{12.34}$$

where m is a coupling constant, $\psi = (\psi_1, \psi_2)^t$ is a spinor, and $\boldsymbol{\sigma}$ is a triplet of Pauli matrices. In this case, the fermions are represented by just one spinor and the bosonic field by a single unit vector $\mathbf{n} = (n_1, n_2, n_3)$, $\mathbf{n} \in S^2$. The latter means that $\mathbf{n}$ takes its values on a two-dimensional sphere, i.e., $\mathbf{n}^2 = 1$. This model can originate, e.g., from electrons interacting with a localized magnetic moment. Then coupling constant $m > 0$ corresponds to a Hund's coupling between electrons (one electron for simplicity) and the direction $\mathbf{n}$ of a localized moment. Notice, that a more complete theory must have the bare action of a moment $\mathbf{n}$ added to a Eq. (12.34). We, however, are interested only in the action of $\mathbf{n}$ induced by an interaction with fermions.

For future convenience, we will use a Euclidean formulation here and in the rest of the paper. It can be obtained by "Wick rotation" $t \to it$. A Euclidean

action obtained from Eq. (12.34) is

$$S_E = \int dt\, \psi^\dagger \left[\partial_t - m\mathbf{n}\cdot\boldsymbol{\sigma}\right]\psi. \tag{12.35}$$

Effective action by chiral rotation trick

We consider partition function

$$Z = \int \mathcal{D}\psi\, \mathcal{D}\bar{\psi}\, \mathcal{D}\mathbf{n}\, e^{-S_E} = \int \mathcal{D}\mathbf{n}\, e^{-S_{\text{eff}}}, \tag{12.36}$$

where the last equality is a definition of an effective action

$$S_{\text{eff}} = -\ln \int \mathcal{D}\psi\, \mathcal{D}\bar{\psi}\, e^{-S_E} = -\ln \det D, \tag{12.37}$$

where we defined an operator $D \equiv \partial_t - m\mathbf{n}\cdot\boldsymbol{\sigma}$. To calculate the logarithm of the fermionic determinant we use "chiral rotation". Namely, we introduce the matrix field $U(t) \in SU(2)$ such that $U^\dagger \mathbf{n}\cdot\boldsymbol{\sigma} U = \sigma^3$ so that

$$\tilde{D} = U^\dagger D U = \partial_t - i\hat{a} - m\sigma^3 = G_0^{-1} - i\hat{a}, \tag{12.38}$$

with

$$\hat{a} \equiv U^\dagger i\partial_t U, \tag{12.39}$$

and

$$G_0 = \left(\partial_t - m\sigma^3\right)^{-1}. \tag{12.40}$$

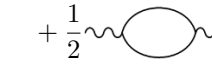
Then we write[7]

$$S_{\text{eff}} = -\ln\det D = -\ln\det\tilde{D} = -\text{Tr}\,\ln\tilde{D}. \tag{12.41}$$

Let us now write $\tilde{D} = G_0^{-1}(1 - G_0 i\hat{a})$ and expand

$$\begin{aligned} S_{\text{eff}} = -\text{Tr}\,\ln\tilde{D} &= \text{Tr}\left[\ln G_0 + G_0 i\hat{a} + \frac{1}{2}(G_0 i\hat{a})^2 + \ldots\right] \\ &= S^{(0)} + S^{(1)} + S^{(2)} + \ldots. \end{aligned} \tag{12.42}$$

The expansion Eq. (12.42) has the following diagrammatic representation

$S_{\text{eff}} = const +$  $+ \frac{1}{2}$  $+ \ldots.$ (12.43)

[7]Notice that the second equality in Eq. (12.41) is the common source of miscalculated topological terms. Quantum anomalies might be present making chiral rotation technique inapplicable. In this case this is a legitimate procedure because of the absence of so-called global anomalies [11, 25].

The zeroth order term $S^{(0)}$ is an (infinite) constant which does not depend on $\hat{a}$. The first order term is given by

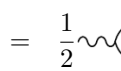
$S^{(1)} =$  $= \mathrm{Tr}\,[G_0 i\hat{a}] = \int \frac{d\omega}{2\pi}\,\mathrm{tr}\left[\frac{1}{-i\omega - m\sigma^3} i\hat{a}_{\omega=0}\right]. \quad (12.44)$

Here $\hat{a}_{\omega=0} = \int dt\,\hat{a}(t)$ and tr is taken over sigma-matrices. Closing the integral in an upper complex ω-plane we obtain

$$S^{(1)} = \int \frac{d\omega}{2\pi}\,\mathrm{tr}\left[\frac{1}{-i\omega - m\sigma^3} i\hat{a}_{\omega=0}\right] = -i a^3_{\omega=0} = -i\int dt\, a^3(t), \quad (12.45)$$

where only the term containing a^3 ($\hat{a} = a^k\sigma^k$, $k = 1, 2, 3$) does not vanish when trace over Pauli matrices is taken.[8]

We may proceed and obtain for the second term of an expansion

$$\begin{aligned} S^{(2)} &= \frac{1}{2}\,\text{  } = \frac{1}{2}\mathrm{Tr}\,[G_0 i\hat{a} G_0 i\hat{a}] \\ &= \frac{1}{2}\int\frac{d\Omega}{2\pi}\int\frac{d\omega}{2\pi}\,\mathrm{tr}\left[\frac{1}{-i\omega - m\sigma^3} i\hat{a}_{-\Omega}\frac{1}{-i(\omega+\Omega) - m\sigma^3} i\hat{a}_{\Omega}\right] \\ &= \frac{1}{8m}\int\frac{d\Omega}{2\pi}\,\mathrm{tr}\left[\hat{a}_{-\Omega}\hat{a}_{\Omega} - \sigma^3\hat{a}_{-\Omega}\sigma^3\hat{a}_{\Omega}\right] + o\left(\frac{1}{m}\right) \\ &= \frac{1}{2m}\int dt\,\left[(a^1)^2 + (a^2)^2\right] + o\left(\frac{1}{m}\right). \end{aligned} \quad (12.46)$$

We neglected here the terms of higher order in $1/m$. Therefore, for an effective action we obtain up to the terms of the order of $1/m$ and omitting constant

$$S_{\text{eff}} = S^{(1)} + S^{(2)} = -i\int dt\, a^3(t) + \frac{1}{2m}\int dt\,\left[(a^1)^2 + (a^2)^2\right]. \quad (12.47)$$

The effective action Eq. (12.47) is expressed in terms of an auxiliary gauge field $\hat{a}$. However, one should be able to re-express it in terms of physical variable $\mathbf{n}$ as it was defined by Eq. (12.37) which contains only $\mathbf{n}$. Let us start with the second term. Using an explicit relation $\mathbf{n}\cdot\boldsymbol{\sigma} = U\sigma^3 U^\dagger$ and the definition Eq. (12.39) one can easily check that $(\partial_t\mathbf{n})^2 = 4\left[(a^1)^2 + (a^2)^2\right]$ and the last term of Eq. (12.47) indeed can be expressed in terms of $\mathbf{n}$ as

$$S^{(2)} = \frac{1}{2m}\int dt\,\left[(a^1)^2 + (a^2)^2\right] = \frac{1}{8m}\int dt\,(\partial_t\mathbf{n})^2. \quad (12.48)$$

Obtaining $S^{(1)}$ is a bit more subtle. The gauge field is defined as Eq. (12.39) with matrix U defined implicitly by $\mathbf{n}\cdot\boldsymbol{\sigma} = U\sigma^3 U^\dagger$. One can see from the

[8] We notice that the ω-integral in Eq. (12.45) is formally diverging. However, being regularized, it becomes the number of fermions in the system. In our model we have exactly one fermion and the regularization procedure in this case is just the closing of the contour of an integration in an upper complex plane.

latter expression that the definition of U is ambiguous. Indeed, one can make a "gauge transformation"

$$U \to U e^{i\sigma^3\psi} \tag{12.49}$$

with $\psi(t)$ any function of t without changing $\mathbf{n}$. Under this transformation the gauge field is transformed as $\hat{a} \to e^{-i\sigma^3\psi}\hat{a}e^{i\sigma^3\psi} - \sigma^3\partial_t\psi$, or

$$a^3 \to a^3 - \partial_t\psi, \tag{12.50}$$

$$a^1 \to a^1\cos 2\psi - a^2 \sin 2\psi, \tag{12.51}$$

$$a^2 \to a^1 \sin 2\psi + a^2 \cos 2\psi. \tag{12.52}$$

Therefore, $S^{(1)} \to S^{(1)} + i\int dt\, \partial_t\psi$, and we notice that $S^{(1)}$ transforms non-trivially[9] under the change of U and therefore, can not be expressed as a simple time integral over the function which depends on $\mathbf{n}$ only. One might question the validity of our derivation because it seems that $S^{(1)}$ defined by Eq. (12.45) is not invariant under the transformation Eq. (12.49) but we notice that $S^{(1)}$ changes only by the integral of a full time derivative. Moreover, if we require periodicity in time, i.e., time changes from 0 to β and $\psi(\beta) = \psi(0) + 2\pi n$ with an integer n, then $S^{(1)} \to S^{(1)} + 2\pi i n$ and "Boltzmann" weight $e^{-S^{(1)}}$ is invariant under Eq. (12.49). Therefore, the contribution to the partition function from the $S^{(1)}$ term depends only on the physical variable $\mathbf{n}$. To understand what is going on let us calculate $S^{(1)}$ explicitly. We parametrize $\mathbf{n} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$. Then, the most general choice of U is

$$U = \begin{pmatrix} \cos\frac{\theta}{2} & e^{-i\phi}\sin\frac{\theta}{2} \\ e^{i\phi}\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \end{pmatrix} e^{i\sigma^3\psi},$$

where $\psi(t)$ is an arbitrary function of t with $\psi(\beta) = \psi(0) + 2\pi n$. It is straightforward to calculate

$$a_3 = -\frac{1-\cos\theta}{2}\partial_t\phi - \partial_t\psi.$$

We see that the last term can be discarded by reasons given above, and we have

$$S^{(1)} = 2\pi i W_0, \tag{12.53}$$

with W_0 defined in Eq. (12.29).

Combining Eq. (12.48) and Eq. (12.53) together we obtain

$$S_{\text{eff}} = 2\pi i W_0 + \frac{1}{8m}\int dt\,(\partial_t\mathbf{n})^2 + o\left(\frac{1}{m}\right). \tag{12.54}$$

In the case where N species of fermions coupled to the same $\mathbf{n}$ field are present, one obtains an overall factor N in effective action, i.e., $S_{\text{eff}} \to NS_{\text{eff}}$.

[9]Notice that Eq. (12.48) does not transform under this gauge transformation.

Topological term

Let us notice that the first term $2\pi i W_0$ of gradient expansion Eq. (12.54) is very different from, say, the second one in the following respects

1. It is imaginary.
2. It does not depend on the value of the mass parameter m.[10]
3. It does not change under reparameterization of time $t \to f(t)$. In particular, it is scale invariant and does not change when $t \to \lambda t$.

The property 3 makes it natural to call the term $2\pi i\, W_0$ topological as it does not depend on time scales but only on the trajectory of $\mathbf{n}(t)$. In fact, properties 1, 2 are consequences of 3 and are the general properties of all topological terms.

Path integral representation of quantum spin

Before going to the next section let us consider some application of derived topological term. We generalize our model slightly so that in Eq. (12.35) ψ denotes N species of fermions which are all coupled to the same bosonic field $\mathbf{n}$. We consider the special limit $m \to \infty$ of the model Eq. (12.35). The Hamiltonian of the model $-m\mathbf{n} \cdot \psi^\dagger \boldsymbol{\sigma} \psi$ in this limit forces all spins of ψ particles to be aligned along $\mathbf{n}$. Therefore, we expect that, in this limit, after an integration over fermions, we will obtain an effective action of a quantum spin $S = N/2$ written in terms of the direction $\mathbf{n}$ of its quantization axis. Multiplying Eq. (12.54) by the number of fermion species N, and taking limit $m \to \infty$, we obtain

$$S = 2\pi i N W_0. \tag{12.55}$$

One can show that upon quantization[11] the components of $\mathbf{n}$ become the components of the quantum spin $S = N/2$ so that $n_a \to \hat{S}_a/S$. The action Eq. (12.55) is explicitly $SU(2)$ invariant and is well-defined for integer $2S = N$, i.e., spin can only be integer or half-integer. Therefore, in path integral formulation the quantization of spin is a consequence of the Wess-Zumino term in the action of the spin.

12.3.3 Derivation of a WZ term from fermionic model without chiral rotation

Here we give an alternative derivation of an effective action Eq. (12.54) from Eq. (12.37) which does not use chiral rotation trick. [10]

[10] If we allow m to be negative this term becomes $2\pi i(\operatorname{sgn} m) W_0$ and depends only on the sign of m, not its magnitude.

[11] The easy way to show that we are dealing with the spin is to add coupling to an external magnetic field $-\int dt\, S\mathbf{h} \cdot \mathbf{n}$ to Eq. (12.55) and write down the classical equation of motion for $\mathbf{n}$ using Eq. (12.31) and constraint $\mathbf{n}^2 = 1$. We obtain $\dot{\mathbf{n}} = [\mathbf{n} \times \mathbf{h}]$ where we assumed $S = N/2$ and changed to the real time $t \to it$. The obtained equation is indeed the classical equation of spin precession.

Consider

$$Z = \int \mathcal{D}\psi\, \mathcal{D}\bar{\psi}\, \mathcal{D}\mathbf{n}\, e^{-S}, \tag{12.56}$$

where

$$S = \int_0^T dt\, \bar{\psi}\, (i\partial_t - im\mathbf{n}\cdot\boldsymbol{\tau})\, \psi. \tag{12.57}$$

Here $\psi = (\psi_1, \psi_2)$ is a Grassmann spinor representing spin 1/2 fermion, $\boldsymbol{\tau}$ the Pauli matrices acting on spinor indices of ψ, and $\mathbf{n}^2 = 1$ is a unit, three-component vector coupled to the spin of fermion $\bar{\psi}\boldsymbol{\tau}\psi$ with the coupling constant m. Integrating out fermions in Eq. (12.56) we obtain

$$Z = \int \mathcal{D}\mathbf{n}\, e^{-S_{\rm eff}(\mathbf{n})}, \tag{12.58}$$

with

$$S_{\rm eff}(\mathbf{n}) = -\ln\det\left(i\partial_t - im\mathbf{n}\cdot\boldsymbol{\tau}\right). \tag{12.59}$$

Let us denote $D = i\partial_t - im\mathbf{n}\cdot\boldsymbol{\tau}$ and $D^\dagger = i\partial_t + im\mathbf{n}\cdot\boldsymbol{\tau}$. We calculate the variation of the effective action

$$\begin{aligned}\delta S_{\rm eff} &= -{\rm Tr}\left\{\delta D D^{-1}\right\} = -{\rm Tr}\left\{\delta D D^\dagger (DD^\dagger)^{-1}\right\} \\ &= im{\rm Tr}\left\{\delta\mathbf{n}\cdot\boldsymbol{\tau}(i\partial_t + im\mathbf{n}\cdot\boldsymbol{\tau})(-\partial_t^2 + m^2 - m\dot{\mathbf{n}}\cdot\boldsymbol{\tau})^{-1}\right\}\end{aligned} \tag{12.60}$$

Expanding the fraction in $\dot{\mathbf{n}}$, calculating the trace, and keeping only lowest orders in $\dot{\mathbf{n}}/m$, we obtain:

$$\delta S_{\rm eff} = \int dt\, \left\{\frac{1}{4m}\delta\dot{\mathbf{n}}\cdot\dot{\mathbf{n}} - \frac{i}{2}\delta\mathbf{n}\cdot[\mathbf{n}\times\dot{\mathbf{n}}]\right\}. \tag{12.61}$$

Restoring the effective action from its variation we have:

$$S_{\rm eff} = \int_0^T dt\, \frac{1}{8m}\dot{\mathbf{n}}^2 - 2\pi i W_0, \tag{12.62}$$

where the *Wess-Zumino action*

$$W_0 = \int_0^1 d\rho \int_0^T dt\, \frac{1}{8\pi}\epsilon^{\mu\nu}\mathbf{n}\cdot[\partial_\mu\mathbf{n}\times\partial_\nu\mathbf{n}]. \tag{12.63}$$

Here ρ is an auxiliary coordinate $\rho \in [0,1]$. Also, the $\mathbf{n}$-field is extended to $\mathbf{n}(t,\rho)$ in such a way that $\mathbf{n}(t,0) = (0,0,1)$ and $\mathbf{n}(t,1) = \mathbf{n}(t)$. Indices μ,ν take values t,ρ.

The Wess-Zumino action, Eq. (12.63), has a very special property. Although it is defined as an integral over two-dimensional disk parameterized by ρ and t its variation depends only on the values of $\mathbf{n}$ on the boundary of the disk—physical time.

12.3.4 Quantum spin as a particle moving in the field of Dirac monopole

Let us think of the action Eq. (12.62) as of the action of a charged particle moving on a surface of two-dimensional sphere with unit radius so that $\mathbf{n}$ is a position of particle on the sphere. Then the first term of Eq. (12.62) is a conventional kinetic energy of the particle. The second term should then be interpreted as a phase picked by particle moving in the field a magnetic charge $2S$ (Dirac monopole) placed in the center of the sphere.

One could have started with the problem of particle of the mass m moving in the field of magnetic monopole of charge $2S$. Then the ground state is $2S$-degenerate and is separated by the gap $\sim 1/m$ from the rest of the spectrum. In the limit $m \to 0$ only the ground state is left and we obtain a quantum spin problem in an approach analogous to the plane rotator from Sec.12.2.4.

12.3.5 Reduction of a WZ term to a theta-term

Let us consider the value of Eq. (12.29) assuming that the polar angle is kept constant at $\theta(\tau) = \theta_0$. Then Eq. (12.29) becomes

$$W_0 = \frac{1-\cos\theta_0}{2} \int_0^\beta \frac{d\tau}{2\pi}\, \partial_\tau \phi, \tag{12.64}$$

and we recognize Eq. (12.53) with Eq. (12.64) as the theta-term Eq. (12.12) corresponding to the particle on a ring with the flux through the ring given by

$$A = \frac{1-\cos\theta_0}{2}. \tag{12.65}$$

In particular, for $\theta_0 = \pi/2$, the topological term in the action of a particle on a ring in magnetic field $A = 1/2$.

12.3.6 Properties of WZ terms

WZ terms

1. do not depend on the metric of spacetime
2. are imaginary in Euclidean formulation
3. do not contribute to stress-energy tensor (and to Hamiltonian).
4. do not depend on m – the scale, below which an effective action is valid (but do depend on $\operatorname{sgn}(m)$)
5. are antisymmetric in derivatives with respect to different space-time coordinates (contain $\epsilon^{\mu\nu\lambda\cdots}$)
6. are written as integrals of (D+1)-forms over auxiliary $(D+1)$-dimensional space - disk D^{D+1} such that $\partial D^{D+1} = S^D$ - compactified space-time

7. are multi-valued functionals. Multi-valuedness results in quantization of coupling constants (coefficients in front of WZ terms)
8. do change equations of motion by changing commutation relation between fields (Poisson's brackets), not by changing Hamiltonian
9. might lead to massless excitations with "half-integer spin" (see Sec. 12.4.5)
10. describe boundary theories of models with θ-terms (see Sec. 12.4.5)
11. being combined (see the spin chains Sec. 12.4.2) produce θ-terms
12. can be calculated by gradient expansion of the variation of fermionic determinants
13. produce θ terms as a reduction of target space (see Sec. 12.3.5). [10]

Among the listed properties the first five 1-5 are the properties of all topological terms while the others are more specific to WZ terms.

12.3.7 Exercises

The exercises Eq. (4-9) were solved in the main text. Try to solve them independently and test your understanding by solving exercises Eq. (10-11).

Exercise 4: WZ term in 0 + 1, preliminaries

Consider a three-dimensional unit vector field $\mathbf{n}(x, y)$ ($\mathbf{n} \in S^2$) defined on a two-dimensional disk D. Define

$$W_0 = \int_D d^2x \, \frac{1}{8\pi} \epsilon^{\mu\nu} \mathbf{n} \cdot [\partial_\mu \mathbf{n} \times \partial_\nu \mathbf{n}] = \int_D \frac{1}{16\pi i} \mathrm{tr}\, [\hat{n} d\hat{n} d\hat{n}], \tag{12.66}$$

where the latter expression is written in terms of differential forms and $\hat{n} = \mathbf{n} \cdot \boldsymbol{\sigma}$.

a) Calculate the variation of W_0 with respect to $\mathbf{n}$. Show that the integral becomes the integral over disk D of the complete divergence (of the exact form).

b) Parametrize the boundary ∂D of the disk by parameter t, apply Gauss-Stokes theorem and express the result of the variation using only the values of $\mathbf{n}(t)$ at the boundary.

We showed that the variation of W depends only on the boundary (i.e., physical) values of $\mathbf{n}$-field. See Eq. (12.31) for the answer.

Exercise 5: WZ term in 0 + 1, definition

Assume that we are given the time evolution of $\mathbf{n}(t)$ field ($\mathbf{n} \in S^2$). We also assume that time can be compactified, i.e., $\mathbf{n}(t = \beta) = \mathbf{n}(t = 0)$. Consider the two-dimensional disk D which boundary ∂D is parametrized by time $t \in [0, \beta]$. The WZ term is defined by

$$S_{WZ} = i4\pi S \, W_0[\mathbf{n}], \tag{12.67}$$

where S is some constant, W_0 is given by Eq. (12.66), and $\mathbf{n}(x, y)$ is some arbitrary smooth extension of $\mathbf{n}(t)$ from the boundary to an interior of the disk.

Let us show that the WZ term is well defined and (almost) does not depend on the extension of $\mathbf{n}(t)$ to the interior of D.

Consider two different extensions $\mathbf{n}^{(1)}(x,y)$ and $\mathbf{n}^{(2)}(x,y)$ of the same $\mathbf{n}(t)$ and corresponding values $W_0^{(1)}$ and $W_0^{(2)}$ of the functional W_0. Show that the difference $W_0^{(1)} - W^{(2)}$ is an integer number - the degree Q of mapping $S^2 \to S^2$. The second S^2 here is a target space of $\mathbf{n}$. How did the first S^2 appear?

We see that $S_{WZ}[\mathbf{n}(t)]$ is a multi-valued functional which depends on the extension of $\mathbf{n}$ to the disk D. However, the weight in partition function is given by $e^{-S_{WZ}}$ and can be made single-valued functional if the coupling constant S is "quantized". Namely, if $2S \in \mathbb{Z}$ (S - half-integer number) the $e^{-S_{WZ}}$ is a well-defined single-valued functional.

For the answer see Eq. (12.32).

Exercise 6: WZ in 0 + 1, spin precession

Let us consider the quantum-mechanical action of the unit vector $\mathbf{n}(t)$ with the (Euclidean) action

$$S_h = S_{WZ}[\mathbf{n}(t)] - S \int dt\, \mathbf{h} \cdot \mathbf{n}(t), \tag{12.68}$$

where S_{WZ} is given by Eq. (12.67) and $\mathbf{h}$ is some constant three-component vector (magnetic field).

Find the classical equation of motion for $\mathbf{n}(t)$ from the variational principle $\delta S_h = 0$. Remember that one has a constraint $\mathbf{n}^2 = 1$ which can be taken into account using, e.g., Lagrange multiplier trick.

The obtained expression is the equation of spin precession and S_{WZ} is a proper, explicitly $SU(2)$ invariant action for the free spin S.

For the answer see Eq. (12.33,12.26).

Exercise 7: WZ in 0 + 1, quantization

Show that the classical equations of motion obtained from S_h correspond to Heisenberg equations (in real time) $\partial_t \hat{\mathbf{S}} = i\left[H, \hat{\mathbf{S}}\right]$ for the quantum spin operator $\hat{\mathbf{S}}$

$$\left[S^a, S^b\right] = i\epsilon^{abc} S^c \tag{12.69}$$

obtained from the Hamiltonian of a spin in magnetic field

$$H = -\mathbf{h} \cdot \hat{\mathbf{S}}. \tag{12.70}$$

Obtain the commutation relations of quantum spin Eq. (12.69) from the topological part S_{WZ}. Notice that this topological action is linear in time derivative and, therefore, does not contribute to the Hamiltonian. Nevertheless, it defines commutation relations between components of the spin operator.

Hint: You can either use local coordinate representation of the unit vector in terms of spherical angles $\mathbf{n} = (\cos\phi \sin\theta, \sin\phi \sin\theta, \cos\theta)$ or use the general formalism of obtaining Poisson bracket from the symplectic form given in S_{WZ}.

Exercise 8: Reduction of the WZ-term to the theta-term in 0 + 1

Let us assume that the field $\mathbf{n}(t)$ is constrained so that it takes values on a circle given in spherical coordinates by $\theta = \theta_0 = const$. Find the value of the topological term S_{WZ} on such configurations (notice that this constraint is not applicable in

the interior of the disk D, only at its physical boundary). Show that the obtained topological term is a theta-term in $0+1$ corresponding to $S^1 \to S^1$.

What is the value of the coefficient in front of that topological term? What is the value of corresponding "magnetic flux" through a ring? For $S = 1/2$, which reduction (value of θ_0) corresponds to the half of the flux quantum?

For the answer see Sec. 12.3.5.

Exercise 9: WZ in 0 + 1, derivation from fermions

Consider a Euclidean action of a fermion coupled to a unit vector

$$S_E = \int d\tau\, \psi^\dagger D\psi, \tag{12.71}$$

where

$$D = \partial_\tau - m\mathbf{n}\cdot\boldsymbol{\tau}, \tag{12.72}$$

with $\mathbf{n} \in S^2$ and $\boldsymbol{\tau}$ the vector of Pauli matrices. We obtain an effective action for $\mathbf{n}$ induced by fermions as

$$e^{-S_{\rm eff}} = \int D\psi D\psi^\dagger e^{-S_E} = \mathrm{Det}\, D, \tag{12.73}$$

or

$$S_{\rm eff} = -\log \mathrm{Det}\, D = -\mathrm{Tr}\, \log D. \tag{12.74}$$

We calculate the variation of $S_{\rm eff}$ with respect to $\mathbf{n}$ as

$$\delta S_{\rm eff} \quad = \quad -\mathrm{Tr}\, \delta D D^{-1} = -\mathrm{Tr}\, \delta D D^\dagger (DD^\dagger)^{-1}, \tag{12.75}$$

where $D^\dagger = -\partial_\tau - m\mathbf{n}\cdot\boldsymbol{\tau}$. We have

$$DD^\dagger \quad = \quad -\partial_\tau^2 + m^2 - m\dot{\mathbf{n}}\cdot\boldsymbol{\tau} = G_0^{-1} - m\dot{\mathbf{n}}\cdot\boldsymbol{\tau}. \tag{12.76}$$

Expand Eq. (12.75) in $1/m$ up to the term m^0 and calculate functional traces. Show that the term of the order m^0 is a variation of the WZ term in 0+1 dimensions. Restore $S_{\rm eff}$ from its variation. What is the coefficient in front of the WZ term? To what value of spin does it correspond?

For the answer see Sec. 12.3.3.

Exercise 10: Fermionic determinant in two dimensions

Let us consider two-dimensional fermions coupled to a phase field $\phi(x)$ ($\phi \equiv \phi + 2\pi$). The Euclidean Lagrangian is given by

$$\mathcal{L}_2 = \bar{\psi}\left[i\gamma^\mu(\partial_\mu - iA_\mu) + ime^{i\gamma^5\phi}\right]\psi, \tag{12.77}$$

where $\mu = 1, 2$ is a spacetime index, $\gamma^{1,2,5}$ is a triplet of Pauli matrices, and A_μ is an external gauge field probing fermionic currents.

We assume that the bosonic field ϕ changes slowly on the scale of the "mass" m. Then one can integrate out fermionic degrees of freedom and obtain an induced effective action for the ϕ-field as a functional determinant.

$$S_{\rm eff} \quad = \quad -\log \mathrm{Det}\, D, \tag{12.78}$$

$$D \quad = \quad i\gamma^\mu(\partial_\mu - iA_\mu) + im\, e^{i\gamma^5\phi}. \tag{12.79}$$

We calculate the effective action using the gradient expansion method. Namely, we calculate the variation of Eq. (12.78) with respect to the ϕ and A-fields and use

$$\delta S_{\text{eff}} = -\delta \log \text{Det}\, D = -\text{Tr}\, \delta \log D = -\text{Tr}\, \delta D\, D^{-1} = -\text{Tr}\, \delta D\, D^{\dagger} (DD^{\dagger})^{-1}. \quad (12.80)$$

a) Calculate $DD^{\dagger}$ for Eq. (12.79). Observe that this object depends only on gradients of ϕ-field.

b) Expand $(DD^{\dagger})^{-1}$ in those gradients. This will be the expansion in $1/m$. (It is convenient to introduce notation $G_0^{-1} = -\partial_\mu^2 + m^2$).

c) Calculate functional traces of the terms up to the order of m^0. Use the plane wave basis to calculate the trace

$$\text{Tr}\,(\hat{X}) \to \int d^2x \int \frac{d^2p}{(2\pi)^2} e^{-i\mathbf{p}\cdot\mathbf{x}} \hat{X} e^{i\mathbf{p}\cdot\mathbf{x}}$$

.

d) Identify the variation of the topological term in the obtained expression. It contains the antisymmetric tensor $\epsilon^{\mu\nu}$ and is proportional to $\text{sgn}\,(m)$.

e) Remove the variation from the obtained expression and find S_{eff} up to the m^0 order.

f) Which terms of the obtained action are topological? Can you write them in terms of differential forms?

For the answer see Ref. [10].

Exercise 11: "Dangers" of chiral rotation

Try to calculate the determinant of the previous exercise using "chiral rotation trick". Namely, consider chiral rotation $\psi \to e^{-i\gamma^5\phi/2}\psi$. Then $\psi^{\dagger} \to \psi^{\dagger} e^{i\gamma^5\phi/2}$ and $\bar{\psi} \to \bar{\psi} e^{-i\gamma^5\phi/2}$. Use the identity $\gamma^\mu\gamma^5 = -i\epsilon^{\mu\nu}\gamma^\nu$ and anti-commutativity of Pauli matrices to show that the operator $D(A_\mu, \phi)$ transforms into

$$\tilde{D}(A_\mu, \phi) = e^{-i\gamma^5\phi/2} D(A_\mu, \phi) e^{-i\gamma^5\phi/2} = D(A_\mu + \frac{i}{2}\epsilon^{\mu\nu}\partial_\nu\phi, 0) = D(\tilde{A}_\mu, 0).$$

Try to calculate $\log \text{Det}\, \tilde{D} = \log \text{Det}\, D(\tilde{A}, 0)$ using expansion in $\tilde{A}$. You will see that the result does not match the effective action obtained in the previous exercise. Why? What one should add to the chiral rotation trick to make the correct calculation?

Answer: The Jacobian of the change of variables corresponding to the chiral rotation. See Refs. [10] and [26].

12.4 Spin chains.

Here we study how topological terms appear in effective theories for quantum spin chains. We emphasize an interplay between different types of topological terms and the effects of topological terms on field dynamics. In addition to original papers, the useful references for this section include: [6, 9, 27].

Let us start with the model of quantum magnet

$$H = \sum_{<kj>} J_{kj} \mathbf{S}_k \cdot \mathbf{S}_j - \sum_j \mathbf{h}_j \cdot \mathbf{S}_j\,. \quad (12.81)$$

Here the summation is taken over the sites k, j of some d-dimensional lattice, J_{kj} are exchange integrals and $\mathbf{h}_j$ is an external (generally space and time-dependent) magnetic field. The quantum spin operators $\mathbf{S}_i$ have $SU(2)$ commutation relations $(a, b, c = 1, 2, 3)$

$$\left[S_j^a, S_k^b\right] = i\delta_{jk}\epsilon^{abc}S_j^c\,. \tag{12.82}$$

Commuting the spin operator S_j^a with the Hamiltonian Eq. (12.81) one obtains Heisenberg equation of motion for the spin operator

$$\partial_t \mathbf{S}_j = i[H, \mathbf{S}_j] = -\sum_k J_{jk}\mathbf{S}_k \times \mathbf{S}_j + \mathbf{h}_j \times \mathbf{S}_j\,. \tag{12.83}$$

12.4.1 Path integral for quantum magnets

Path integral for the magnet on a lattice

The classical action for the magnet Eq. (12.81,12.82) can be written as

$$S = -4\pi i S \sum_j W_0[\mathbf{n}_j] + \int d\tau\, H\,, \tag{12.84}$$

where we introduced classical unit vectors $\mathbf{n}_i$ and summed the terms Eq. (12.63) for each spin. The classical Hamiltonian used in Eq. (12.84) is obtained from Eq. (12.81) substituting $\mathbf{S}_i$ by $S\mathbf{n}_i$. Variation of the action Eq. (12.84) over n_j with the use of Eq. (12.31) produces classical equation of motion

$$-iS\,\partial_\tau \mathbf{n}_j \times \mathbf{n}_j + S^2 \sum_k J_{kj}\mathbf{n}_k - S\sum_i \mathbf{h}_j = 0\,. \tag{12.85}$$

Taking a cross-product with $\mathbf{n}_j$ gives a classical analogue of Eq. (12.83)

$$-i\partial_\tau \mathbf{n}_j = S\sum_k J_{kj}\mathbf{n}_k \times \mathbf{n}_j - \sum_i \mathbf{h}_j \times \mathbf{n}_j\,. \tag{12.86}$$

Remember that $-i\partial_\tau = \partial_t$.

The path integral over trajectories of unit vectors $\mathbf{n}_j(\tau)$ with the amplitude e^{-S} corresponding to the classical action Eq. (12.84) gives the quantization corresponding to Eqs. (12.81,12.82). Our goal now is to find a continuum quantum field theory description of this lattice magnet.

Here important remark is in order. A given lattice theory does not necessarily have a reasonable continuum description. One needs a special reason for continuum approximation to be applicable. Such reasons could be the vicinity to a second order phase transition where correlation length becomes much bigger than the lattice spacing or some other reasons for scale separation. In the following we will try to first derive a continuum limit for the theory and then check the self-consistency of the continuum approximation. Another important point is that the way to take a continuum limit depends crucially on the state of the system. In the following we first consider the ferromagnetic state and then go to the collinear antiferromagnetic state.

Continuum limit for Quantum Ferromagnet

Let us assume for simplicity that $J_{jk} = J < 0$ for nearest neighbor sites j, k of a spin chain in 1d, square lattice in 2d and cubic lattice in 3d and magnetic field is constant $\mathbf{h}_j = \mathbf{h}$. The classical Hamiltonian is then

$$H = -|J|S^2 \sum_{<kj>} \mathbf{n}_k \cdot \mathbf{n}_j - S \sum_j \mathbf{h} \cdot \mathbf{n}_j \,. \tag{12.87}$$

We assume that there is a short range ferromagnetic order, i.e., nearest neighbor spins are almost perfectly aligned. We replace spins $\mathbf{n}_j$ at lattice sites by a continuous field $\mathbf{n}(x)$ and proceed as follows. Up to a constant, $\mathbf{n}_{j+e_x} \cdot \mathbf{n}_j \to -\frac{1}{2}(\mathbf{n}_{j+e_x} - \mathbf{n}_j)^2 \to -\frac{1}{2}a^2(\partial_x \mathbf{n})^2$ etc. Here a is the lattice constant. Replacing the summation over j by the integration over space, we obtain the continuum limit of the Hamiltonian Eq. (12.87)

$$H = -\frac{1}{2}|J|S^2 a^2 \int \frac{d^d x}{a^d} (\partial_\mu \mathbf{n})^2 - S \int \frac{d^d x}{a^d} \mathbf{h} \cdot \mathbf{n} \,. \tag{12.88}$$

Similarly, we have for the full action Eq. (12.84)

$$\begin{aligned} S[\mathbf{n}] \;\; &= \;\; -4\pi i S \int \frac{d^d x}{a^d} W_0[\mathbf{n}(x,\tau)] \\ &\quad -\frac{1}{2}|J|S^2 a^2 \int d\tau \frac{d^d x}{a^d} (\partial_\mu \mathbf{n})^2 - S \int d\tau \frac{d^d x}{a^d} \mathbf{h} \cdot \mathbf{n} \,. \end{aligned} \tag{12.89}$$

Variation of this action with respect to the continuous unit vector field $\mathbf{n}(x, \tau)$ produces the well-known classical Landau-Lifshitz equation for magnetization (to go to real time one should replace $i\partial_\tau \to -\partial_t$)

$$i\partial_\tau \mathbf{n} = |J|Sa^2 \, (\mathbf{n} \times \Delta \mathbf{n}) - \mathbf{h} \times \mathbf{n} \,. \tag{12.90}$$

Let us remark here that $W_0[\mathbf{n}_j]$ is a topological term for an individual spin $\mathbf{n}_j$ on the site j of the lattice. However, due to the integration over space, the first term of the continuum action Eq. (12.89) depends on the spatial metric (e.g., distortions of the lattice will change it). Therefore, this term cannot be considered topological. Nevertheless, it is linear in time derivative and therefore time-reparameterization invariant. Therefore, it remains imaginary after Wick's rotation and results in a very essential interference even in imaginary time path integral.

Bloch's law: the dispersion of spin waves in ferromagnet

As an application of the continuum theory for magnetization in ferromagnets, let us derive the dispersion of spin waves starting from Eq. (12.90). We assume that magnetic field is constant and uniform $\mathbf{h} = (0, 0, h)$, and that there is a long range ferromagnetic order with spins oriented in the same direction. We

also assume that spin fluctuations are small and write $\mathbf{n} = (u_1, u_2, 1)$, where $u_{1,2}$ are components of $\mathbf{n}$ in x, y directions in spin space that are assumed to be small so that $\mathbf{n}^2 = 1 + u_1^2 + u_2^2 \approx 1$ up to quadratic terms in u. Substituting all this in Eq. (12.90) we obtain

$$-i\omega u = -i|J|Sa^2\mathbf{k}^2 u - ihu\,. \tag{12.91}$$

Here we introduced complex notation $u = u_1 + iu_2$ and made Fourier transform $\Delta \to -\mathbf{k}^2$ and $i\partial_\tau \to -\partial_t \to i\omega$. We immediately obtain the dispersion of spin waves

$$\omega = |J|Sa^2\mathbf{k}^2 + h, \tag{12.92}$$

the result known as the Bloch's law. In the absence of external magnetic field the dispersion of spin waves is quadratic in wave vector k.

To conclude our brief discussion of the ferromagnetic case we have to recall that the continuum theory was derived under the condition that fluctuations of $\mathbf{n}$ are small compared to 1 or $|u| \ll 1$. Given a temperature and other parameters of the theory, one should calculate the average value of those fluctuations. The condition $\langle |u|^2 \rangle \ll 1$ is then the necessary condition for the self-consistency of the continuum approximation.

12.4.2 Continuum path integral for Quantum Antiferromagnet

Let us consider a more subtle case of quantum antiferromagnet. We again start with the Hamiltonian Eq. (12.87). However, we assume now that $J > 0$ and write

$$H = JS^2 \sum_{<kj>} \mathbf{n}_k \cdot \mathbf{n}_j - S\sum_j \mathbf{h} \cdot \mathbf{n}_j\,. \tag{12.93}$$

Although Eq. (12.93) looks very similar to Eq. (12.87) the unit vectors $\mathbf{n}_j$ tend to be antiparallel on nearest sites (we again assume square lattice here so that the antiferromagnetic order is not frustrated). One cannot use the continuous field $\mathbf{n}(x)$ instead of lattice vectors $\mathbf{n}_j$. Taking continuum limit is more involved and can be achieved through the following substitution

$$\mathbf{n}_j = (-1)^j \mathbf{m}(x) + a\mathbf{l}(x)\,. \tag{12.94}$$

Here we assume that both fields $\mathbf{m}(x)$ and $\mathbf{l}(x)$ are good continuous (smooth) fields. [12] The former represents the smooth staggered magnetization while the latter is a ferromagnetic component. It is expected that the ferromagnetic component is small and the corresponding rescaling by the lattice constant a is made. As $\mathbf{n}_j^2 = 1$ we have

$$\mathbf{n}_j^2 = \mathbf{m}^2 + 2(-1)^j a(\mathbf{m} \cdot \mathbf{l}) + a^2\mathbf{l}^2 = 1\,. \tag{12.95}$$

[12] We emphasize that in the following we take a particular continuum limit which assumes short range ordered antiferromagnetic state. It is believed to be appropriate for large S Heisenberg antiferromagnets. However, it is not appropriate, e.g. for spin chains at so-called Bethe Ansatz integrable points.

We solve this condition to the order of a^2 by two conditions

$$\mathbf{m}^2 = 1\,, \qquad \mathbf{m}\cdot\mathbf{l} = 0\,. \tag{12.96}$$

Using Eq. (12.94) we have up to constants

$$\begin{aligned} \mathbf{n}_{j+e_x}\cdot\mathbf{n}_j &\to \frac{1}{2}a^2\left[(\partial_x\mathbf{m})^2 + 4\mathbf{l}^2 + 4(-1)^j\partial_x\mathbf{m}\cdot\mathbf{l}\right]\,, \\ \mathbf{h}\cdot\mathbf{n}_j &\to (-1)^j\mathbf{h}\cdot\mathbf{m} + a\mathbf{h}\cdot\mathbf{l}\,. \end{aligned}$$

Substituting these expressions into Eq. (12.93) we obtain

$$\begin{aligned} H &= JS^2\sum_j\frac{1}{2}a^2\left[(\partial_\mu\mathbf{m})^2 + 4d\mathbf{l}^2 + 4(-1)^j\partial_\mu\mathbf{m}\cdot\mathbf{l}\right] \\ &\quad -S\sum_j((-1)^j\mathbf{h}\cdot\mathbf{m} + a\mathbf{h}\cdot\mathbf{l})\,. \\ &\to JS^2a^2\frac{1}{2}\int\frac{d^dx}{a^d}\left[(\partial_\mu\mathbf{m})^2 + 4d\mathbf{l}^2\right] - Sa\int\frac{d^dx}{a^d}\mathbf{h}\cdot\mathbf{l}\,. \end{aligned} \tag{12.97}$$

In the last step, we dropped all oscillating terms and replaced summation by integration over space.

The next step is to do a similar procedure with the term in the action coming from the summation of topological terms. We proceed as follows

$$\begin{aligned} \sum_j W_0[\mathbf{n}_j] &= \sum_j W_0[(-1)^j\mathbf{m}(x) + a\mathbf{l}(x)] \\ &\approx \sum_j(-1)^j W_0[\mathbf{m}(x) + (-1)^j a\mathbf{l}(x)] \\ &\approx \sum_j(-1)^j W_0[\mathbf{m}(x)] + \int\frac{d^dx}{a^d}\int d\tau\, a\mathbf{l}(x)\frac{\delta W_0[\mathbf{m}]}{\delta\mathbf{m}} \end{aligned}$$

We now use the variation formula Eq. (12.31) and its consequence

$$W_0[\mathbf{m}(x+e_x)] - W_0[\mathbf{m}(x)] \approx \frac{1}{4\pi}\int d\tau\,(a\partial_x\mathbf{m})\cdot(\mathbf{m}\times\partial_\tau\mathbf{m})$$

and obtain

$$\begin{aligned} \sum_j W_0[\mathbf{n}_j] \approx\ & \frac{1}{4\pi}\int\frac{d^dx}{a^d}\int d\tau\, a d\,\mathbf{l}(x)\cdot(\mathbf{m}\times\partial_\tau\mathbf{m}) \\ &+\sum_{j_y,j_z}(-1)^{j_y+j_z}\frac{1}{8\pi}\int\frac{dx}{a}\int d\tau\,(a\partial_x\mathbf{m})\cdot(\mathbf{m}\times\partial_\tau\mathbf{m})\,. \end{aligned} \tag{12.98}$$

The first term of Eq. (12.98) is written for any spatial dimension d. In the second term, we assumed the three-dimensional case. Notice that while the

sign alternation was taken into account in x direction there is still a sum to be taken with the factor $(-1)^{j_y+j_z}$ in other two directions. That summation will suppress this term and, therefore, it is relevant only in one spatial dimension. We summarize for the topological contribution

$$-4\pi i S \sum_j W_0[\mathbf{n}_j] \approx -iSa^{1-d}d \int d\tau\, d^d x\, \mathbf{l}(x)\cdot(\mathbf{m}\times\partial_\tau \mathbf{m}) \\ -i\frac{S}{2}\delta_{d,1}\int d\tau\, dx\, \partial_x \mathbf{m}\cdot(\mathbf{m}\times\partial_\tau\mathbf{m})\,. \quad (12.99)$$

Collecting all terms together to get the continuum limit of the action Eq. (12.84) we obtain

$$\begin{aligned} S[\mathbf{m},\mathbf{l}] &= i\frac{S}{2}\delta_{d,1}\int d\tau\, dx\, \partial_x\mathbf{m}\cdot(\mathbf{m}\times\partial_\tau\mathbf{m}) \\ &\quad -iSa^{1-d}d\int d\tau\, d^d x\, \mathbf{l}(x)\cdot(\mathbf{m}\times\partial_\tau\mathbf{m}) \\ &\quad +JS^2a^{2-d}\frac{1}{2}\int d\tau\, d^d x\, \left[(\partial_\mu\mathbf{m})^2+4d\mathbf{l}^2\right] - Sa^{1-d}\int d\tau\, d^d x\, \mathbf{h}\cdot\mathbf{l} \\ &= i\frac{S}{2}\delta_{d,1}\int d\tau\, dx\, \partial_x\mathbf{m}\cdot(\mathbf{m}\times\partial_\tau\mathbf{m}) + JS^2a^{2-d}\frac{1}{2}\int d\tau\, d^d x\,(\partial_\mu\mathbf{m})^2 \\ &\quad +\int d\tau\,\frac{d^d x}{a^d}\,\Big(2JS^2a^2d\mathbf{l}^2 - Sa^1\mathbf{l}\cdot\Big[\mathbf{h}+id\,(\mathbf{m}\times\partial_\tau\mathbf{m})\Big]\Big). \quad (12.100) \end{aligned}$$

The obtained expression is the continuum limit of Eq. (12.84) derived in the antiferromagnetic regime with the assumption of small fluctuations around the short range collinear antiferromagnetic order. The field $\mathbf{l}$ describing the magnetization of the magnet enters the action in a very simple way and can be "integrated out". For details of derivation see the Appendix B. Here we present the results dropping the external magnetic field for simplicity. In two and higher spacial dimensions, $d>1$, we have

$$S[\mathbf{m}] = \frac{1}{2g}\int d\tau\,\frac{d^d x}{a^{d-1}}\,\left[\frac{1}{v_s}(\partial_\tau\mathbf{m})^2 + v_s(\partial_\mu\mathbf{m})^2\right], \quad (12.101)$$

where

$$v_s = \frac{2JSa}{\sqrt{d}}\,, \qquad g=\frac{2}{S\sqrt{d}}\,. \quad (12.102)$$

The one-dimensional case is special and has an additional topological term in the action

$$S[\mathbf{m}] = \frac{1}{2g}\int d\tau\, dx\,\left[\frac{1}{v_s}(\partial_\tau\mathbf{m})^2 + v_s(\partial_x\mathbf{m})^2\right] \\ +i\theta\int d\tau\, dx\,\frac{1}{4\pi}\mathbf{m}\cdot(\partial_\tau\mathbf{m}\times\partial_x\mathbf{m})\,, \quad (12.103)$$

where

$$v_s = 2SaJ\,, \qquad g = \frac{2}{S}\,, \qquad \theta = 2\pi S\,. \tag{12.104}$$

The model, Eq. (12.103), is known as $O(3)$ nonlinear sigma model with topological theta-term. It is the low energy, long distance description of the antiferromagnetic Heisenberg spin chain with large spin $S \gg 1$ with the correspondence between parameters of the model and the parameters of the spin chain given by Eq. (12.104). Let us start with the discussion of the nonlinear sigma model without topological term.

12.4.3 RG for $O(3)$ NLSM

The model, Eq. (12.103), without topological term ($\theta = 0$) can be re-written as

$$S[\mathbf{m}] \;=\; \frac{1}{2g}\int d^2x\,(\partial_\mu \mathbf{m})^2\,, \tag{12.105}$$

where $\mu = \tau, x$ and we re-defined $\tau \to \tau/v_s$. The action Eq. (12.105) with the constraint $\mathbf{m}^2 = 1$ is known as $O(3)$ nonlinear sigma model (NLSM). It is relativistically invariant with spin wave velocity v_s playing the role of the speed of light. This relativistic invariance is emergent and we should remember that the next order gradient corrections to the model and various perturbations are generally not relativistically invariant.

At small values of the coupling constant g corresponding to large values of S one can treat Eq. (12.105) perturbatively and ask how the coupling constant renormalizes when one goes to longer distances. It turns out [28] that g increases with the scale. The increase of g signals the tendency of the $\mathbf{m}$-field to disorder. More precisely, the effective coupling of Eq. (12.105) at the length L satisfies renormalization group (RG) equation

$$\frac{dg}{d\log L} = \frac{1}{2\pi}g^2 + O(g^3), \tag{12.106}$$

and gives

$$g(L) = \frac{g_0}{1 - \frac{g_0}{2\pi}\log(L/a)}\,, \tag{12.107}$$

where $g_0 = g(a)$ is the coupling constant at UV (lattice) scale a. At the scale $L \sim \xi$ with

$$\xi \sim a e^{2\pi/g_0}, \tag{12.108}$$

the effective coupling constant $g(\xi)$ becomes of the order of unity and we cannot trust RG equation Eq. (12.106) at this point.

We stress here that RG analysis is not conclusive. The only conclusion we can make is that the effective length ξ given by Eq. (12.108) emerges. At this scale, the $\mathbf{m}$ field is somewhat disordered, but we cannot say anything about the nature of the phase and about the long distance behavior of $\mathbf{m}$-field correlation functions. There are essentially two scenarios. The first one is that

the actual model has a gap of the order of v_s/ξ, the field $\mathbf{m}$ is disordered with all correlations decaying exponentially with the correlation length ξ Eq. (12.108). The second possibility is that the RG flow leads to a new fixed point and behavior of the model at scales larger than ξ is governed by that fixed point (in particular, long range correlation functions might decay as power laws etc.). It turns out that this is the former scenario that is realized for 2d $O(3)$ nonlinear sigma model Eq. (12.105). We know this because $O(3)$ NLSM has been solved exactly by Bethe Ansatz [29] and has a gap separating the ground state from excitations. In the next section we will argue that the second scenario might be relevant when topological term is present in NLSM.

Bethe Ansatz solution of the model Eq. (12.105) is outside of the scope of these lectures. Instead, to have some understanding of how finite gap (correlation length) appears in NLSM we refer the reader to the Exercise 13 "$O(N)$ NLSM" below where the correlation length is obtained for the $O(N)$ NLSM in the limit of large N.

12.4.4 $O(3)$ NLSM with topological term

Let us now consider the $O(3)$ NLSM with topological theta term

$$S[\mathbf{m}] \;=\; \frac{1}{2g}\int d^2x\,(\partial_\mu \mathbf{m})^2 + i\theta Q\,, \tag{12.109}$$

where

$$Q \;=\; \int d^2x\,\frac{1}{4\pi}\mathbf{m}\cdot(\partial_\tau \mathbf{m}\times\partial_x \mathbf{m})\,. \tag{12.110}$$

We assume that the boundary conditions $\mathbf{m}(x)\to\mathbf{m}_0$ =const, as $x\to\infty$ so that the winding number Q is an integer. The parameters corresponding to the AFM spin chain are given by Eqs. (12.103,12.104), i.e., $g=2/S$ and $\theta=2\pi S$. Following Haldane [30,31], we notice that the topological term in Eq. (12.109) contributes the complex weight to path integral given by $e^{i\theta Q}=(-1)^{2SQ}$. This weight depends crucially on the integer-valuedness of spin. If the spin S is half-integer, the weight is non-trivial $(-1)^Q$, and results in interference of topological sectors characterized by different topological charges Q. On the other hand, if S is integer, the weight is unity and does not affect the path integral. [13] Based on this observation Haldane conjectured that AFM spin chains with integer spin S have singlet ground states separated by finite gap from all excitations similar to $O(3)$ NLSM without topological term. On the other hand, AFM spin chains with half-integer spin have gapless excitations similar to the spin-1/2 chain. For the latter, the spectrum of excitations has been known from the exact solution by Bethe [32]. The Haldane's conjecture has been supported by numerical simulations and experiments.

[13] This statement is not quite correct. The topological term can still be important for various boundary conditions and due to the presence of singularities.

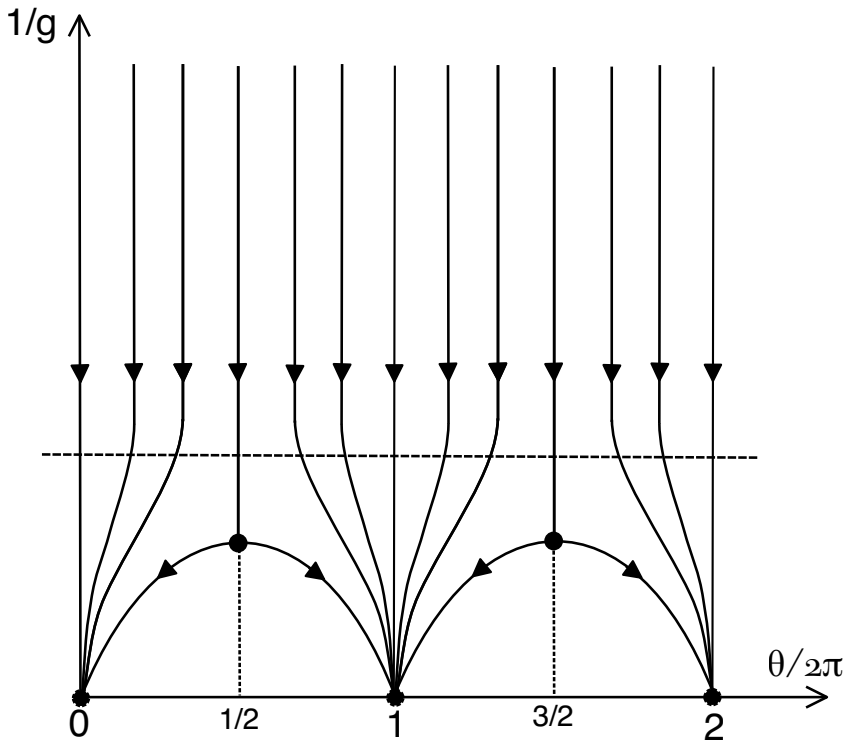


Figure 12.4: The schematic RG flow for nonlinear sigma model with theta term is shown. The vertical axis is the inverse coupling constant g while the horizontal axis is the topological angle θ. The picture has exact symmetry with respect to $\theta \to \theta + 2\pi$ and with respect to reflection relative to half integer $\theta/2\pi$. The horizontal dashed line corresponds to $g \sim 1$. Below this line the perturbative RG is not working and interpretation of the flow is more subtle.

It is instructive to think about RG flow for the NLSM with theta term. The model has two parameters g and θ and it is appealing to think about RG flow in the plane labeled by those parameters. First of all, we notice that starting with dimerized spin chain one obtains values of θ which are not necessarily multiples of π (see Exercise 12 "Dimerized spin chain" below). Therefore, it is tempting to draw the flow diagram similar to the one for integer quantum Hall effect [33, 34] (see figure 12.4).

It is not difficult to see that conventional perturbative calculation results in non-renormalizability of θ. This is reflected by strictly vertical flow at small g (large $1/g$) in figure 12.4. One can argue following [33] that taking into ac-

count instanton configurations (configurations with $Q \neq 0$) will result in the deviations of flow from the vertical one as shown in the figure. However, the instanton contribution is suppressed by factors of the order of $e^{-2\pi/g}$ and it is very hard to develop consistent perturbation theory taking into account exponentially small terms but neglecting terms of higher order in coupling g. More straightforward interpretation of figure 12.4 is that it shows the flow of some observables which correspond to $1/g$ and θ at small values of g and change with scale as shown in the figure (in analogy with observable conductances σ_{xx} and σ_{xy} of [33]). We will not dwell on the interpretation of figure Eq. (12.4). However, it is believed that as shown on that figure the long distance behavior of spin chains of half-integer $\theta/(2\pi)$ is governed by new infrared fixed points while the spin chain with any other θ flows to the models with finite gap in the spectrum. In fact, it was clarified that the critical description of half-integer spin chains is given by Wess-Zumino-Witten (WZW) model (see Ref. [35]). In particular, the staggered magnetization correlation decays as power law at large distances.

In the limit of large S, we have the following picture for the correlator $\langle \mathbf{m}(x) \cdot \mathbf{m}(0) \rangle$. At distances shorter than $\xi \sim a e^{\pi S}$ (obtained form Eq. (12.108) at $g = 2/S$) there is a short range antiferromagnetic order $\langle \mathbf{m}(x) \cdot \mathbf{m}(0) \rangle \sim 1$. At larger distances $x \gg \xi$, the order is destroyed, and there is no long-range order in agreement with the Mermin-Wagner theorem. However, the way in which correlations decay at large distances depend crucially on S. If S is integer, the decay is exponential $\langle \mathbf{m}(x) \cdot \mathbf{m}(0) \rangle \sim e^{-x/\xi}$, while at half-integer S, it is algebraic (power-law) with $\langle \mathbf{m}(x) \cdot \mathbf{m}(0) \rangle \sim x^{-1}$. [14] This power-law signals the emergent scale invariance (in fact, conformal invariance) and is captured by the infrared fixed point of effective WZW model. Interestingly, for half-integer spins the symmetry of the spin chain is enlarged at the critical point. In addition to the fluctuations of the staggered magnetization $\mathbf{m}(x)$, fluctuations of dimerization become soft as well. They combine into the effective $SU(2)$ "order parameter" which is an $SU(2)$ degree of freedom of WZW model. We refer the reader to [35] for more detailed analysis of critical behavior of spin chains.

12.4.5 Boundary states for spin 1 chains with Haldane's gap

We argued at the beginning of Sec. 12.4.4 that for integer S the topological term is "ineffective" and the model Eq. (12.109) behaves similarly to the $O(3)$ NLSM without topological term Eq. (12.105). This is not quite correct. The argument is essentially based on the integer-valuedness of the topological charge Q. However, it is necessarily integer only for smooth $\mathbf{m}$-field configurations for compact boundary conditions (e.g., $\mathbf{m} \to \mathbf{m}_0$ at $x, t \to \infty$). Let us consider an example when boundary conditions are different.

[14] At non-zero temperature there is another length scale $\xi_T \sim aJS^2/T$, and the power-law decay of correlations will be eventually replaced by an exponential decay at large distances due to thermal fluctuations.

We assume that the spin chain is long but has a finite length L with free boundary conditions (boundary spins at $x = 0, L$ can take arbitrary values). Then the topological term can be written as

$$\begin{aligned} S_{top}[\mathbf{m}] &= i\theta \int d\tau \int_0^L dx \, \frac{1}{4\pi} \mathbf{m} \cdot (\partial_\tau \mathbf{m} \times \partial_x \mathbf{m}) \\ &= i\theta \left[\frac{\Omega_L}{4\pi} + Q - \frac{\Omega_0}{4\pi} \right] . \end{aligned} \tag{12.111}$$

Here we decomposed the winding number Eq. (12.110) into integer part Q and the difference of solid angles Ω_0, Ω_L subtended by boundary vectors $\mathbf{m}(0)$ and $\mathbf{m}(L)$, respectively. Assuming $\theta = 2\pi$, we drop the integer part and end up with the contribution

$$S_{top} = i \left[\frac{\Omega_L}{2} - \frac{\Omega_0}{2} \right] , \tag{12.112}$$

which is written in terms of boundary vectors only. Comparing with Eq. (12.29) we recognize Eq. (12.112) as the action of two spin-1/2 located at the ends of the spin chain. As we expect that the bulk degrees of freedom have a gap in the spectrum ($S = 1$ is an integer) we conclude that the $S = 1$ AFM spin chain in the gapful phase (Haldane's phase) should have gapless boundary spin 1/2 excitations.

We would like to stress again how unusual is the conclusion we have just made. It is well known that "adding" finite number of spin 1s one can get only superposition of integer valued spins. For example, two spin-1 particles can only have sectors with total spins $0, 1$, and 2. However, we managed adding large number of spin 1s to get two boundary spin 1/2 as low lying excitations of the finite spin chain with free boundary conditions! [15] Most amazingly, these boundary spin 1/2 have been observed in experiment. [36, 37]

This example illustrates very interesting connection between θ and WZW topological terms. The gapful model with θ terms might produce massless boundary theory of one dimension lower described with the use of WZ terms.

12.4.6 AKLT model

To understand better how the bulk gap and boundary spin 1/2 states are formed in the Haldane's phase of $S = 1$ antiferromagnetic spin chain let us consider the deformation of the Heisenberg Hamiltonian known as AKLT model [38, 39]

$$H = J \sum_j \left(\mathbf{S}_j \cdot \mathbf{S}_{j+1} + \frac{1}{3} (\mathbf{S}_j \cdot \mathbf{S}_{j+1})^2 \right) . \tag{12.113}$$

[15] Actually, for the finite length of the spin chain those two boundary spin 1/2 are effectively interacting with the strength of interaction $\sim e^{-L/\xi}$. Only in the limit of an infinite length we obtain truly non-interacting spin 1/2 degrees of freedom.

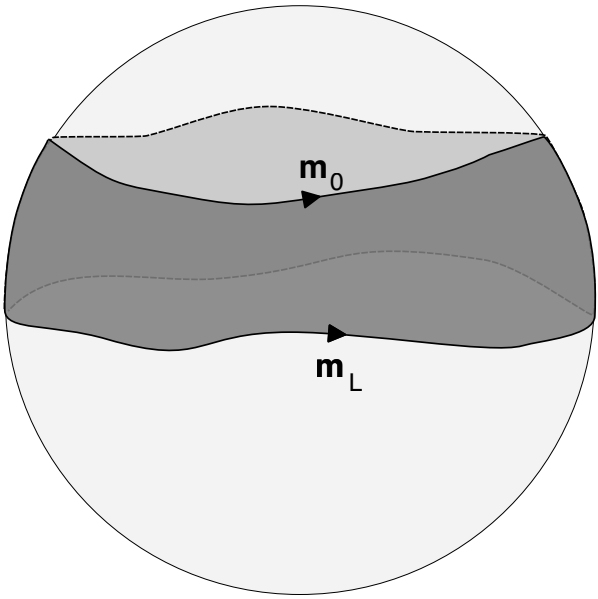


Figure 12.5: Time evolution of boundary magnetizations (unit vectors) $\mathbf{m}_0(\tau)$ and $\mathbf{m}_L(\tau)$ is shown. For periodic time these vectors sweep solid angles Ω_0 and Ω_L, respectively. The shaded area is given by Eq. (12.111) which is up to integer Q is given by the difference of solid angles Ω_L and Ω_0 as in Eq. (12.112).

One should think about this model as fine tuned model with the coefficient J' in front of the biquadratic term being exactly 1/3 of the coefficient in front of Heisenberg exchange term. Changing J' from 0 to $J/3$ interpolates between Heisenberg spin chain and AKLT model Eq. (12.113). It can be shown that the ground state of Eq. (12.113) is separated by the finite gap from excitations. If this gap is not closed in the process of changing J' from $J/3$ to zero then the Heisenberg spin chain has a gap as well and we say that AKLT and Heisenberg model are adiabatically connected and are in the same phase.

Let us now see why AKLT model is much easier to analyze than the Heisenberg model for $S = 1$. It turns out that the Hamiltonian Eq. (12.113) up to constant can be written as

$$H = 2J \sum_j P_2(\mathbf{S}_j + \mathbf{S}_{j+1}) \,, \tag{12.114}$$

where

$$P_2(\mathbf{S}_1 + \mathbf{S}_2) = \frac{1}{2}\mathbf{S}_j \cdot \mathbf{S}_{j+1} + \frac{1}{6}(\mathbf{S}_j \cdot \mathbf{S}_{j+1})^2 + \frac{1}{3} \tag{12.115}$$

is a projector on the state with the total spin $S = 2$ (see the Exercise 15 "Projector to $S = 2$"). Then schematically (see Refs. [38, 39] for details) one

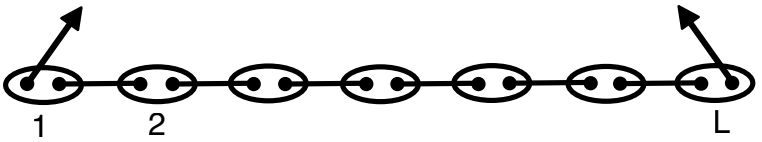


Figure 12.6: The schematic picture of the ground state of AKLT model. Ovals represent the sites of the spin chain. Spin 1 at each site is split in two spin 1/2s. The horizontal segments represent singlet states of spin 1/2-s from neighbor sites. Notice that one of spin 1/2-s at the site 1 and at the site L are not parts of singlets. They are the boundary spin 1/2 states of the AKLT spin 1 chain.

can think about the ground and excited states of Eq. (12.114) in the following way. We split every spin S_j into two spins 1/2 (let us call them A and B. Taking one of those two spin 1/2 from the site j and one of spin 1/2 from the site $j+1$ we form a singlet state and make a product of all these singlets. Then we project the result back to the total spin 1 on each site. The obtained state

$$\mathcal{P}\prod_j \Big(|+\rangle_{j,A}|-\rangle_{j+1,B} - |-\rangle_{j,A}|+\rangle_{j+1,B}\Big)$$

has a property that the sum of two neighbor spins is never 2 and is, therefore, the ground state of Eq. (12.114) and of Eq. (12.113). [16] It is also physically clear that to excite this system one should break one of the singlets and pay finite energy. Schematically we illustrate the construction in figure 12.6. In particular, one can see that the boundary spin 1/2 in this case are just uncoupled "halves" of the end spins of the chain. In some sense AKLT provides a "chemical" model with singlets being covalent bonds and spin 1/2-s being physical electrons. We refer to AKLT model as to a strong coupling limit of the Haldane's phase of a spin-1 chain.

12.4.7 Exercises

Exercise 12: Dimerized spin chain

Start with the spin-chain Hamiltonian

$$H = \sum_k J_k\, \mathbf{S}_k \cdot \mathbf{S}_{k+1}$$

with $J_k = J > 0$ if k even and $J_k = J' > 0$ if k odd. Repeat the derivation of the section 12.4.2 and show that the continuum limit is still given by $O(3)$ NLSM with

[16] Of course, these speculations are not rigorous at all. For rigorous treatment see Ref. [38,39] or modern treatment using the technique of matrix product states (MPS).

theta term Eq. (12.103) but instead of Eq. (12.104) the parameters of the model are given by

$$v_s = 2Sa\sqrt{JJ'}\,, \qquad g = \frac{1}{S}\left(\sqrt{\frac{J}{J'}} + \sqrt{\frac{J'}{J}}\right)\,, \qquad \theta = 2\pi S\left[1 - \frac{J-J'}{J+J'}\right]\,. \quad (12.116)$$

Notice that the model is self-dual with respect to the duality transformation $J \leftrightarrow J'$, $\theta \to 4\pi S - \theta$ corresponding to the reflection symmetry of the 1d lattice and the Hamiltonian.

Exercise 13: $O(N)$ NLSM

Let us consider the model given by the action Eq. (12.105), where the unit vector field $\mathbf{m}^2 = 1$ has N components $\mathbf{m} = (m_1, m_2, \ldots, m_N)$. This model is known as $O(N)$ nonlinear sigma model. We replace Eq. (12.105) by

$$S[\mathbf{m}, \lambda] \quad = \quad \frac{1}{2g_0}\int d^2x\, \left[(\partial_\mu \mathbf{m})^2 + i\lambda(\mathbf{m}^2 - 1)\right]\,. \quad (12.117)$$

We introduce here the field $\lambda(x,t)$ so that functional integration over it gives the local constraint $\mathbf{m}^2 = 1$. Assuming that the functional integral over $\lambda(x,t)$ is performed exactly one should not worry about the constraint in Eq. (12.117). Let us now make an approximation assuming that the functional integral corresponding to Eq. (12.117) is dominated by the saddle point $i\lambda = M^2 = const$. Then the path integral over $\mathbf{m}$ is Gaussian.[17]

Compute this Gaussian integral and find $\langle \mathbf{m}^2 \rangle$ for a given M. Assume that all divergent integrals can be cut off by the lattice scale a. Write down the consistency equation $\langle \mathbf{m}^2 \rangle = 1$. This is the so-called *gap equation* which determined the saddle point value of M. Show that it is given by

$$M \sim \frac{1}{a}\exp\Big(-\frac{4\pi}{Ng_0}\Big)\,. \quad (12.118)$$

Consider now fluctuations of λ around the saddle point, and show that these fluctuations are suppressed by the parameter $1/N$, i.e., one needs large N limit to make saddle point approximation self-consistent.

As a result of this exercise we obtained that the correlation functions of $\mathbf{m}$-field behave as the ones for the field with the mass M given by Eq. (12.118), i.e., $\langle \mathbf{m}(x) \cdot \mathbf{m}(0) \rangle \sim \exp(-Mx)$.

For the answer to this exercise see Ref. [6].

[17] At this step we essentially replace the local constraint $\mathbf{m}^2 = 1$ for the global one $\int d^2x\, \mathbf{m}^2 = const$ and hope that fluctuations of local magnitude of $\mathbf{m}$ are not important.

Exercise 14: Boundary spin 1/2 states of a Haldane's chain

Consider the "action" of a two-dimensional $O(3)$ non-linear sigma model with topological term

$$S = S_{NLSM} + S_\theta \,, \tag{12.119}$$

$$S_{NLSM} = \frac{1}{2g}\int d^2x \left(\partial_\mu \mathbf{n}\right)^2 , \tag{12.120}$$

$$S_\theta = i\theta Q \,, \tag{12.121}$$

$$Q = \int d^2x \, \frac{1}{8\pi}\epsilon^{\mu\nu}\mathbf{n}\cdot\left[\partial_\mu \mathbf{n}\times\partial_\nu \mathbf{n}\right]. \tag{12.122}$$

This action can be derived as a continuum limit of the Heisenberg spin chain with large spins on sites. In the latter case, $g = 2/S$ and $\theta = 2\pi S$. In the case of integer S, the spin chain is massive, and there are no bulk excitations at low energies (smaller than the gap).

Let us assume that the action Eq. (12.119) is defined on the open chain of the length L. Show that the topological theta-term Eq. (12.121,12.122) formally defined on the open chain reduces to two WZ (0+1-dimensional) terms at the boundary of spacetime, i.e., at the ends of the spin chain. This means that we expect two quantum spins living at the ends of the spin chain. Show that the coefficient in front corresponds to the value of those spins $S/2$. In particular, it means that the boundary states of $S = 1$ spin chain correspond to spin-1/2!

Remark: Neglecting the NLSM part of the action is possible in this exercise only because of the gap in the bulk at the integer value of spin.

For the answer see Sec. 12.4.5.

Exercise 15: Projector to $S = 2$

Show that the operator Eq. (12.115) projects any state of two spins $S = 1$ onto the state with total spin $S = 2$.

Hint: use the identity $(\mathbf{S}_1 \cdot \mathbf{S}_2) = \frac{1}{2}\left[(\mathbf{S}_1 + \mathbf{S_2})^2 - 4\right]$ for two spin-1 operators $\mathbf{S}_1^2 = \mathbf{S}_2^2 = S(S+1) = 2$.

12.5 Conclusion

In these lectures, we considered a few examples of topological terms that appear in effective actions used in condensed matter theory. We discussed some of the properties of those terms and their physical consequences. We also showed how these terms can be generated by fermionic degrees of freedom as phases of fermionic determinants. Of course, these lectures can serve only as an introduction to a quickly developing field of topological phases of matter. Because of limited time we focused on topological terms in sigma-models and did not consider the ones made out of background and dynamic gauge fields. The latter are related to the physics of quantum Hall effect (see Ref. [40] for some recent developments) and to the physics of topological insulators and superconductors (e.g., see Ref. [41]). Other interesting but much less understood topics include topological terms for singular processes (defects in space-time configurations,

such as monopoles, hedgehogs and vortices), the role of topology for phases with gapless fermions, topology and physics out of equilibrium etc. My hope is that these lectures were stimulating enough to encourage the reader to study and to work on applications of topology to physics.

12.6 Acknowledgements

I was inspired to work on topological terms in condensed matter physics by Paul Wiegmann. I have learned most of what is written in these lectures working with Paul and I am infinitely grateful to him for this experience. I would also like to thank Patrick Lee who encouraged my first attempt to lecture on topological terms at MIT. Some of these lectures were shaped in the "Topics" course on topological terms for Stony Brook University's physics graduate students. I thank all students and faculty attending those lectures for their attention and questions and to Paul Wiegmann and Artem Abanov for reading the manuscript and providing a very useful feedback. Finally, these notes would never be finished without Somen Bhattacharjee who invited me to SERC School on Topology and Condensed Matter Physics (2015, Kolkata, India) to give these lectures and gently but constantly reminded me to work on the notes.

Appendix A: Topological defects and textures

In this appendix we collect few exercises on the topic of homotopy classifications of topological defects and textures in media with spontaneously broken symmetry. A detailed exposure of the topic could be found in the classic reference [7]. Some of known homotopy groups for various spaces are collected in Appendix C.

Exercise 16: Nematic

Nematic is a liquid crystal characterized by an order parameter which is the unit three-component vector $\mathbf{n} = (n_1, n_2, n_3)$, $\mathbf{n}^2 = 1$ with an additional condition $\mathbf{n} \sim -\mathbf{n}$. The latter means that two unit vectors which are opposite to each other describe the same state. Such an order parameter is called "director".

What are the types of topological defects and textures allowed for three-dimensional nematic? What about two-dimensional one?

Exercise 17: Crystal

One can view a crystalline state as a continuous translational symmetry broken to the subgroup of discrete translations. Then the order parameter space should be identified (for three-dimensional crystal) with $M = G/H = \mathbb{R}^3/(\mathbb{Z} \times \mathbb{Z} \times \mathbb{Z})$.

a) What (geometrically) is the order parameter space for this system?

b) What are the homotopy groups of this manifold $\pi_{0,1,2,3}(M)$?

c) What types of topological defects and textures are allowed in such a system?

Exercise 18: Superfluid $^3He - A$

The order parameter of superfluid $^3He - A$ can be represented by two mutually orthogonal unit vectors $\mathbf{\Delta}_1$, $\mathbf{\Delta}_2$. That is, at each point in three-dimensional space one has a pair of vectors with properties $\mathbf{\Delta}_1^2 = \mathbf{\Delta}_2^2 = 1$ and $\mathbf{\Delta}_1 \cdot \mathbf{\Delta}_2 = 0$.

a) What is the order parameter space for this system?

b) What are the homotopy groups of this manifold $\pi_{0,1,2,3}(M)$?

c) What are the types of topological defects and textures allowed in such a system?

Exercise 19: Heisenberg model

What topological defects and textures should one expect in the ordered state of a three-dimensional classical Heisenberg model? What changes if the order parameter is a director instead of a vector? A "director" means a vector without an arrow, i.e., one should identify $\mathbf{S} \equiv -\mathbf{S}$. The models with a director as an order parameter are used to describe nematic liquid crystals (see Ex. 16).

Exercise 20: Continuum limit of XY model

Let us start with the XY model defined on a cubic d-dimensional lattice. The allowed configurations are parameterized by a planar unit vector $\mathbf{n}_i = (\cos\theta_i, \sin\theta_i)$ on each site i of the lattice. The energy is given by

$$E = -\sum_{\langle ij \rangle} J \cos(\theta_i - \theta_j)\,. \tag{12.123}$$

We assume that the most important configurations are smooth on a lattice scale and one can think of θ_i as of smooth function $\theta(\mathbf{x})$ defined in R^d - continuous d-dimensional space. Show that the energy is given in this continuous limit by

$$E = \frac{J}{2} \int \frac{d^d x}{a^d}\, a^2 (\partial_\mu \theta)^2\,, \tag{12.124}$$

where a is the lattice constant. The combination $\rho_s^{(0)} = Ja^{2-d}$ is referred to as *bare spin-wave stiffness* (or *bare superfluid density*).

Compute the energy of the vortex in such a model. Remember that the divergent integrals should be cut off by lattice constant a and by the size of the system L at small and large distances, respectively.

For the answer see Ref. [6].

Exercise 21: Correlation function $\langle(\theta(x) - \theta(0))^2\rangle$

Calculate the correlation function $\langle(\theta(x) - \theta(0))^2\rangle$ in the XY model in d dimensions neglecting the topology of θ, i.e., neglecting vortices, and thinking about θ as of real number without periodicity. Divergencies at small distances should be cut off by the lattice constant a.

Hint: consider $\int d^d q\, \frac{e^{i\mathbf{q}\cdot\mathbf{x}}}{q^2}$ with proper cutoffs. For the answer see Ref. [6].

Exercise 22: Correlation function $\langle \mathbf{n}(x)\mathbf{n}(0)\rangle$. Low temperatures.

Using the result of the previous exercise calculate the correlation function $\langle \mathbf{n}(x)\mathbf{n}(0)\rangle$ in the XY model in d dimensions neglecting the topology of θ. Write $\langle \mathbf{n}(x)\mathbf{n}(0)\rangle = \langle \cos(\theta(x)-\theta(0))\rangle = \mathbb{R}\langle e^{i(\theta(x)-\theta(0))}\rangle$ and use the properties of Gaussian integrals.

Make the conclusion about the existence of a true long range order in XY model in 2d and relate it to the Mermin-Wagner theorem.

For the answer see Ref. [6].

Exercise 23: Correlation function $\langle \mathbf{n}(x)\mathbf{n}(0)\rangle$. High temperatures.

Let us consider high temperatures. Assume that $J/T \ll 1$. Using high temperature expansion for XY-model Eq. (12.123) show that correlation function $\langle \mathbf{n}(x)\mathbf{n}(0)\rangle$ decays exponentially. Find the correlation length at high temperatures.

For the answer see Ref. [6].

Exercise 24: Vortex unbinding

Make an estimate of the BKT phase transition temperature in 2d XY model. Use the energy of the vortex calculated previously, the estimate of the entropy of the vortex, and the condition $F = 0$ for the free energy of the vortex.

For the answer see Ref. [6].

Appendix B: Integrating out l field

Consider the $\mathbf{l}$ dependent part of the action Eq. (12.100)

$$\int d\tau \, \frac{d^d x}{a^d} \left(2JS^2 a^2 d\,(\mathbf{l})^2 - Sa^1 \mathbf{l} \cdot \Big[\mathbf{h} - \lambda \mathbf{m} + id\,(\mathbf{m} \times \partial_\tau \mathbf{m}) \Big] \right) . \qquad (12.125)$$

Here we added the Lagrange multiplier λ. The variation with respect to λ produces the constraint $\mathbf{l}\cdot\mathbf{m} = 0$. The field $\mathbf{l}$ enters the action quadratically and can be integrated out just by substituting $\mathbf{l}$ given by the variational principle. After adding the Lagrange multiplier we can vary over $\mathbf{l}$ without any constraints and obtain

$$4JS^2 a^{2-d} d\mathbf{l} - Sa^{1-d} \Big[\mathbf{h} - \lambda \mathbf{m} + id\,(\mathbf{m} \times \partial_\tau \mathbf{m}) \Big] = 0$$

or

$$\mathbf{l} = \frac{1}{4JSad} \Big[\mathbf{h} - \lambda \mathbf{m} + id\,(\mathbf{m} \times \partial_\tau \mathbf{m}) \Big] .$$

Substituting the latter expression into Eq. (12.125) we obtain

$$\frac{1}{8Ja^d d} \int d\tau \, d^d x \, \Big[\mathbf{h} - \lambda \mathbf{m} + id\,(\mathbf{m} \times \partial_\tau \mathbf{m}) \Big]^2 . \qquad (12.126)$$

Variation over λ gives

$$0 = \mathbf{m} \cdot \Big[\mathbf{h} - \lambda \mathbf{m} + id\,(\mathbf{m} \times \partial_\tau \mathbf{m}) \Big] = \mathbf{m} \cdot \mathbf{h} - \lambda$$

which determines $\lambda = \mathbf{h}\cdot\mathbf{m}$ and gives for the Eq. (12.126)

$$\begin{aligned} - &\quad \frac{1}{8Ja^d d}\int d\tau\, d^dx \left[\mathbf{h} - \mathbf{m}(\mathbf{h}\cdot\mathbf{m}) + id\,(\mathbf{m}\times\partial_\tau\mathbf{m})\right]^2 \\ = &\quad -\frac{1}{8Ja^d d}\int d\tau\, d^dx \left[\mathbf{h}^2 - (\mathbf{h}\cdot\mathbf{m})^2 - d^2\,(\mathbf{m}\times\partial_\tau\mathbf{m})^2 - 2id\,\mathbf{h}\cdot(\mathbf{m}\times\partial_\tau\mathbf{m})\right] \\ \to &\quad \frac{1}{8Ja^d}\int d\tau\, d^dx \left[d\,(\partial_\tau\mathbf{m})^2 + 2i\,\mathbf{h}\cdot(\mathbf{m}\times\partial_\tau\mathbf{m})\right]. \end{aligned} \tag{12.127}$$

In the last step we dropped terms independent of $\mathbf{m}$ and the term quadratic in $\mathbf{h}$. Putting all results together, we obtain the effective action which is the functional of the $\mathbf{m}$ field only.

$$\begin{aligned} S[\mathbf{m}] &= i\frac{S}{2}\delta_{d,1}\int d\tau\, dx\, \mathbf{m}\cdot(\partial_\tau\mathbf{m}\times\partial_x\mathbf{m}) \\ &\quad + JS^2a^{2-d}\frac{1}{2}\int d\tau\, d^dx\,(\partial_\mu\mathbf{m})^2 + \frac{d}{8Ja^d}\int d\tau\, d^dx\,(\partial_\tau\mathbf{m})^2 \\ &\quad + i\frac{1}{4Ja^d}\int d\tau\, d^dx\,\mathbf{h}\cdot(\mathbf{m}\times\partial_\tau\mathbf{m}). \end{aligned}$$

Appendix C: Homotopy groups often used in physics

In this appendix, we collect some of the homotopy groups often used in physics. Many of these groups can be found in [42].

Generalities

If M and N are two topological spaces then for their direct product we have

$$\pi_k(M\times N) = \pi_k(M)\times\pi_k(N).$$

If M is a simply-connected topological space ($\pi_0(M) = \pi_1(M) = 0$) and group H acts on M then one can form topological space M/H identifying points of M which can be related by some element of H ($x \equiv hx$). Then we have the following relation

$$\pi_1(M/H) = \pi_0(H).$$

In particular, if H is a discrete group, $\pi_0(H) = H$, and

$$\pi_1(M/H) = H.$$

For higher homotopy groups, we have the relation

$$\pi_k(M/H) = \pi_k(M), \quad \text{if } \pi_k(H) = \pi_{k-1}(H) = 0. \tag{12.128}$$

Homotopy groups of spheres

For a circle

$$\begin{aligned} \pi_1(S^1) &= \mathbb{Z}, \\ \pi_k(S^1) &= 0, \quad \text{for } k \geq 2. \end{aligned}$$

For higher-dimensional spheres, it is true that

$$\begin{aligned} \pi_n(S^n) &= \mathbb{Z}, \\ \pi_k(S^n) &= 0, \quad \text{for } k < n. \end{aligned}$$

Homotopy groups of spheres $\pi_{n+k}(S^n)$ do not depend on n for $n > k+1$ (homotopy groups stabilize). In the table below we show in shaded cells the entries from which homotopy groups remain stable (constant along the diagonal of the table).

Homotopy groups of spheres									
	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	π_9
S^1	$\mathbb{Z}$	0	0	0	0	0	0	0	0
S^2	0	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{12}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_3$
S^3	0	0	$\mathbb{Z}$	$\mathbb{Z}_{\mathbf{2}}$	$\mathbb{Z}_2$	$\mathbb{Z}_{12}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_3$
S^4	0	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_{\mathbf{2}}$	$\mathbb{Z} \times \mathbb{Z}_{12}$	$\mathbb{Z}_2 \times \mathbb{Z}_2$	$\mathbb{Z}_2 \times \mathbb{Z}_2$
S^5	0	0	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{\mathbf{24}}$	$\mathbb{Z}_2$
S^6	0	0	0	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{24}$
S^7	0	0	0	0	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$
S^8	0	0	0	0	0	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$

Here and thereon we denote by $\mathbb{Z}$ the group isomorphic to the group of integer numbers with respect to addition. $\mathbb{Z}_n$ is a finite Abelian cyclic group. It can be thought of as a group of n-th roots of unity with respect to a multiplication. Alternatively, it is isomorphic to a group of numbers $\{0, 1, 2, \ldots, n-1\}$ with respect to addition modulo n. Or simply $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$.

Homotopy groups of Lie groups

Unitary groups

Bott periodicity theorem for unitary groups states that for $k > 1$, $n \geq \frac{k+1}{2}$

$$\pi_k(U(n)) = \pi_k(SU(n)) = \begin{cases} 0, & \text{if } k\text{-even}; \\ \mathbb{Z}, & \text{if } k\text{-odd}. \end{cases}$$

The fundamental group $\pi_1(SU(n)) = 0$ and $\pi_1(U(n)) = 1$ for all n.

In the following table, we show in shaded cells the entries from which Bott periodicity theorem "starts working" and table entries become the same further down the column.

Homotopy groups of unitary groups												
	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	π_9	π_{10}	π_{11}	π_{12}
$U(1)$	$\mathbb{Z}$	0	0	0	0	0	0	0	0	0	0	0
$SU(2)$	0	**0**	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{12}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_3$	$\mathbb{Z}_{15}$	$\mathbb{Z}_2$	$\mathbb{Z}_2 \times \mathbb{Z}_2$
$SU(3)$	0	0	$\mathbb{Z}$	**0**	$\mathbb{Z}$	$\mathbb{Z}_6$	0	$\mathbb{Z}_{12}$	$\mathbb{Z}_3$	$\mathbb{Z}_{30}$	$\mathbb{Z}_4$	$\mathbb{Z}_{60}$
$SU(4)$	0	0	$\mathbb{Z}$	0	$\mathbb{Z}$	**0**	$\mathbb{Z}$	$\mathbb{Z}_{24}$	$\mathbb{Z}_2$	$\mathbb{Z}_{120} \times \mathbb{Z}_2$	$\mathbb{Z}_4$	$\mathbb{Z}_{60}$
$SU(5)$	0	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	**0**	$\mathbb{Z}$			

Orthogonal groups

Bott periodicity theorem for orthogonal groups states that for $n \geq k + 2$

$$\pi_k(O(n)) = \pi_k(SO(n)) = \begin{cases} 0, & \text{if } k = 2, 4, 5, 6 \ \ (mod\ 8); \\ \mathbb{Z}_2, & \text{if } k = 0, 1 \ \ (mod\ 8); \\ \mathbb{Z}, & \text{if } k = 3, 7 \ \ (mod\ 8). \end{cases}$$

In the following table, we show in shaded cells the entries from which Bott periodicity theorem "starts working".

Homotopy groups of orthogonal groups								
	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8
$SO(2)$	$\mathbb{Z}$	0	0	0	0	0	0	0
$SO(3)$	$\mathbb{Z}_{\mathbf{2}}$	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{12}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$
$SO(4)$	$\mathbb{Z}_2$	**0**	$(\mathbb{Z})^{\times 2}$	$(\mathbb{Z}_2)^{\times 2}$	$(\mathbb{Z}_2)^{\times 2}$	$(\mathbb{Z}_{12})^{\times 2}$	$(\mathbb{Z}_2)^{\times 2}$	$(\mathbb{Z}_2)^{\times 2}$
$SO(5)$	$\mathbb{Z}_2$	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0	$\mathbb{Z}$	0
$SO(6)$	$\mathbb{Z}_2$	0	$\mathbb{Z}$	**0**	$\mathbb{Z}$	0	$\mathbb{Z}$	$\mathbb{Z}_{24}$
$SO(n > 6)$	$\mathbb{Z}_2$	0	$\mathbb{Z}$	0	**0**	0		

Symplectic groups

Bott periodicity theorem for symplectic groups states that for $n \geq \frac{k-1}{4}$

$$\pi_k(Sp(n)) = \begin{cases} 0, & \text{if } k = 0, 1, 2, 6 \ \ (mod\ 8); \\ \mathbb{Z}_2, & \text{if } k = 4, 5 \ \ (mod\ 8); \\ \mathbb{Z}, & \text{if } k = 3, 7 \ \ (mod\ 8). \end{cases}$$

In the following table, we show in shaded cells the entries from which Bott periodicity theorem "starts working".

Homotopy groups of symplectic groups												
	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	π_9	π_{10}	π_{11}	π_{12}
$Sp(1)$	0	**0**	$\mathbb{Z}$	$\mathbb{Z}_{\mathbf{2}}$	$\mathbb{Z}_{\mathbf{2}}$	$\mathbb{Z}_{12}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_3$	$\mathbb{Z}_{15}$	$\mathbb{Z}_2$	$\mathbb{Z}_2 \times \mathbb{Z}_2$
$Sp(2)$	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	**0**	$\mathbb{Z}$	**0**	**0**	$\mathbb{Z}_{120}$	$\mathbb{Z}_2$	$\mathbb{Z}_2 \times \mathbb{Z}_2$
$Sp(n \geq 3)$	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0	$\mathbb{Z}$	0	0	**0**	$\mathbb{Z}$	$\mathbb{Z}_{\mathbf{2}}$

Exceptional groups

Homotopy groups of exceptional groups												
	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	π_9	π_{10}	π_{11}	π_{12}
G_2	0	0	$\mathbb{Z}$	0	0	$\mathbb{Z}_3$	0	$\mathbb{Z}_2$	$\mathbb{Z}_6$	0	$\mathbb{Z} \times \mathbb{Z}_2$	0
F_4	0	0	$\mathbb{Z}$	0	0	0	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0	$\mathbb{Z} \times \mathbb{Z}_2$	0
E_6	0	0	$\mathbb{Z}$	0	0	0	0	0	$\mathbb{Z}$	0	$\mathbb{Z}$	$\mathbb{Z}_{12}$
E_7	0	0	$\mathbb{Z}$	0	0	0	0	0	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$
E_8	0	0	$\mathbb{Z}$	0	0	0	0	0	0	0	0	0

Homotopy groups of some other spaces

Tori

An n-dimensional torus can be defined as a direct product of n circles $T^n = (S^1)^{\times n}$. One can immediately derive that

$$\begin{aligned} \pi_1(T^n) &= (\mathbb{Z})^{\times n}, \\ \pi_k(T^n) &= 0, \quad \text{for } k \geq 2\,. \end{aligned}$$

Projective spaces

The real projective space $\mathbb{R}P^n$ can be represented as $\mathbb{R}P^n = S^n/\mathbb{Z}_2$. For $n = 1$ $\mathbb{R}P^1 = S^1$. We have from Eq. (12.128)

$$\begin{aligned} \pi_1(\mathbb{R}P^1) &= \mathbb{Z}\,, \\ \pi_1(\mathbb{R}P^n) &= \mathbb{Z}_2, \quad \text{for } n \geq 2\,, \\ \pi_k(\mathbb{R}P^n) &= \pi_k(S^n), \quad \text{for } k \geq 2\,. \end{aligned}$$

Homotopy groups of real projective spaces									
	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	π_9
$\mathbb{R}P^1$	$\mathbb{Z}$	0	0	0	0	0	0	0	0
$\mathbb{R}P^2$	$\mathbb{Z}_2$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{12}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_3$
$\mathbb{R}P^3$	$\mathbb{Z}_2$	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{12}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_3$
$\mathbb{R}P^4$	$\mathbb{Z}_2$	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z} \times \mathbb{Z}_{12}$	$\mathbb{Z}_2 \times \mathbb{Z}_2$	$\mathbb{Z}_2 \times \mathbb{Z}_2$

Similarly, for complex projective spaces $\mathbb{C}P^n$ we have $\mathbb{C}P^1 = S^2$ and, generally, $CP^n = S^{2n+1}/S^1$. We have for homotopy groups

$$\begin{aligned} \pi_1(\mathbb{C}P^n) &= 0, \\ \pi_2(\mathbb{C}P^n) &= \mathbb{Z}, \\ \pi_k(\mathbb{C}P^n) &= \pi_k(S^{2n+1}), \quad \text{for } k \geq 3. \end{aligned}$$

Homotopy groups of complex projective spaces												
	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	π_9	π_{10}	π_{11}	π_{12}
$\mathbb{C}P^1$	0	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{12}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_3$	$\mathbb{Z}_{15}$	$\mathbb{Z}_2$	$\mathbb{Z}_2 \times \mathbb{Z}_2$
$\mathbb{C}P^2$	0	$\mathbb{Z}$	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{24}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{30}$
$\mathbb{C}P^3$	0	$\mathbb{Z}$	0	0	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{24}$	0	0
$\mathbb{C}P^4$	0	$\mathbb{Z}$	0	0	0	0	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{24}$

References

[1] R. B. Laughlin and D. Pines, Proceedings of the National Academy of Sciences of the United States of America **97**, 28 (2000).

[2] P. W. Anderson et al, Science, **177**(4047),393 (1972).

[3] M. E. Peskin and D. V. Schroeder, *Quantum field theory* (The Advanced Book Program, Perseus Books Reading, Massachusetts, 1995).

[4] M. Gell-Mann and M. Lévy, Il Nuovo Cimento, **16**, 705 (1960).

[5] S. L. Adler and R. F. Dashen, *Current algebras and applications to particle physics*, volume 30. (Benjamin, 1968).

[6] A. M. Polyakov, *Gauge Fields and Strings* (CRC, September 1987).

[7] N. D. Mermin, Rev. Mod. Phys., **51**, 591 (1979).

[8] A. Altland and B. D. Simons, *Condensed matter field theory* (Cambridge University Press, 2010).

[9] E. Fradkin, *Field theories of condensed matter physics* (Cambridge University Press, 2013).

[10] A. G. Abanov and P. B. Wiegmann, Nucl. Phys. B, **570**, 685 (2000).

[11] A. G. Abanov, Phys. Lett. B, **492**, 321 (2000).

[12] F. Wilczek and A. Shapere, *Geometric phases in physics*, volume 5. (World Scientific, 1989).

[13] S. Treiman and R. Jackiw, *Current algebra and anomalies* (Princeton University Press, 2014).

[14] B. A. Dubrovin, A. T. Fomenko, and S. P. Novikov, *Modern geometry - methods and applications: Part II: The geometry and topology of manifolds*, volume 104 (Springer Science & Business Media, 2012).

[15] M. Nakahara, *Geometry, topology and physics* (CRC Press, 2003).

[16] M. Monastyrsky, *Topology of gauge fields and condensed matter* (Springer Science & Business Media, 2013).

[17] V. I. Arnold, *Mathematical methods of classical mechanics* (Graduate Texts in Mathematics, 60:229–234, 1991).

[18] M. Stone and P. Goldbart, *Mathematics for physics: a guided tour for graduate students* (Cambridge University Press, 2009).

[19] L. D. Landau and E. M. Lifshitz, *Mechanics* (Course of theoretical physics, Vol 1), pages 84–93, 1976.

[20] J. Wess and B. Zumino, Phys. Lett. B, **37**, 95 (1971).

[21] E. Witten, Nucl. Phys. B, **223**, 422 (1983).

[22] S. P. Novikov, Russian mathematical surveys, **37**, 1 (1982).

[23] E. Witten, Nucl. Phys. B, **223**, 433 (1983).

[24] M. Stone, Phys. Rev. D, **33**, 1191 (1986).

[25] E. Witten, Phys. Lett. B, **117**, 324 (1982).

[26] K. Fujikawa and H. Suzuki, *Path integrals and quantum anomalies.* (Number 122. Oxford University Press on Demand, 2004).

[27] I. K. Affleck, *Field theory methods and quantum critical phenomena*, Technical report, PRE-31353, 1988.

[28] A. M. Polyakov, Phys. Lett. B, **59**, 79 (1975).

[29] P. B. Wiegmann, Phys. Lett. B, **152**, 209 (1985).

[30] F. D. M. Haldane, Phys. Rev. Lett., **50**, 1153 (1983).

[31] F. D. M. Haldane, Phys. Lett. A, **93**, 464 (1983).

[32] H. Bethe, Zeitschrift für Physik,**71**, 205 (1931).

[33] H. Levine, S. B. Libby, and A. M. M. Pruisken, Phys. Rev. Lett., **51**, 1915 (1983).

[34] D. E. Khmel'nitskii, JETP lett, **38**, 552 (1983).

[35] I. Affleck and F. D. M. Haldane, Phys. Rev. B, **36**, 5291 (1987).

[36] M. Hagiwara, K. Katsumata, I. Affleck, B. I. Halperin, and J. P. Renard, Phys. Rev. Lett., **65**, 3181 (1990).

[37] M. Kenzelmann, G. Xu, I. A. Zaliznyak, C. Broholm, J. F. DiTusa, G. Aeppli, T. Ito, K. Oka, and H. Takagi, Phys. Rev. Lett. **90**, 087202 (2003); Erratum Phys. Rev. Lett. **90**, 109902 (2003).

[38] I. Affleck, T. Kennedy, E. H. Lieb, and H. Tasaki, Phys. Rev. Lett., **59**, 799 (1987).

[39] I. Affleck, T. Kennedy, E. H. Lieb, and H. Tasaki, In *Condensed Matter Physics and Exactly Soluble Models*, pages 253–304. Springer, 1988.

[40] A. Gromov, G. Y. Cho, Y. You, A. G. Abanov, and E. Fradkin, Phys. Rev. Lett., **114**, 016805 (2015).

[41] S. Ryu, J. E. Moore, and A. W. W. Ludwig, Phys. Rev. B, **85**, 045104 (2012).

[42] K. Itô, *Encyclopedic dictionary of mathematics*, volume 1. (MIT press, 1993).

13

Dirac quasiparticles and Majorana modes in condensed matter systems

K. Sengupta

We present a pedagogical introduction to Dirac fermions and Majorana bound states which appear as emergent quasiparticles in several condensed matter systems.

13.1 Introduction

The concept of quasiparticles has been central to understanding correlated condensed matter systems for many decades. The quantum many-body ground states of such systems result from competition between the kinetic energy and the (screened) Coulomb interactions of their constituent 10^{23} electrons. Such a competition is known to lead to extremely complicated ground states specially when the interaction dominates over the kinetic energy of the electrons. A specific example of such systems is a Quantum Hall system in the ultra strong magnetic field regime. The presence of such a strong magnetic field leads to a complete quenching of the kinetic energy of the electrons and leads to a huge ground state degeneracy provided that the magnetic field is strong enough to create more states in the lowest Landau level than the number of electrons $N_{\rm el}$ in the system (fractional quantum Hall effect regime). This degeneracy is lifted by the Coulomb interaction leading to exotic many-body ground states; indeed, when the number of flux quanta N_ϕ, passing through the two dimensional (2D) plane where the electrons are confined, satisfies $N_{\rm el} = \nu N_\phi$, one observes Hall conductance plateaus for specific values of ν: $\sigma_{\rm xy} = \nu e^2/h$, where e is the electron charge and $h = 2\pi\hbar$ is the Planck's constant and we shall restrict our discussion to fractions where $\nu = 1/(2n+1)$ with integer n. The quasiparticles

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_13

of such correlated ground state turns out to be significantly different from electrons; they happen to have fractional charge νe. Moreover, these quasiparticles turn out to be anyons; exchanging two such quasiparticles lead to an exchange phase of $\exp(i\theta)$ with $\theta = \nu\pi$. These quasiparticles, which are detectable in realistic experiments, therefore serve as a tool for unraveling the complex nature of the many-body correlated ground states.

The physics of the fractional quantum Hall effect and several other related spin and fermionic models made it clear a few decades back that strong interactions may lead to exotic quasiparticles whose properties differ significantly from those of dressed electrons. However, in the last decade or so, it has been realized that novel quasiparticles whose properties differ significantly from that of electrons may also appear in weakly interacting condensed matter systems either due to specific nature of their band structure or as a consequence of novel topological feature of the system ground state. Examples of such quasiparticles involve Dirac fermions and Majorana modes. The former mimics several properties of relativistic Dirac electrons which are qualitatively different from those of their non-relativistic counterparts while the latter are non-Abelian anyonic quasiparticles localized, for example, at the end of a 1D superconducting wire, with zero energy. In the rest of the article, we shall focus on the properties of these quasiparticles.

The plan of the rest of this article is as follows. In Sec. 13.2, we detail out the origin and some of the properties of Dirac quasiparticles. This is followed by Sec. 13.3, where we outline the origin and properties of localized Majorana modes in unconventional superconductors.

13.2 Dirac Fermions in Graphene and Topological Insulators

13.2.1 Graphene band structure

The emergence of Dirac-like quasiparticles in condensed matter may most simply occur due to specific properties of the band structure of the system. The simplest example of this is graphene, a well-known 2D sheet of carbon atoms which is essentially a single layer of the well-known carbon allotrope graphite. In graphene, the carbon atoms form a honeycomb lattice. The electrons of these atoms undergo sp^2 hybridization leaving behind a single p_z orbital. Ab-initito calculations confirm that the all the transport properties of graphene comes from electrons from these p_z orbitals.Also such calculations lead to a tight-binding model for the kinetic energy of graphene electrons which we are going to use in what follows.

Such a tight-binding model mimics the kinetic energy electrons to comprise of hopping between near-neighbor sites; in what follows, for simplicity, we shall assume that only nearest-neighbor hoppings are non-zero. This is certainly not the case for realistic graphene systems; however, the basic qualitative features of the graphene quasiparticles that we would like to point out do not change due to this assumption. Within this approximation the tight-binding

Hamiltonian for the graphene electrons can be written as

$$H = -t \sum_{\langle \mathbf{r},\mathbf{r}' \rangle} \psi^\dagger_{\mathbf{r}} \psi_{\mathbf{r}'} + \text{h.c.}, \tag{13.1}$$

where t denotes the hopping amplitude and $\psi_{\mathbf{r}}$ denotes the electron annihilation operator at site $\mathbf{r}$. At this point, we make an important observation. The honeycomb lattice of graphene posses a property that any two nearest neighbor sites of the lattice are inequivalent from a crystallographic point of view. This is most easily seen by noticing that if the surrounding sites of any given site form a Y, then the surrounding sites of all its neighbor form an inverted Y. These two group of sites form two different sublattice which is commonly known as A and B sublattice of graphene. In the presence of such a sublattice structure, one can meaningly talk about probability amplitude of electrons being on an A or a B sublattice. Thus one can write the electron wavefunction as $\Psi(\mathbf{r}) \equiv (\Psi_A(\mathbf{r}), \Psi_B(\mathbf{r}))$, where Ψ_A and Ψ_B denotes probability amplitudes of the electron being on the A or B sublattice. This situation is analogous to the having a (fictitious) spin index whose up and down states correspond to the electrons being on the A and the B sublattice. This analogy has led to the terminology "pseudospin" for describing the sublattice degree of greedom for electrons in graphene.

Since the nearest-neighbor hopping of electrons always happens between different sublattices, the tight-binding Hamiltonian of the electrons can be written as

$$H = -t \sum_{\langle \mathbf{r}\mathbf{r}' \rangle} \psi^\dagger_A(\mathbf{r}) \psi_B(\mathbf{r}') + \text{h.c.}, \tag{13.2}$$

where the sum over $\mathbf{r}$ is now over all A sublattice sites. In momentum space, H can be written as

$$\begin{aligned} H &= -\sum_{\mathbf{k}} \psi^\dagger_A(\mathbf{k}) \phi_{\mathbf{k}} \psi_B(\mathbf{k}) + \text{h.c.}, \qquad (13.3) \\ \phi_{\mathbf{k}} &= t \left(2 \exp(-i\sqrt{3} a k_y/2) \cos(3k_x a/2) + \exp(i\sqrt{3} k_y a) \right), \end{aligned}$$

where a is the lattice spacing. The energy spectrum can be obtained by diagonalizing H and is given by

$$\begin{aligned} E_{\pm\mathbf{k}} &= \pm|\phi_{\mathbf{k}}|, \quad |\phi_{\mathbf{k}}| = t\sqrt{3 + g_{\mathbf{k}}} \\ g_{\mathbf{k}} &= 2\cos(\sqrt{3} k_y a) + 4\cos(\sqrt{3} k_y a/2)\cos(3k_x a/2). \end{aligned} \tag{13.4}$$

The two energy bands $E_{+\mathbf{k}}$ and $E_{-\mathbf{k}}$ are called conduction and valence band of graphene. Note that they can intersect only at zero energy. Such an intersection takes places at six corners of the hexagonal Brillouin zone of graphene. Out of these six points, only two are inequivalent; the rest are connected by reciprocal lattice vectors. Furthermore, around these two points, commonly called as the

$$K = (0, 4\pi/(3\sqrt{3}a)), \quad \text{and} \quad K' = (0, 4\pi/(3\sqrt{3}a))$$

points or valleys, the Hamiltonian has a 2×2 matrix structure corresponding to the pseudospin degree of freedom and a linear dispersion given by $E_{\mathbf{k}} = \pm v_F|k|$ where $v_F = \sqrt{3}ta/2$ and $\mathbf{k}$ is now measured from the K or K' valleys. Around these points, the Hamiltonian can be written (by expanding $\phi_{\mathbf{k}}$ around K or K' points) in terms of a two component pseudospinor $\psi = (\psi_A(\mathbf{k}), \psi_B(\mathbf{k}))$ as

$$H_{\text{Dirac}} = \hbar v_F \int \frac{d^2k}{(2\pi)^2} \psi^\dagger(\mathbf{k})(\pm\sigma_x k_x + \sigma_y k_y)\psi(\mathbf{k}). \tag{13.5}$$

Thus the low-energy quasiparticles around K and K' obey Dirac-like equation with v_F replacing the speed of light c.

The above discussion brings out how the details of band structure and the underlying lattice structure of a non-interacting system of electrons may lead to effective Dirac-like nature of its low-energy quasiparticles. These quasiparticles have qualitatively different properties compared to their Schrodinger counterpart. We shall list some of them in Sec. 13.2.3; but before that let us look into one more system where the topology of the bulk band structure of a 3D solid may lead to Dirac-like quasiparticles at the edge.

13.2.2 Topological Insulators

The topological properties of the bulk electronic spectrum often lead to interesting properties at the edge of a system; this perispomenon goes by the name of the bulk-boundary correspondence. There are several important examples of this phenomenon in condensed matter system.

The first and probably the simplest example of this phenomenon is the well-known quantum Hall effect. In these systems, the presence of an external magnetic field breaks time-reversal symmetry in the bulk and lead to a finite Hall conductance $\sigma_0 = ne^2/h$, where n is an integer for integer Hall effect and could be fractional for fractional Hall effect. The action describing the system must therefore, when probed by external electromagnetic field, contain a term

$$S \sim \sigma_0 \int d^2r\ dt\ \epsilon_{\mu\nu\lambda}\ A_\mu \partial_\nu A_\lambda, \tag{13.6}$$

where $\epsilon_{\mu\nu\lambda}$ is the anti-symmetric tensor. Now let us consider that the quantum Hall sample is finite and let us demand that S must be gauge-invariant. Making the transformation $A_\mu \to A_\mu + \partial_\mu \Lambda$, we find that we are left with an additional surface term

$$\delta S \sim \sigma_0 \epsilon_{\mu\nu\lambda} \int dx_\mu\ dt\ \Lambda\ \partial_\nu A_\lambda. \tag{13.7}$$

This term is usually non-zero since there is no reason (unlike the case in infinite systems) for the applied electromagnetic fields to vanish at the sample edge. However, this seems to lead us to a gauge-dependent action which is clearly nonsense. The only resolution out of this conundrum is to allow for additional

modes localized at the edge such that its action, under the same gauge transformation, cancels δS. Thus we are forced by the gauge invariance principle to the conclusion that time reversal symmetry breaking in the bulk of a system must lead to additional modes at the sample edge.

Once this connection between symmetry breaking in the bulk and the presence of chiral edge modes is understood, it is natural to look for instances where this can take place. In particular, the question asked by the community at this point was whether such a bulk-edge correspondence be achieved without breaking time-reversal symmetry. The answer to this query is affirmative and can be understood by considering systems with strong spin-orbit coupling. To see the connection, let us consider a 2D system of electrons in the x-y plane with a spin-orbit coupling term with strength α

$$\begin{aligned} H_{\rm SO} &= \alpha \int d^2r\ \psi^\dagger L_z S_z \psi(r), \\ L_z &= (xp_y - yp_x), \quad \text{and} \quad \psi^\dagger(r) = (\psi^\dagger_\uparrow(r), \psi^\dagger_\downarrow(r)). \end{aligned} \tag{13.8}$$

The total Hamiltonian of the system can be written as

$$\begin{aligned} H &= \int d^2r\ \psi^\dagger(r) \left(\frac{p^2}{2m} I + \alpha L_z S_z \right) \psi(r) \\ &= \int d^2r\ \psi^\dagger(r) \frac{\left[(p_x - m\alpha y s_z)^2 + (p_y + m\alpha x s_z)^2 \right]}{2m} \psi(r), \end{aligned} \tag{13.9}$$

where we have neglected terms $O(\alpha^2)$ which are usually small. Thus we find that H takes the form of a Hamiltonian of electrons in the presence of an effective magnetic field along z but with opposite direction for each spin species. Thus, just as electrons in the presence of a magnetic field exhibits chiral edge states, the present systems will have two chiral edge states carrying opposite spins and with opposite velocities. Note that this system does not break time-reversal symmetry; the edge states do not carry any net charge. However, they do carry a net spin-current. Such systems are called spin-Hall systems and they exhibit a quantized spin-Hall conductance.

Topological insulators turn out to be 3D counterparts of spin-Hall insulators. In these systems the presence of strong spin-orbit coupling and time reversal symmetry results in energy bands to occur in pair (known as Kramer's pairs) which satisfy $E_\sigma(\mathbf{k}) = E_{\bar\sigma}(-\mathbf{k})$. Such bands can only cross at momenta $\mathbf{k}^* = 0, \mathbf{G}/2$, where $\mathbf{G}$ is the reciprocal lattice vector of the system. These crossings turn out to play an important part in band topology and can be shown to lead to gapless states at the edge whose quasiparticles obey an effective Dirac Hamiltonian. The details of how this happens can be found in Ref. [1]. Here, we merely note that the bulk-band correspondence in these systems do lead to presence of Dirac fermions at the surface. The Hamiltonian of the Dirac fermions atop a topological insulator surface can be written as

$$H_{\rm Dirac} = \hbar v_F \int \frac{d^2k}{(2\pi)^2} \psi^\dagger(\mathbf{k}) [\hat z \cdot (\boldsymbol{\sigma} \times \mathbf{k})] \psi(\mathbf{k}) \tag{13.10}$$

where $\psi^\dagger(\mathbf{k}) = (\psi^\dagger_\uparrow(\mathbf{k}), \psi^\dagger_\downarrow(\mathbf{k}))$ is a two-component electron creation operator, v_F is the Fermi velocity of the surface electrons, $\boldsymbol{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$ is the Pauli matrix, and we have denoted the direction orthogonal to the surface as $\hat{z}$.

13.2.3 Properties of Dirac quasiparticles

Several properties of the Dirac quasiparticles discussed in the last section is significantly different from their Schrodinger counterparts. We shall discuss three such properties in this section.

The first of these concerns helicity of the Dirac electrons which is well-known for massless relativistic fermions in particle physics. To see how this appears in the present context, let us consider Dirac electrons on the surface of graphene whose Hamiltonian is given by Eq. 13.5. The eigenfunctions of the quasiparticles obeying this Hamiltonian can be obtained by solving the associated Dirac equation $H_{\text{Dirac}}\psi = E\psi$ and is given by

$$\psi = (u, v) = (1, \pm \exp(\pm i\theta)), \tag{13.11}$$

where $+(-)$ signs correspond to electrons in the $K(K')$ valleys and $\theta = \arctan(k_y/k_x)$. Note that if the electrons moves along x, $k_y = 0$ and in this case its eigenfunctions is $(1, \pm 1)$. This implies that the pseudospin of a quasiparticle in $K(K')$ valley points along (opposite to) x. One can repeat this calculation for any direction of electrons opposite with the same result, namely, the pseudospin of these quasiparticles in $K(K')$ valley point along (opposite to) their direction of motion. This property is qualitatively different from that of a Schrodinger quasiparticle for which there is no connection between spin and direction of motion. An analogous calculation with the Dirac Hamiltonian for quasiparticles atop a topological insulator surface (Eq. 13.10) yields

$$\psi = (u, v) = (1, i \exp(i\theta)). \tag{13.12}$$

In this case, the spin of the electron points orthogonal to its direction of motion.

The second property concerns Dirac electrons atop a topological insulator surface in the presence of a magnetic field parallel to the surface. To see how the behavior of Dirac electrons differ from their Schrodinger counterparts in such a situation, we consider the Dirac Hamiltonian given in Eq. 13.10 in the presence of a magnetic field along x: $B = B_0\hat{x}$. Since the field is applied parallel to the plane, it is not going to affect the orbital motion of the electrons irrespective of whether they obey Schrodinger or Dirac equations. This can be best understood by noting that the vector potential of the applied field may be chosen to just have $A_z \neq 0$ in which case the field does not couple to any in plane momenta operators. However, the spin of the electrons still sees the applied field through a Zeeman term; for Schrodinger electrons, one may simply choose the spin quantization axis of the electrons along the field direction leading to

$$H_Z^{\text{Sch}} = \int d^2k \; \psi^\dagger(\mathbf{k}) g\mu_B B_0 \sigma_z \psi(\mathbf{k}),$$

where μ_B is the Bohr magneton and g is the gyromagnetic ratio. This term is responsible for Zeeman splitting of electron energy levels. However, for a quasiparticle obeying Dirac equations, one can not choose the spin quantization axis along the magnetic field since in writing Eq. 13.10, one has already chosen that axis to be along z. Thus the form of the Zeeman term becomes

$$H_Z^{\rm Dir} = \int d^2k\ \psi^\dagger(\mathbf{k}) g\mu_B B\sigma_x \psi(\mathbf{k}).$$

Thus the total Hamiltonian in the presence of the applied field becomes

$$H_{\rm Dirac} = \hbar v_F \int \frac{d^2k}{(2\pi)^2} \psi^\dagger(\mathbf{k})[\sigma_x(k_y - \beta B_0) - \sigma_y k_y]\psi(\mathbf{k}),$$

where $\beta = g\mu_B/(\hbar v_F)$. This demonstrates that a constant applied field along the plane simply leads to a constant shift in k_x (or k_y or both depending on the field direction); thus such a field appears as a constant gauge potential to the Dirac quasiparticles and can not change its energy spectrum. We note here, however, that if the applied field has appropriate spatial dependence (for example if $B_0 \equiv B_0(y)$), the Dirac quasiparticle will perceive its effect as the gauge potential of a (fictitious) magnetic field along z. Here the word "fictitious" is to be understood in the sense that the field is not prepared in the laboratory; however, we note that its effect can lead to measurable effect in motion of the Dirac quasiparticles. We also note that for this effect to occur, the Dirac electrons must have a matrix structure in spin space; thus an applied magnetic field do not lead to similar behavior for graphene electrons which are Dirac pseudo-spinors. In their case, an applied strain field may have similar effect.

Finally, we discuss the transmission properties of the Dirac quasiparticles in the presence of a potential barrier. The simplest barrier that can be used to demonstrate the difference between Schrodinger and Dirac quasiparticles is the square potential barrier: $V = V_0$ for $0 \le x \le d$ and $V = 0$ otherwise. In what follows, we wish to compute the transmission probability of an electron approaching the barrier with a fixed energy $E < V_0$.

For a classical particle which undergoes deterministic Newtonian dynamics the transmission probability turns out to be zero since the particle will always bounce back after scattering from the potential barrier. This is exactly analogous to throwing a ball at a wall; if the height of the ball is less that of the wall, the ball will bounce back with unit probability. For Schrodinger electrons, which obeys non-relativistic Schrodinger equation, the probability turns out to be finite. A straightforward textbook calculation shows that the transmission probability is given by

$$\begin{aligned} T_{\rm Sch} &= \left[1 + \left(\frac{k^2+\chi^2}{2\chi k}\right)^2 \sinh^2(\chi)\right]^{-1}, \qquad (13.13)\\ \chi &= \frac{2d(V_0 - E)}{\hbar v_d}, \quad v_d = \frac{\hbar}{md}, \quad k^2 = \frac{2mEd^2}{\hbar^2}. \end{aligned}$$

We note that for $V_0 \gg E$, $T_{\rm Sch}$ is a monotonically decreasing function of V_0 which seems to give us the intuitively expected answer that the probability of tunneling through a barrier decreases with its height.

This intuition essentially fails for a quasiparticle which obeys Dirac equation. Let us consider that a Dirac quasiparticle travels form region I to the barrier and it gets transmitted with a probability $T_{\rm Dir}$. A straightforward calculation shows that in the thin barrier limit ($V_0 \to \infty$ and $d \to 0$ with $\chi_0 = V_0 d/(\hbar v_F)$ held finite) yields

$$T_{\rm Dir} = \frac{\cos^2\gamma}{1-\cos^2\chi\sin^2\gamma}, \tag{13.14}$$

where $\gamma = \arctan(k_y/k_x)$ is the angle at which the incident particle hits the barrier. We note that $T_{\rm Dir}$ is an oscillatory function of the barrier strength and it reaches unity for $\chi = n\pi$; this phenomenon is known as transmission resonance of Dirac electrons which has no analog for Schrodinger electrons in the presence of a single barrier.

Finally, we note from the expression of $T_{\rm Dir}$ that the transmission becomes unity for *any barrier strength* provided the electrons approaches the barrier with normal incidence ($\gamma = 0$). This paradoxical result goes by the name of Klein paradox and was discovered by Klein right after the discovery of Dirac equation. One way to understand this paradoxical result is the following. Consider an electron approaching the barrier at normal incidence. This implies, that its spin or pseudospin points at a fixed angle with its direction of motion. If the barrier has to reflect the electron which has a fixed helicity, it will also have to flip its spin or pseudospin upon scattering. However, the barrier potential that we have chosen does not depend on the electron's spin or pseudospin; in other words, its a scalar in spin or pseudospin space. Its an elementary result in quantum scattering theory that scattering from such barriers can never flip spin or pseudospin. Thus the barrier has no choice but to let the electron with $\gamma = 0$ pass which leads to unit transmission. This phenomenon depends the presence of a fixed helicity of an electron and thus does not occur in case of Schrodinger electrons.

13.3 Majorana modes in unconventional superconductors

Majorana fermions were first proposed in the late thirties by Ettore Majorana who found real valued wavefunctions as solutions of Dirac equations which represented fermionic particles which are their own antiparticles. Majorana fermions in the context of high energy physics have not been detected yet; they, however, are candidate particles for several cases including neutrinos and dark matter. In all of these instances, they are neutral fermionic particles which can be represented by real fields.

In contrast Majorana modes in condensed matter are emergent fermionic states at zero energy in a many-body system in 2D systems which have non-Abelian statistics and can be represented by real-valued wavefunctions. To see

how such modes may appear, we shall look at the edge states of unconventional superconductors with p-wave pairing symmetry. We shall also assume that the pairing occurs in the triplet channel with $m_z = 1$ (equal spin pairing); in this case the Bogoliubov- de Gennes (BdG) equations describing the quasiparticles of the superconductor can be written as

$$E_{\mathbf{k}}\psi_{\mathbf{k}} = [(\epsilon_{\mathbf{k}} - \mu)\tau_3 + \Delta_{\mathbf{k}}\tau_1]\psi_{\mathbf{k}}, \tag{13.15}$$

where $\epsilon_{\mathbf{k}} = \hbar^2 k^2/2m$ is the energy dispersion of the electrons in the superconductor, μ is the chemical potential, $\Delta_{\mathbf{k}} = \Delta_0 k_x/k_F$ is the superconducting pair-potential, k_x is the x component of k_F, τ_3 and τ_1 are Pauli matrices in the particle-hole space, and $\psi_{\mathbf{k}} = (u_{\mathbf{k}}, v_{\mathbf{k}})$ is the two component wavefunction describing the BdG quasiparticles. These quasiparticles, in the bulk of the superconductor has an energy $E_{\mathbf{k}} = \pm\sqrt{(\epsilon_{\mathbf{k}} - \mu)^2 + \Delta_{\mathbf{k}}^2}$. Note the spectrum of these quasipartciles is gapped at all points on the Fermi surface except at $(k_x, k_y) = (\pm k_F, 0)$ where the gap closes. These states are extended and can not occur for $E_{\mathbf{k}} < |\Delta_{\mathbf{k}}|$. In what follows, we are going to take an example.

This situation changes when we try to solve the BdG equation with an edge at $x = 0$. Here we need to use the boundary condition that $\psi(x = 0) = 0$. The most natural way to incorporate such a boundary condition is to write the wavefunction as a superposition of left and right moving waves generated due to reflection from the edge at $x = 0$. It turns out that such wavefunctions admit a zero energy solution of the Bogoliubov equations with $E = 0$ and $\psi \sim \exp[-\kappa x]$, where $\kappa = \Delta_0/\hbar v_F$ is the inverse of the localization length. The presence of such states crucially depend on the fact that a quasiparticle reflected from the barrier leading to a reversal of the sign of its momenta ($k_x \to -k_x$) sees a different sign of the pair-potential. These states do not mix with the bulk state due to the presence of the superconducting energy gap, and are hence stable.

The creation of operator for a BdG quasipartciles are given by $\gamma_{\mathbf{k}}^\dagger = u_{\mathbf{k}}^* c_{\mathbf{k}}^\dagger + v_{\mathbf{k}} c_{-\mathbf{k}}$. For the edge modes, it turns out that $u = v = 1/\sqrt{2}$, so that $\gamma^\dagger = \gamma$. This indicates that these quasiparticles are their own anti-particles. Thus these localized states represents Majorana modes. Also these modes are charge neutral; this can be understood by noting the fact that u and v represent the electron and the hole amplitudes of a BdG quasiparticle.

The detection of such Majorana modes are done through several ways. We shall first discuss the possibility of the detection of such modes using tunneling conductance (G) measurement in this article. In a typical experiment such measurement is carried across a N-B-S junction where a normal metallic region (N) is fused with the superconductor (S) with a barrier (B) region separating the two. The tunneling conductance of such a junction for a given energy eV of the incident electrons, in the limit of large barrier strength, reflects the density of states $\rho(eV)$. Clearly, in the absence of any zero energy states, the tunneling conductance will then be zero for subgap voltages ($eV < \Delta_0$)., In contrast, if these states are present $G(eV)$ will show a peak at zero energy. Theoretical calculations, following standard Landauer-Butticker approach, predicts that $G(eV = 0) = 2e^2/h$ if a Majorana mode is present and $G(eV) \simeq 0$

if it is absent. This prediction has recently been tested in experiments where the presence of a midgap state is verified in $1D$ wires with strong spin-orbit coupling and proximity-induced superconductivity. However, it turns out that it is not entirely clear if the peak occurs due to presence of Majorana modes; for example, the amplitudes of these peaks turn out, for yet unexplained reason, to be $0.2e^2/h$.

Another phase sensitive detection of Majorana modes occur through Josephson current measurements. It turns out that for Josephson junctions made out of superconductors which hosts zero energy states, the Josephson current becomes a 4π periodic function of the phase difference between the superconductors: $I_J^{dc} = \Delta_0 \sin(\phi/2)$. Since in any standard junction, the presence of a DC voltage leads to $\phi = 2eVt/\hbar$, these junctions exhibit AC Josephson effect at a fractional frequency $I_J^{AC} = \Delta_0 \sin(\omega_J t)$ with $\omega_J = eV/\hbar$. This leads to what is dubbed as fractional Josephson effect. Moreover, when such junctions are irradiated with microwave radiation of frequency ω, only even Shapiro steps occur at $2eV = 2n\hbar\omega$ (as opposed to standard Shapiro steps for conventional superconductors at $2eV = n\hbar\omega$). The absence of odd Shapiro steps has been seen in recent experiments with 1D superconducting wires and this constitutes the most robust experimental signature of Majorana modes.

In conclusion, we have provided a brief overview of the physics of Dirac quasiparticles and Majorana modes in condensed matter systems. These quasiparticles and modes appear as emergent modes; the reason for their appearance may be traced to either properties of band-structure of the system or topological properties of the bulk which, via bulk-boundary correspondence, lead to gapless states at the edge of the system. These quasiparticle and modes have strikingly different properties from those of their Schrodinger counterparts. Such properties can be tested in table-top experiments thus leading us to first experimental observation of Dirac and Majorana physics. Moreover, the presence of interaction between these quasiparticles allows us to study the result of combination of many-body interaction and Dirac/Majorana nature of quasiparticles. For more details on these subject we refer the readers to several review articles on the subject [2–5].

References

[1] R. Roy, arXiv:1004.3507.

[2] A. H. Castro Neto, F. Guinea, N. M. R. Peres, K. S. Novoselov, and A. K. Geim, Rev. Mod. Phys. **81**, 109 (2009).

[3] S. Das Sarma, Shaffique Adam, E. H. Hwang, and Enrico Rossi, Rev. Mod. Phys. **83**, 407 (2011).

[4] M. Z. Hasan and C. L. Kane, Rev. Mod. Phys. **82**, 3045 (2010).

[5] S. R. Elliott and M. Franz, Rev. Mod. Phys. **87**, 137 (2015).

14

Vertex Models and Knot invariants

P. Ramadevi

Knots are closed non-self-intersecting curves in three dimensional space. In order to classify knots, we require quantitative description which are called 'knot invariants.' Well-known knot invariants are Alexander polynomial, Jones' polynomial, Kauffman and HOMFLY-PT polynomials. We will briefly recapitulate the salient features of knots . Then we will present the computation of Jones' polynomial for a knot. As these knots can be obtained from braids, we can reproduce knot invariants using braid groups and their representations. Six vertex model corresponds to placing spin half states on the edges of a square lattice. Using the Boltzmann weights corresponding to the six vertex model, we review a braid group representation and the construction of Jones' polynomials. This procedure can be applied to higher spins placed on the edges of the square lattice which we will briefly summarize.

14.1 Introduction

In the 19th century, Lord Kelvin [1,2] thought that the atoms are to be viewed as vortex lines intertwined in the homogeneous ether medium. Contemporary physicist Tait attempted writing the periodic table of vortex like atoms as periodic table of knots leading to knot tables. Once the presence of ether medium was ruled out, the study of knots were no longer pursued by physicists. In fact, these ideas led to the emergence of a new research area called 'knot theory' investigated by mathematicians leading to the construction of knot polynomials [3–5]. The exciting work of Witten [6] showed that a topological quantum field theory known as 'Chern-Simons field theory' provided a natural arena for the study of knots. This probably motivated physicists to work again in the

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_14

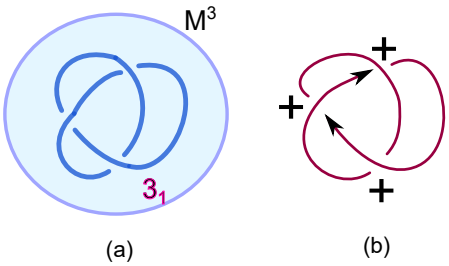


Figure 14.1: Trefoil knot

area of knot theory from late eighties. Particularly, the physicists attempted the construction of the knot polynomials from statistical mechanical models [7], topological quantum field theory and rational conformal field theory. Thus the area of knot theory has been a meeting ground for both physicists and mathematicians. Our theme in this set of lectures is to highlight the construction of knot polynomials from vertex models. In the following section, we will briefly recapitulate salient features of knots and links and the computation of knot polynomials.

14.2 Salient Features of Knots

Mathematically, a *knot* is an embedding of a circle S^1 in a three-manifold M^3. In Fig. 14.1(a), we have illustrated a knot called 'trefoil' embedded in a three manifold. Clearly the topological nature of the knot is independent of shape or size. *Links* are collection of many such non-intersecting knots. These knots can be drawn as projections onto a two-dimensional plane with over crossing and under crossings as drawn in Fig. 14.1(b). The number of minimal crossings is called crossing number. Rolfsen knot tables [8] gives the table of knots in the increasing order of crossing number. For instance, the three crossing trefoil is denoted as 3_1 and subscript denotes a specific inequivalent knot. There are in fact two inequivalent five crossing knots denoted as $5_1, 5_2$ in the Rolfsen table. Clearly, crossing number is not a good knot invariant.

In Fig. 14.1(b), we have placed an arrow(orientation) on the trefoil and assigned $\sigma = +1$ to each of the crossing. For a general knot, we can assign a sign $\sigma = \pm 1$ depending upon whether the crossing is right handed (+) or left handed (-). Writhe ω of any knot is defined as the sum of the crossing signs of the knot. Even though 'writhe' is a knot invariant, there are more than one sharing the same value. Therefore, writhe ω are referred to as 'weak invariants'.

It is one of the challenging problems to determine whether any two knots are topologically equivalent or inequivalent. For example, see Fig. 14.2(a) and deduce whether the two knots are equivalent or inequivalent. Interestingly, there

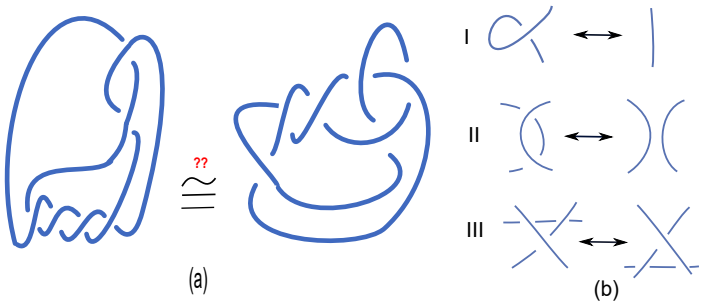


Figure 14.2: (a) Equivalence of knots through (b) Reidmeister moves

is a prescribed set of moves called Reidmeister moves as shown in Fig. 14.2(b) performed on one knot. If the second knot is obtained by this set of three moves, we consider that the two knots are equivalent. Definitely, this is a tedious process. We need a quantitative description for deducing topological equivalence of two knots. Such a quantity is called 'knot invariant'. The construction of knot invariant must incorporate the fact that the invariant remains same under these Reidmeister moves.

There are skein/recursion relation giving well known polynomial invariants for knots and links. They were put forth by mathematicians. Alexander polynomial $\Delta(K)$ [3] is the first one variable knot invariant which distinguished many inequivalent knots K. Unfortunately, the polynomial is same for many knots and their mirror images. After nearly 60 years, Jones' [4] came up with another recursive relation whose polynomials $J(K)$are different for many knots and their mirror images. Later on, two-variable generalization of skein relations of Jones' were given independently by six mathematicians and the polynomial is known as HOMFLY-PT [5] $P(K)$. Around the same time, unoriented knot polynomials with a modified recursion relation involving two variables was given by Kauffman [9].

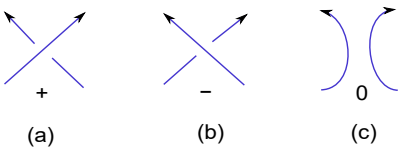


Figure 14.3: (a) Overcrossing (b) Undercrossing (c) No crossing

Concentrating on any particular crossing of a knot, we can write the skein relations between involving three knots K_+, K_- and K_0 where the subscript denotes overcrossing, undercrossing and no crossing as indicated in Fig. 14.3. The skein relations of Alexander polynomial, Jones' and HOMFLY-PT are as

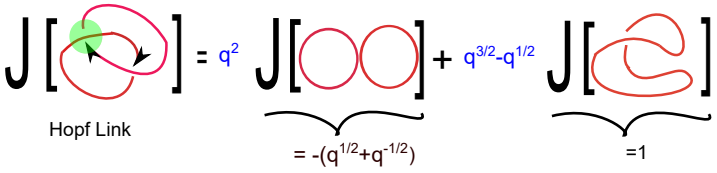
follows:

$$\Delta(K_+) - \Delta(K_-) = (q^{\frac{1}{2}} - q^{-\frac{1}{2}})\Delta(K_0) \quad (14.1)$$

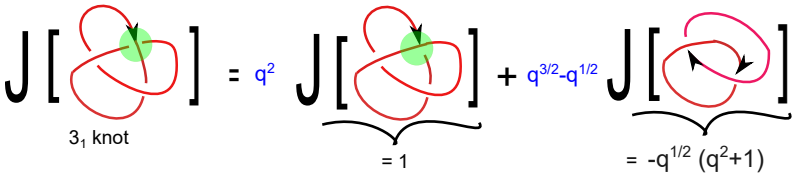
$$q^{-1}J(K_+) - qJ(K_-) = (q^{\frac{1}{2}} - q^{-\frac{1}{2}})J(K_0) \quad (14.2)$$

$$a^{-1}P(K_+) - a(K_-) = (q^{\frac{1}{2}} - q^{-\frac{1}{2}})P(K_0) \quad (14.3)$$

We observe from the tables for these well known knot polynomials, this recursive approach takes care of same polynomial invariants for equivalent knots but does not guarantee the converse. There are many inequivalent knots sharing the same polynomials. In this sense, these well known polynomial invariants do not give one to one onto map to inequivalent knots. This is the famous 'classification problem of knots' which calls for some radical approach. It appears that the methods in physics suggest a possibility leading to improving classification of knots. For completeness and some hands on exposure on evaluating Jones' polynomial, we show the calculation of trefoil knot polynomial. We normalize the unknot polynomial $J[\bigcirc] = 1$ and compute Hopf link and then the trefoil as illustrated below:



$$= -q^{1/2}(q^2+1)$$



$$= q^3+q-q^4$$

14.3 Knots from braids

These knots can be constructed from closure of braids. Take n identical particles on a plane, there are two independent exchanges between ith particle and $i+1$th particle (See Ref. [10]). In fact, the clockwise and anticlockwise exchange is indicated as under crossing and over crossing respectively. Mathematically,

we denote this exchange as b_i for clockwise exchange and inverse b_i^{-1} for anti-clockwise exchange. Any arbitrary exchange is given by a word involving b_i's and their inverses. These b_i's are called generators of a group B_n commonly known as braid group. All braid words are not independent and are subjected to the following defining relations:

$$b_i b_j = b_j b_i \text{ if } |\mathrm{i} - \mathrm{j}| > 1; \quad b_i b_{i+1} b_i = b_{i+1} b_i b_{i+1} \ . \tag{14.4}$$

We will see in the next section that the trilinear relation is similar to the Yang-Baxter equation which appears in the context of integrable vertex models [11].

The knots can be obtained from braids by identifying the initial state of n identical particles before exchange with the final state of these particles after exchange (also known as closure). This is sometimes denoted as a trace operation of a braid word.

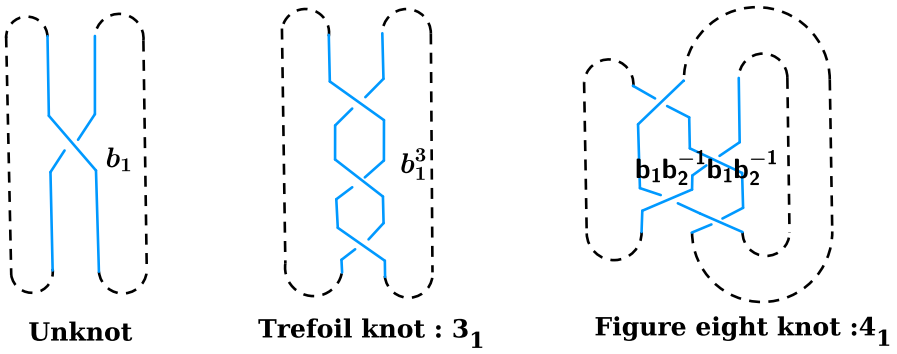


Figure 14.4: Closure of braids

We have illustrated above the closure of the braid word $b_1 \in B_2$ giving a simple circle known as 'unknot'. Similarly the braid word $b_1^3 \in B_2$ gives trefoil knot and braid word $b_1 b_2^{-1} b_1 b_2^{-1} \in B_3$ gives figure-eight knot.

It is important to stress that the braid to knot is not one to one but there is a definite set of moves called Markov moves which removes this non-uniqueness as indicated in Fig. 14.5. Concatenating braid word $A \in B_n$ with $B \in B_n$ will give braid word AB. Similarly concatenating B with A will give different braid word BA. These two braid words are related by Markov move I. Under closure, both braid words will give the same knot. Markov move II corresponds to combining $b_n^{\pm 1}$ to any braid word $A \in B_n$ and this operation will not change the knot.

Hence construction of knot invariants involves taking a braid group representation and defining a trace operation which is unchanged under Markov moves. In the next section we will elaborate on the braid group representations from vertex models and the formula enabling evaluation of knot invariants.

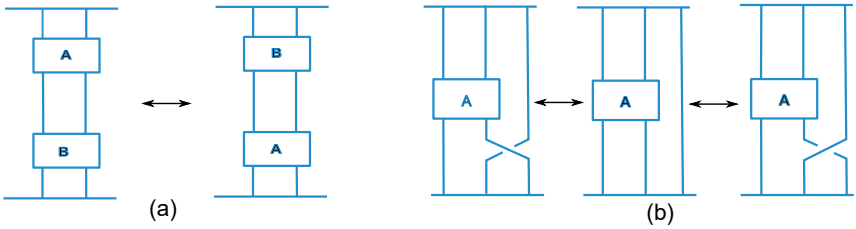


Figure 14.5: (a) Markov move I (b) Markov move II

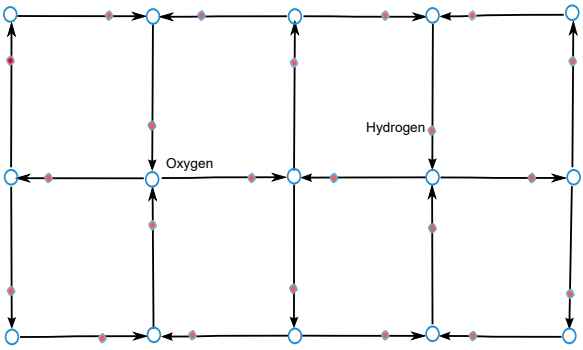


Figure 14.6: Ice-type model

14.4 Vertex model

In the following section, we will briefly review a statistical mechanical model with the aim of constructing braid group representation. This will enable computation of knot polynomials directly by postulating a suitable trace operation on the braid group elements which remain unaltered under Markov moves.

14.5 Ice Type model

Ice-type model is a well known square lattice model where every vertex in the lattice has an oxygen atom. The hydrogen atoms are placed on the four edges intersecting every vertex such that two of them are near and the other two atoms are far away from oxygen atom. This is depicted in Fig. 14.6 where the arrows provide equivalent presentation. That is, near hydrogen atoms is replaced by arrows going into the vertex and similarly far away hydrogen atoms is replaced by outgoing arrow from the vertex. Clearly, the arrows on the four

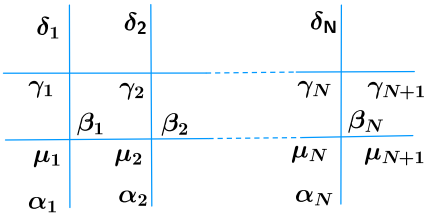


Figure 14.7: N-state Vertex model

edges intersecting every vertex satisfy the following arrow rule: the number of ingoing arrows equals the number of outgoing arrows.

The above lattice model can be generalized to N state vertex model where we place the possible states of spin $n = (N-1)/2$ on the edges denoted by Greek letter $\mu_i, \alpha_i, \beta_i, \nu_i, \gamma_i \in \{n, n-1, \ldots -n\}$ in Fig. 14.7. The arrow rule is generalized to the conservation of spin states for the edges intersecting every vertex. The Boltzmann weight associated to the vertex in Fig. 14.8 will be $w(\mu_1, \alpha_1|\beta_1, \mu_2)$ which is non-zero if and only if the spin conservation $\mu_1 + \alpha_1 = \beta_1 + \mu_2$ is satisfied. We will discuss in detail the vertex model with spin $n = 1/2$. There are two possible states at every edge denoted as $\pm \equiv \pm 1/2$. Spin conservation allows six possible vertex configurations as drawn in Fig. 14.9.

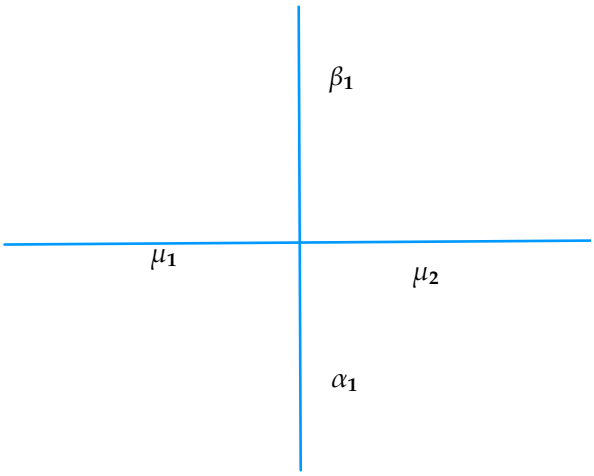


Figure 14.8: Boltzmann weight $w(\mu_1, \alpha_1|\beta_1, \mu_2)$ for the vertex

Sometimes, the $N = 2$ vertex model, where the edges have spin $1/2$ states, is also referred to as six vertex model as there are six non-zero Boltzmann weights.

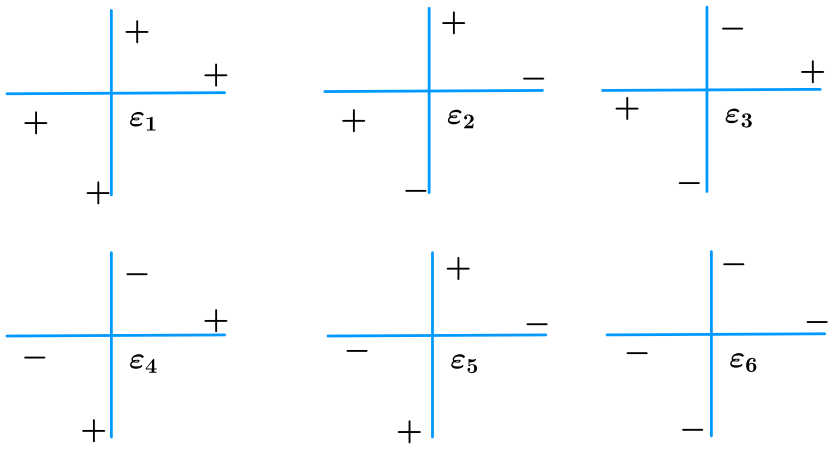


Figure 14.9: Allowed Boltzmann weights for six vertex model

14.5.1 Six vertex model

We will now study the construction of braid group representation from the Boltzmann weights of the vertex model. For completeness, let us tabulate the non-zero Boltzmann weights for this model after imposing symmetry(valid in the absence of external field) $w(\mu_1, \alpha_1|\beta_1, \mu_2) = w(-\mu_1, -\alpha_1| - \beta_1, -\mu_2)$:

$\mu_1, \alpha_1 \backslash \beta_1, \mu_2 \rightarrow$	$++$	$+-$	$-+$	$--$
$++$	a	0	0	0
$+-$	0	b	c	0
$-+$	0	c	b	0
$--$	0	0	0	a

where the explicit form of a, b, c in terms of energies ϵ_i at temperature $T = \beta^{-1}/k_B$ (k_B is Boltzmann constant) are

$$\begin{aligned} w(++|++) &= w(--|--) = a = \exp(-\epsilon_1\beta) \\ w(+-|+-) &= w(-+|-+) = b = \exp(-\epsilon_2\beta) \\ w(+-|-+) &= w(-+|+-) = c = \exp(-\epsilon_3\beta) \end{aligned} \tag{14.5}$$

Formally, the partition function Z describing the vertex model will be

$$Z = \sum_i e^{-\beta E_i}$$

where the summation i is over all possible configurations of the spin states $\pm$ on the edges and E_i is the energy for the ith configuration. Equivalently, we can introduce row to row transfer matrix $V_{\alpha\beta}$

$$V_{\boldsymbol{\alpha\beta}} = \sum_{\mu_1\mu_2\ldots\mu_N} w(\mu_1\ \alpha_1|\beta_1\ \mu_2)w(\mu_2\ \alpha_2|\beta_2\ \mu_3)\ldots$$

and rewrite Z as

$$Z = \sum_{\alpha_1^i\alpha_2^i\ldots\alpha_N^i} V_{\boldsymbol{\alpha^1\alpha^2}}\ V_{\boldsymbol{\alpha^2\alpha^3}}\ldots V_{\boldsymbol{\alpha^N\alpha^1}} = Tr(V^N) = \lambda_{max}^N$$

where λ_{max} is the maximum eigenvalue of the transfer matrix $V_{\alpha\beta}$. Attempting to find the eigenvectors and eigenvalues of the transfer matrix shows that the Boltzmann weights are parametrized in the following manner [11]:

$$a\ :\ b\ :\ c = \sinh(\lambda - u)\ :\ \sinh(\lambda)\ :\ \sinh(\lambda + u)$$

Further the solutions for the eigenvectors are dependent on

$$\Delta = (a^2 - b^2 + c^2)/(2ac) = -\cosh(\lambda) \tag{14.6}$$

This gives the freedom of choosing infinite number of transfer matrices $V(u)$ parametrized by u possessing the same eigenvectors. The parameter u is usually called spectral parameter. Equivalently, we can say that there are infinite family of commuting transfer matrices

$$V(u)V(v) = V(v)V(u)\ . \tag{14.7}$$

The above commutative condition on the transfer matrices will imply trilinear relation involving Boltzmann weights (see Chapter 9 in [11]) known as Yang-Baxter equation:

$$X_{i+1}(u)X_i(u+v)X_{i+1}(v) = X_i(v)X_{i+1}(u+v)X_i(u)\ , \tag{14.8}$$

where $X_i(u)$, known as Yang-Baxter operator, in terms of Boltzmann weights is

$$(X_i)_{\boldsymbol{\alpha\beta}} = \delta_{\alpha_1\beta_1}\delta_{\alpha_2\beta_2}\ldots w(\alpha_i, \alpha_{i+1}|\beta_i\beta_{i+1})\delta_{\alpha_{i+2}\beta_{i+2}}\ldots$$

Interestingly, the Yang-Baxter equation (14.8) resembles braid group relation (14.4) if we remove the dependence on the spectral parameter. We need to take a suitable limit of the spectral parameters and check whether the Yang-Baxter operator approaches a finite value. Taking the limit $u, v, u+v \to \infty$ on Yang-Baxter operator $X_i(u)$, we need to obtain finite operator:

$$\lim_{u\to\infty} X_i(u) = b_i$$

where b_i is the braid group generator. In order achieve such a finite operator b_i, we perform the following asymmetrization on the Boltzmann weights

$$\tilde{w}(\alpha_1, \alpha_2|\beta_1\beta_2)) = \exp\{\mu u(\alpha_2 - \alpha_1 - \beta_1 + \beta_2)\}w(\alpha_1, \alpha_2|\beta_1\beta_2) \tag{14.9}$$

which satisfies Yang-Baxter equation (14.8). We will take a suitable choice for the parameter μ so that we get a finite b_i as spectral parameter u tends to infinity. Using this asymmetrization condition, the new Boltzmann weights $\tilde{w}(\alpha_1, \alpha_2|\beta_1\beta_2)$ will be

$\alpha_1, \alpha_2 \backslash \beta_1\beta_2 \rightarrow$	$++$	$+-$	$-+$	$--$
$++$	a	0	0	0
$+-$	0	$\tilde{c}_1$	c	0
$-+$	0	c	$\tilde{c}_2$	0
$--$	0	0	0	a

where $\tilde{c}_1 = b\exp\{-2\mu u\}$; $\tilde{c}_2 = b\exp\{2\mu u\}$. Taking $u \rightarrow \infty$, $\mu = 1/2$ and $\exp(2\lambda) = q$ and suitable normalization, we get following finite limit on the Boltzmann weights:

$$\lim_{u\to\infty} \frac{w(\alpha_1, \alpha_2|\beta_1\beta_2)}{w(++|++)} = \sigma(\alpha_1, \alpha_2|\beta_1\beta_2)$$

which is tabulated below:

$\sigma(\alpha_1, \alpha_2|\beta_1\beta_2) =$

$\alpha_1, \alpha_2 \backslash \beta_1\beta_2$	$++$	$+-$	$-+$	$--$
$++$	1	0	0	0
$+-$	0	0	$-q^{1/2}$	0
$-+$	0	$-q^{1/2}$	$1-q$	0
$--$	0	0	0	1

The above 4×4 matrix satisfies the following relation:

$$\sigma^2 = (1-q)\sigma + q \ . \tag{14.10}$$

We can define the braiding generators b_i's involving σ- matrices as

$$b_i = \lim_{u\to\infty} X_i(u) = \mathbb{I} \otimes \mathbb{I} \ldots \otimes \mathbb{I}_{i-1} \otimes \sigma_{4\times 4} \otimes \mathbb{I}_{i+2} \otimes \ldots \tag{14.11}$$

From the above two equations, we can deduce

$$b_i^2 = (1-q)b_i + q\mathbb{I} \ .$$

Applying this relation on any undercrossing of a knot, drawn as closure of braid, will reproduce skein relation (14.3) giving Jones' polynomials.

So far, we have reviewed the derivation of matrix representation for a braid group generators from the six vertex model Boltzmann weights. We would like to attempt a direct construction of a polynomial invariant for knots without going through the skein relation approach. Basically, an algebraic formula for knots, obtained from braids, must be written such that the formula is unchanged when we perform Markov moves I and II. For a braid word $A \in B_n$

the following algebraic formula $\alpha(A)$ respects Markov moves and hence can be called knot invariant [7]:

$$\alpha(A) = (\tau\bar{\tau})^{-(n-1)} \left(\frac{\bar{\tau}}{\tau}\right)^{(e/2)} \phi(A) \tag{14.12}$$

where e is the exponent sum of the b_i's appearing in the braid word A, $\tau = 1/(1+q)$ and $\bar{\tau} = q/(1+q)$. Here $\phi(A)$, referred to as Markov trace, can be computed as

$$\phi(A) = \mathrm{Tr}\{HA\} \tag{14.13}$$

where H is a tensor product of 2×2 matrix h

$$H = \underbrace{h \otimes h \otimes h \otimes \ldots h}_{n}$$

with the entries of h matrix as follows:

$$h = \frac{1}{(1+q)} \begin{pmatrix} 1 & 0 \\ 0 & q \end{pmatrix} .$$

14.5.2 Knot Polynomials

We shall evaluate the knot polynomials directly using the above data in formula (14.12) for some knots drawn in Fig. 14.4. In fact, unknot and trefoil are obtained from closure of two-strand braid. So we will have only powers of the generator $b_1 \in B_{n=2}$ to compute $\alpha(A)$ where $A \in B_{n=2}$. For braid group B_2, $b_1 \equiv \sigma$:

$$b_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & -\sqrt{q} & 0 \\ 0 & -\sqrt{q} & 1-q & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

For trefoil, whose braid word is b_1^3, $e = 3$. It is straightforward to write the matrix for braid word b_1^3. Multiply matrix b_1^3 with $h \otimes h$

$$H = h \otimes h = \begin{pmatrix} \frac{1}{(1+q)^2} & 0 & 0 & 0 \\ 0 & \frac{q}{(1+q)^2} & 0 & 0 \\ 0 & 0 & \frac{q}{(1+q)^2} & 0 \\ 0 & 0 & 0 & \frac{q^2}{(1+q)^2} \end{pmatrix}$$

to give a 4×4 matrix whose trace will give $\phi(b_1^3)$. The final answer for the trefoil (Eq. 14.12) is

$$\alpha(b_1^3) \;=\; q + q^3 - q^4 \tag{14.14}$$

For unknot, whose braid word is $A = b_1$ in Fig. 14.4, the invariant $\alpha(b_1) = 1$.

The next knot in the knot table is figure-eight with four crossings. Here the braid word involves both b_1, b_2 and hence the braid group is $B_{n=3}$. The

explicit matrix form for the two generators ($b_1 = \sigma \otimes I_{2X2}, b_2 = I_{2X2} \otimes \sigma$) of $B_{n=3}$ will be 8×8 matrix:

$$b_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\sqrt{q} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\sqrt{q} & 0 & 0 \\ 0 & 0 & -\sqrt{q} & 0 & 1-q & 0 & 0 & 0 \\ 0 & 0 & 0 & -\sqrt{q} & 0 & 1-q & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} ;$$

$$b_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\sqrt{q} & 0 & 0 & 0 & 0 & 0 \\ 0 & -\sqrt{q} & 1-q & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\sqrt{q} & 0 \\ 0 & 0 & 0 & 0 & 0 & -\sqrt{q} & 1-q & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The 8×8 matrix form for the braid word $A = b_1 b_2^{-1} b_1 b_2^{-1}$, corresponding to figure-eight knot, can be obtained using the b_1 and b_2 matrices. Multiplying the matrix corresponding to A with the $H = h \otimes h \otimes h$

$$H = \begin{pmatrix} \frac{1}{(1+q)^3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{q}{(1+q)^3} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{q}{(1+q)^3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{q^2}{(1+q)^3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{q}{(1+q)^3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{q^2}{(1+q)^3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{q^2}{(1+q)^3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{q^3}{(1+q)^3} \end{pmatrix}$$

and taking trace will determine $\mathrm{Tr}(HA)$. The exponent sum $e = 0$ for figure-eight knot and the polynomial invariant turns out to be

$$\alpha(4_1) = 1 + q^{-2} - q^{-1} - q + q^2 \ . \tag{14.15}$$

We hope the direct evaluation of knot polynomials which we have elaborated for trefoil and figure-eight clarifies the necessary mathematical steps to determine polynomials for arbitrary knots obtained from braids. These polynomials, obtained from braid group representation derived using six-vertex model, are the Jones' polynomials.

This methodology can be generalized to N-state vertex model. We will briefly discuss the construction of braid group representation from $N = 3$ state vertex model where we place spin 1 states $\pm 1, 0$ on the edges of the square lattice. It is a simple exercise to check that there are nineteen allowed vertices obeying spin conservation for $N = 3$ vertex model. Hence nineteen non-zero Boltzmann weights leading to the nomenclature 'nineteen vertex model'.

14.5.3 Nineteen vertex model (Spin-1 particles)

Just like the way we did a suitable asymmetrization and normalization to six-vertex model Boltzmann weights to obtain braid generators b_i's, we could obtain the braid generators from nineteen vertex model Boltzmann weights. The matrix form for $\sigma(\alpha_1, \beta_1 | \alpha_2, \beta_2)$ will be 9×9 matrix

$$\sigma = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -q & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & q^2 & 0 & 0 \\ 0 & -q & 0 & 1-q^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & q & 0 & -\sqrt{q}+q^{5/2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -q & 0 \\ 0 & 0 & q^2 & 0 & -\sqrt{q}+q^{5/2} & 0 & 1-q-q^2+q^3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -q & 0 & 1-q^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The matrix h involved in Markov trace for the three state system will be 3×3 matrix:

$$h = \frac{1}{(1+q+q^2)} \begin{pmatrix} 1 & 0 & 0 \\ 0 & q & 0 \\ 0 & 0 & q^2 \end{pmatrix}$$

Similar to six vertex model, we will take $H = h \otimes h$ which is 9×9 matrix for evaluation of Markov trace $\mathrm{Tr}(HA)$ for braid words $A \in B_2$. The values of τ and $\bar{\tau}$ will be

$$\frac{1}{1+q+q^2}, \quad \text{and} \quad \frac{q^2}{1+q+q^2},$$

respectively. With this data, we can work out the knot polynomials for trefoil using the braid group representation obtained from nineteen vertex model Boltzmann weights.

For figure-eight, the braid word is an element of braid group B_3. We need to use 27×27 matrix representation for the two braiding generators: $b_1 = \sigma \otimes \mathbb{I}_{3\times 3}$ and $b_2 = \mathbb{I}_{3\times 3} \otimes \sigma$ and $H = h \otimes h \otimes h$ which is again 27×27. Using these matrix representation, knot invariant for any braid word $A \in B_3$ can be computed. For completeness, we tabulate the polynomials obtained from braid group representation using nineteen vertex model Boltzmann weights.

Knot	braid word A	n, e	$\alpha(A)$
Unknot	b_1	$n = 2, e = 1$	1
Trefoil	b_1^3	$n = 2, e = 3$	$q^2(1 + q^3 - q^5 + q^6 - q^7 - q^8 + q^9)$
Figure- 8	$b_1 b_2^{-1} b_1 b_2^{-1}$	$n = 3, e = 0$	$3 + q^{-6} - q^{-5} - q^{-4} + 2q^{-3} - q^{-2}$ $-q^{-1} - q - q^2 + 2q^3 - q^4 - q^5 + q^6$

These polynomials are different from Jones' polynomial and they agree with two-variable Kauffman polynomials for a specific choice of the variables. In ref. [12], for knots obtained from braid words in B_3, polynomials are computed and listed.

14.6 Summary and Discussions

In this chapter we have discussed the construction of braid group representation using the Boltzmann weights of N-state vertex model. In particular, we have elaborated on the matrix representation of braiding generators from six vertex model. Using an algebraic formula (Eq. 14.12) for knot invariants, we have explicitly evaluated Jones' polynomial for trefoil and figure-eight. Finally, we have indicated the approach for nineteen vertex models and tabulated the knot polynomials for unknot, trefoil and figure-eight.

There is a neat reverse method of constructing spectral parameter dependent Yang-Baxter operator for many new exactly solvable models beyond the already known six, nineteen, forty-four vertex models. This is achieved by using infinitely many new braiding eigenvalues obtained from the study of knots and links through Chern-Simons theory(see section 4 in Ref. [13]). In fact, this reverse approach suggest a way of deriving Yang-Baxter operators $X_i(u)$'s for new solvable models which appears difficult to be obtained by conventional methods [11].

Acknowledgments
PR would like to thank Somen Bhattacharjee for the excellent SERC school organized at SN Bose center (organized by RKM Vivekananda University, Belur

Math) during December 2015. PR would also like to thank Vivek Kumar Singh for converting the lectures delivered into a Latex file and drawing all the figures.

References

[1] William Thomas Kelvin, First Baron, Mathematical and Physical Papers, vol. IV, Hydrodynamics and General Dynamics, Cambridge University, Cambridge 1910.

[2] R. K. Kaul, "Topological quantum field theories: A Meeting ground for physicists and mathematicians", *Quantum field theory*, Ed. A. N. Mitra, p211-232; e-Print: hep-th/9907119.

[3] J.W. Alexander, Trans. Am. Math. Soc. **30**, 275 (1928).

[4] V.F.R. Jones, Bull. AMS 103-112 (1985); Ann. Math. **126**, 335-388 (1987).

[5] P. Freyd, D. Yetter, J. Hoste, W.B.R. Lickorish, K. Mill et and A. Ocneanu, Bull. AMS. **12**, 239 (1985); J.H. Przytycki and K.P. Traczyk, Kobe J. Math. **4**, 115 (1987).

[6] E. Witten, Commun. Math. Phys. **121**, 351-399 (1989).

[7] Y. Akutsu and M. Wadati, J. of Phys. Soc. of Japan **56**, 3039 (1987).

[8] D. Rolfsen, *Knots and links* (Publish or Perish, Berkeley, 1976).

[9] L.H. Kauffman, Topology **26**, 395 (1987).

[10] P. Ramadevi,"Exchange of Identical Particles", Resonance, Feb 2001, pg 23-28.

[11] R.J. Baxter, *Exactly Solved Models in Statistical Mechanics* (Academic Press, London 1982).

[12] Y. Akutsu, T. Deguchi and M. Wadati, J. of Phys. Soc. of Japan **56**, 3464 (1987).

[13] R.K. Kaul, "Chern-Simons theory, knot invariants, vertex models and three manifold invariants", Conference proceedings C96-12-12.1, p.45-63; e-Print: hep-th/9804122.

15

Concepts of polymer statistical topology

Sergei Nechaev

This chapter reviews a few conceptual steps in an analytic description of topological interactions involving topology and statistical physics of fluctuating non-phantom rope-like objects. The main ingredient here is the statistics of Brownian bridges in a non-Euclidean space of constant negative curvature. After an introduction to the role of knots and topological constraints in polymeric systems, following topics are discussed, (i) the conformal methods for entangled random walks, (ii) conditional Brownian bridges in hyperbolic spaces, and (iii) the crumpled globule phase of polymers.

15.1 What are we talking about?

How to peel an orange, without removing its skin? Can one make an omelet from unbroken eggs? Can one smoothly (without tops) comb hairy billiard ball (the sphere), or a donut (the torus)? Why cannot one tie a knot on a telegraph wire, stretched along the railway line? These and similar, often entertaining and seemingly naive, questions are directly related to topology. I will be interested in a rather narrow range of problems associated with the so-called low-dimensional topology, i.e., with the topology of systems containing long linear hurly-burly threads of different physical nature. As these objects, one can play with polymeric chains, vortex lines in superconductors, strings in quantum field theory, etc.

It is worth noting that the low-dimensional topology is a pretty insidious, or, better to say "serpentine" science, because the daily use of ropes and

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_15

wires, being trivial for "kitchen tasks", becomes almost useless for mathematical description of knots. Careful tying of shoe laces does not help much in constructing topological invariants, determination of physical and geometric properties of highly entangled polymeric networks, folded proteins, DNAs in chromosomes, etc. Moreover, even obvious notions from everyday's experience seem to be incorrect: for example, any child knows what a knotted rope is, and what its difference is from an unknotted one. However, this knowledge fails being translated into mathematically rigorous terms: one cannot define a knot on any open curve, and to talk seriously about knots, only closed paths should be dealt with. Indeed, having free ends of the thread, one can always transform any two arbitrary conformations of threads into each other by a continuous deformation. Therefore, for open threads it would be more correct to speak of quasi-knots (instead of knots) tied on a closed curve consisting of the thread itself, and, say, the segment joining its open extremities. In the case when the knot has a size substantially smaller than the thread's length, the difference between the knot and quasi-knot becomes negligible.

History has brought to us the name of one of the first topologists-experimenters, Alexander the Great (IV century BC), who, being unable to untie the Gordian knot, just slashed it with the sword. Surprisingly, modern algebraic topology partly borrows ideas of Alexander the Great to build topological invariants by splitting intersections of wires on a plane knot projection. For introduction to these constructions, please look up Ref [1].[1] Nobody knows how the Gordian knot looked exactly, however, there is an opinion that it resembled the so-called celtic knot, whose typical plane projection (knot diagram) is shown in the Fig.15.1.

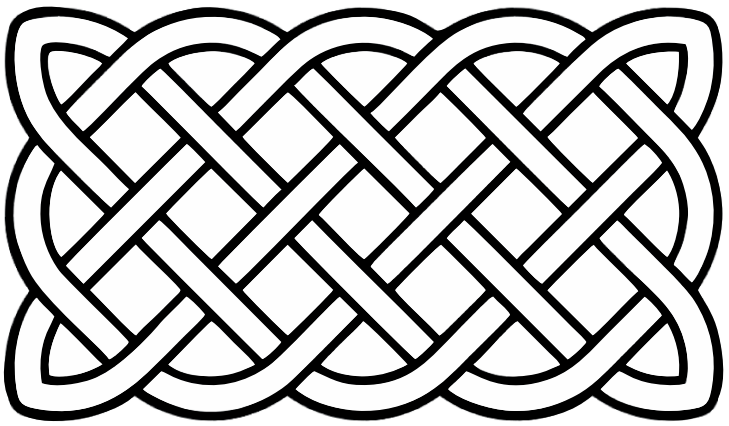


Figure 15.1: Sample of celtic knot projection to the plane (knot diagram).

[1] See Chaps.7 and 14 of this book.

Unknotting heavily tangled ropes, we are not even surprised that long threads, left to themselves, have the feature of getting entangled most unpleasant way of all. We just heave a bitter sigh, when an almost unravelled ball of wool accidentally falls into the clutches of a curious kitten, and our hard work goes to hell...

As a rule, we do not think about physical reasons for such "injustice", considering it as one of manifestations of the general law of life: unpleasant things happen more often than pleasant ones, and the sandwich usually falls butter side down. And we are quite aware that the spontaneous knotting of long strands, managed by the laws of probability theory, is a consequence of non-Euclidean geometry of the phase space of knots, i.e., such hypothetical space that contains all possible knots and in which there is a concept of the metric, or the distance between "topologically similar" knots. The more different topological types of two knots, the greater the distance between them in this space. In general, the construction of this space, called the universal covering, and the description of its properties are extremely challenging. However, for some special cases, to which I address a little later, this space can be exhaustively described in geometric terms. So, it turns out that, in order to untie ropes in a purely scientific way, one has first to understand the Lobachevsky-Riemann geometry and its relation to the knot theory, and then to learn how the probability theory works in this non-Euclidean space. Before going further, let me briefly digress and discuss some of physical questions in which the topological interactions play a crucial role.

Mathematicians primarily are interested in the question how to construct characteristics of entangled curves, that depend only on their topological state, but not on their shapes. Besides the traditional fundamental topological issues concerning the construction of new knot invariants, description of knot homologies, homotopic classes etc, there exists an important set of adjoint but much less studied problems related to probability theory and statistical physics. First of all, I mean the problem of so-called "knot entropy" evaluation. Suppose we know everything about a set of entangled curves and know how to classify their topology. In order to understand which issue is interesting for physicists, imagine that one has a sample of rubber, i.e., the collection of polymer chains crosslinked at their extremities such that they form a connected network. When the network is deformed as shown in the Fig.15.2, the ensemble of available conformations of each individual subchain in the network changes, resulting in an essential decrease of the entropy, S. So, one says that the elasticity of stretched rubber is of entropic nature.

Once being prepared, the network sample cannot change its topology, i.e., is *topologically frozen*. If the applied deformations do not break the subchains, the topological constraints like entanglements between different subchains, at high extensions enter into the game as new effective quasi-links, providing additional restrictions on the ensemble of available subchain conformations. Therefore, the topological constraints contribute to the stress-strain dependence of polymeric

Figure 15.2: Emotional view on elastic deformation. This sculpture, displayed at the Ujazdowski Castles Center for Contemporary Art in Warsaw (Poland), was made by Agnieszka Kalinowska in 2002.

networks, being the origin of so-called Mooney-Rivlin corrections to the classical Hooke's law [2].

In general, the non-phantomness of a polymer chain causes two types of interactions: i) volume interactions vanishing for infinitely thin chains, and ii) topological interactions, which survive even for chains of zero thickness. For sufficiently high temperatures, a polymer molecule strongly fluctuates without a reliable thermodynamic state called a coil state. However for temperatures below some critical value, θ, the polymeric chain forms a weakly fluctuating dense globular (drop-like) structure [2], and one may expect that just in the globular state the topological interaction manifest themselves in all their glory. The crucial difficulty in description of topological interactions comes from their non-locality: the entropic part of a polymeric chain free energy, $F = -TS$, strongly depends on the global chain topology. Saying more formally, the topological interactions in dense polymer systems cannot be treated in a *perturbative* way and new ideas of *nonperturbative* description are demanded.

The general problem we are dealing with, can be formulated as follows. Consider a 3D cubic grid, and let Ω be the ensemble of all possible closed non-self-intersecting N-step paths on this grid with one point fixed. Denote by ω ($\omega \in \Omega$) some particular realization of a path. Our aim is to calculate the partition function, $Z\{\mathrm{Inv}\}$, for a knot to be in a specific topological state characterized by the topological invariant Inv (yet non-specified). This can be

written formally as

$$Z\{\mathrm{Inv}\} = \sum_{\{\omega\in\Omega\}} \Delta\Big\{\mathrm{Inv}(\omega) - \mathrm{Inv}\Big\}, \tag{15.1}$$

where $\mathrm{Inv}(\omega)$ is a functional representation of the invariant for the path ω, Inv is a specific topological invariant of the knot which we would extract, and Δ is the Kronecker delta-function. The entropy, $S\{\mathrm{Inv}\}$, of a topological state, Inv, is defined as

$$S\{\mathrm{Inv}\} = \ln Z\{\mathrm{Inv}\}. \tag{15.2}$$

Based on the above definition, we can see that the statistical-topological problems are similar to those encountered in physics of disordered systems and in particular, of spin glasses. Indeed, the topological state of a path plays the role of a "quenched disorder" and the entropy, $S\{\mathrm{Inv}\}$, is averaged over the ensemble of trajectories fluctuating at the "quenched topological state" [3]. In the context of this analogy, it seems challenging to extend the concepts and methods developed over the years for disordered systems to the scopes of statistical topology. The main difference between the systems with topological disorder and the standard spin systems with disorder in the coupling constant, deals with the strongly nonlocal character of interactions in the first case: a topological state is determined for the entire closed path and is its "global" property.

Below I review a few conceptual steps in analytic description of topological interactions, which constitute the basis of a new interdisciplinary branch in mathematical physics, emerged at the edge of topology and statistical physics of fluctuating non-phantom rope-like objects. This new branch is called statistical (or probabilistic) topology. Yet its most fascinating manifestation is connected with the nonperturbative description of DNA packing in chromosomes in a form of a crumpled globule [4,5]. After experimental works of the MIT-Harvard team in 2009 [6], the concept of crumpled globule became a kind of a new paradigm allowing us to understand the mathematical origin of many puzzled features of DNA structuring and functioning in a human genome.

To intrigue the reader, I can say that the mathematical background of the crumpled globule deals with the statistics of Brownian bridges in the non-Euclidean space of constant negative curvature. Forthcoming sections are written to uncover this abracadabra.

15.2 Milestones

15.2.1 Abelian epoch

In 1967 S.F. Edwards laid the foundation of the statistical theory of topological interactions in polymer physics. In [7] he proposed the way of exact computation of the partition function of a single self-intersecting random walk topologically interacting with the infinitely long uncrossable string (in 3D), or

obstacle (in 2D). Sir Sam Edwards was the first to recognize the deep analogy between Abelian topological problems in statistical mechanics of Markov chains and quantum-mechanical problems (like Bohm-Aharonov ones) of the charged particles in the magnetic field. The review of classical results is given in physical context in [8], some rigorous results, including application in financial mathematics were discussed in [9], and modern advances are summarized in [10]. The works of S.F. Edwards opened the "Abelian" epoch in the statistical theory of topological interactions.

In his work S.F. Edwards used the path integral formalism combined with the functional representation of the Gaussian linking number. All these steps have been many times reproduced in the literature, so we do not discuss the details, just recall that one finally arrives at the quantum problem of a free charged particle (with an imaginary magnetic charge) in a solenoidal magnetic field. If the magnetic flux (the obstacle) is located at the origin and is orthogonal to the plane, then for the probability $P(r_1, r_N, \theta, n, N)$ to find an N-step polymer chain, whose extremities are located at the distances r_1 and r_N with $\widehat{\mathbf{r}_1, \mathbf{r}_N} = \theta$ and which makes n full turns around the origin, we get:

$$P(r_1, r_N, \theta, n, N) = \frac{1}{\pi N a^2}\, e^{-\frac{r_1^2 + r_N^2}{2Na^2}} \int_{-\infty}^{\infty} d\nu\, I_{|\nu|}\left(\frac{r_1^2 + r_N^2}{2Na^2}\right) e^{i\nu(2\pi n + \theta)}, \quad (15.3)$$

where $I_{|\nu|}(...)$ is the modified Bessel function of order $|\nu|$, and a is the size of the monomer (the typical step of the random walk). Obviously, the normalization condition is fulfilled,

$$\sum_{n=-\infty}^{\infty} P(r_1, r_N, \theta, n, N) = \frac{1}{\pi N a^2}\, e^{-\frac{(\mathbf{r}_1 - \mathbf{r}_N)^2}{Na^2}}, \quad (15.4)$$

which means that the summation over all windings, n, $(-\infty < n < \infty)$ gives the Gaussian distribution. I will reproduce the result Eq. (15.3) in the next Section using the conformal approach. Though it is less popular in polymer physics than the path-integral formalism, it can be straightforwardly generalized to the non-Abelian multi-obstacle case.

It should be noted that the exact computation of the partition function of the self-intersecting random walk topologically entangled with *two* uncrossable obstacles in the plane, despite the huge number of works since 1967, still is an open problem. However, apparently this gap will be filled soon, because its quantum-mechanical counterpart, the Abelian problem of Bohm-Aharonov scattering in presence of *two* magnetic fluxes in the plane, has been solved in 2015 by E. Bogomolny [11]. To my point of view, his solution has opened the Pandora's Box in the field, since he showed the deep mathematical connection of this particular problem to the theory of Painleve equations, integrable systems etc.

15.2.2 Non-Abelian epoch

Each time when we consider statistics of sufficiently dense polymer system, we encounter the extremely difficult problem of classification of topological states of polymer chains. Even the simplest physically relevant questions dealing with the knotting probability of a polymer chain, cannot be answered using the Gauss invariant due to its weakness. The Gauss linking number, becomes inapplicable since it does not reflect the sequence in which a given topological state was formed. For example, when some trial trajectory encloses two obstacles, the path is entangled with two obstacles simultaneously, while being not entangled with them separately, as shown in Fig.15.3.

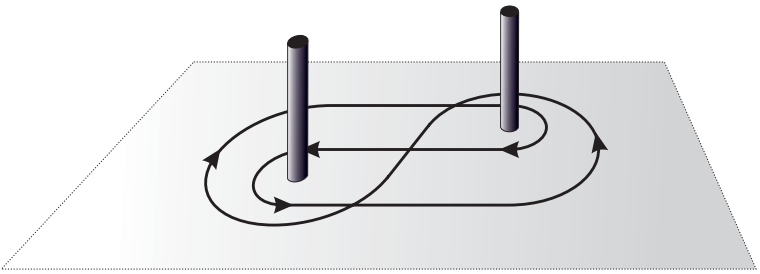


Figure 15.3: Topological configuration of a polymer entangled with two obstacles simultaneously but not entangled with any one of them separately (so-called Pochhammer contour).

Interestingly, I have not found in the literature any example of the path entangled with *three* obstacles simultaneously, while not entangled with any separate obstacle and any pairs of obstacles. Formally the question can be formulated as follows: find the element $X \in F_3$ of a free group of three generators, F_3, such that X belongs both to the commutant of F_3 and to the commutant of commutant of F_3.

For dense polymer systems, one encounters many configurations as shown in Fig.15.3, and they should be properly classified and treated using the invariants stronger than the Gauss linking number. So, to summarize, the Abelian (commutative) invariants become inapplicable for dense polymers and should be replaced by the non-Abelian (noncommutative) ones.

Methods of algebraic topology in polymer statistics. Topology as quenched disorder

A very useful and powerful method of knots classification has been offered by a polynomial invariant introduced by Alexander in 1928. The breakthrough in the field of polymer statistics was made in 1975-1976 when the algebraic polynomials were used for the topological state identification of closed random walks generated by the Monte-Carlo method [12]. It has been recognized that the Alexander polynomials being much stronger invariants than the Gauss linking

number, could serve as a convenient tool for the calculation of the thermodynamic properties of entangled random walks. To my point of view, 1975 is the year of the second birth of the probabilistic polymer topology, since the main part of our modern knowledge on knots and links statistics in dense polymer systems is obtained with the help of these works and their subsequent developments.

Other polynomial invariants for knots and links were suggested by V.F.R. Jones [13] (Jones polynomials) and by J. Hoste, A. Ocneanu, K. Millett, P.J. Freyd, W.B.R. Lickorish, and D.N. Yetter [14] (HOMFLY polynomials). The Jones invariant arise from the investigation of the topological properties of braids [15]. V.F.R. Jones succeeded in finding a profound connection between the braid group relations and the Yang-Baxter equations representing a necessary condition for commutativity of the transfer matrix offered the relation with integrable systems [16].[2] It should be noted that neither the Alexander, Jones, and HOMFLY invariants, nor their various generalizations are complete; however, these invariants are successfully used to solve many statistical problems in polymer physics. A clear geometric meaning of Jones invariant was provided by the works of Kauffman, who demonstrated that Jones invariant can be rewritten in terms of the partition function of the Potts spin model [1]. Later Kauffman and Saleur showed that the Alexander invariants are related to a partition function of the free fermion model [17]. The list of knot invariants used in polymer physics would be incomplete without mentioning Vassiliev invariants [18] and Khovanov homologies [19].

How to define the knot complexity?

There are many definitions of knot complexity. Some authors use the concept of *minimal number of crossings* [20–24]. In other works (see, for example [24, 25]) knot complexity is associated with a properly normalized logarithm of a kind of knot torsion, $\log|\Delta_K(-1)|$, where $\Delta_K(t)$ is the Alexander polynomial of the knot K. The estimate of knot complexity using the *knot energy* was discussed in the works [26–28] and to my point of view is yet underappreciated concept by polymer physicists.

Another approach deals with the fashionable concept of *knot inflation* [29]. This topological invariant is defined as the quotient, μ, of the contour length of the knot made of an elastic tube to its diameter in the maximal uniformly inflated configuration as shown in the Fig.15.4 for two different torus knots of same tube length. One sees that the more complex the knot is, the thinner is the limiting tube. Such approach was introduced and exploited in [5, 29].

Knot invariants like the minimal number of crossings, as well as those built on the basis of the knot inflation concept, are similar to the invariants defined as the degree of algebraic polynomial used in the works [30, 31]. All of them have one common ancestor – the so-called *primitive path*, appeared in

[2] See Chaps. 7, 14 of this book.

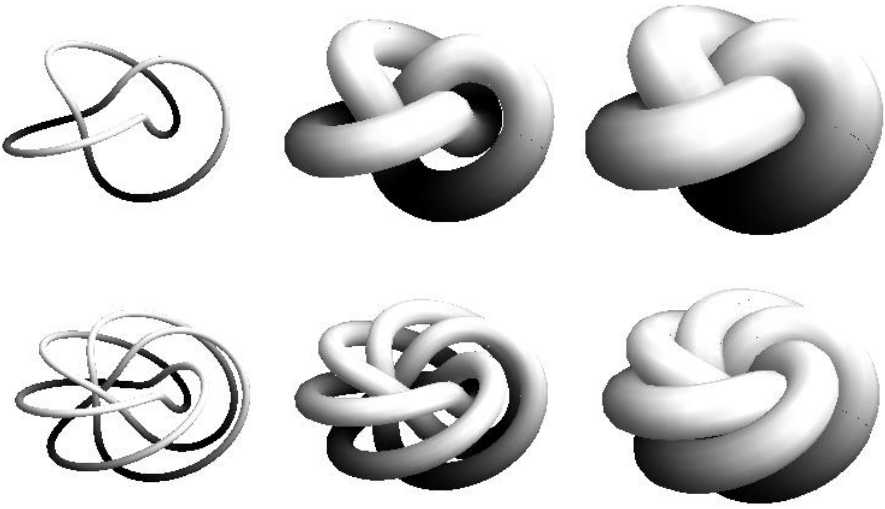


Figure 15.4: Affine inflation of a torus knot.

physical literature in 1970s in the works on entanglements in polymer melts. Introduced by P. de Gennes (see, for example [32]), the primitive path was served to describe topological effects in the dynamics of individual chains in concentrated polymer solutions. Later on, the same concept has been used in the computation of equilibrium properties of polymer chains in lattices of topological obstacles [33, 34] (I discuss this issue in detail in the next Section).

The notion of a primitive path and its relation to the knot inflation concept can be elucidated as follows. Consider a closed path of fixed length entangled with the lattice of obstacles (see Fig.15.5a). Performing an affine extension (inflation) of the lattice of obstacles (preserving the length of the path), one arrives eventually at the unfolded fully stretched configuration, see Figs.15.5b-c. Just the configuration in the Fig.15.5c is called the *primitive path* and it characterizes the topological state of a path with respect to the lattice of obstacles. Let us associate the properly normalized length of the chain and the spacing between obstacles, with the length of an elastic tube in the "knot inflation concept" and its diameter. In that way, the relationship between the primitive path, the minimal number of crossings and the quotient μ in the maximal inflated tube configuration becomes intuitively clear by construction. In more detail this relationship was discussed in [5].

In 1991 it was realized [35] that the concept of a primitive path has a straightforward interpretation in terms of a geodesics in a space of constant negative curvature. In the forthcoming section, we show how the geodesic length, in turn, may be related via its matrix representation to the degree of the polynomial invariant. Though our construction is restricted to the particular case of knots on narrow strips, the very idea can be used to attack more general

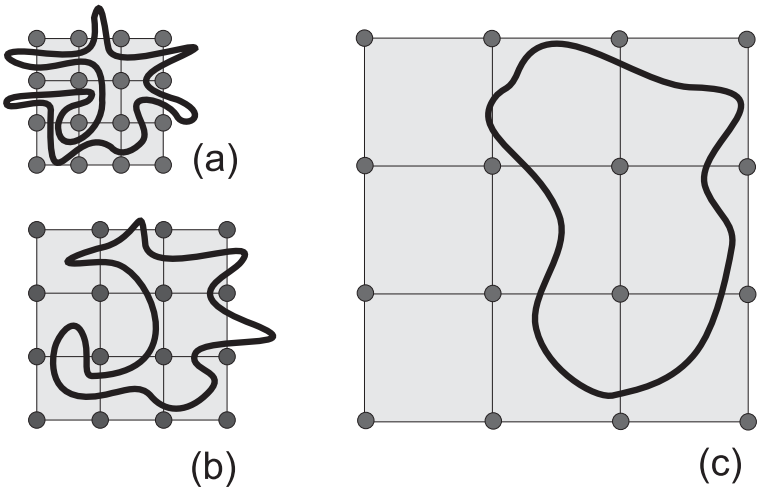


Figure 15.5: Affine inflation of a lattice of obstacles. The corresponding contour becomes "less folded".

models. The relationship between primitive paths and the maximal degree of the corresponding polynomial invariants has been discussed in [36] and will be overviewed in the next Section.

Conformal methods in statistics of entangled random walks

In 1985-1988 we have adopted the knowledge of stochastic processes in the Riemannian geometry to the statistical topology of polymers in multi-punctured spaces [34,37,38]. In particular, we have shown that the probability for a random walk (a polymer chain without volume interactions) to be unentangled with the regular lattice of topological obstacles in the plane is asymptotically described by the probability to form a Brownian bridge (BB) in the Lobachevsky plane (the non-Euclidean plane with a constant negative curvature). To have a transparent geometric image, though rather naive, the problem can be formulated as a "snake in a night forest". Suppose that a very long snake, lost in a dense forest at night, would like to grasp randomly its own tail in such a way, that the formed ring is not entangled with any tree. What is the probability, $P(L, d)$, of such an event, if L is the length of the snake and d is the average distance between the trees? What is the typical size, $R_g(N, d)$, of the snake (in polymer statistics, where a closed snake is replaced by a polymer ring, R_g is called the "gyration radius")?

Conformal methods provide straightforward answers to these questions. The key idea is to find the conformal transformation $w(z)$ which maps the complex plane, $z = x + iy$, with obstacles (branch points), to the "universal covering" space, $w = u + iv$, free of branchings in any finite domain. The main ingredient of this approach is the "conformal invariance" of Laplace operator,

$\Delta(z)$. Under the conformal mapping $w(z)$ the Laplacian $\Delta(z) = \partial^2_{xx} + \partial^2_{yy}$ is transformed to $\Delta(w) = \partial^2_{uu} + \partial^2_{vv}$ as follows

$$\Delta(w) = \partial^2_{uu} + \partial^2_{vv} = |w'(z)|^2(\partial^2_{xx} + \partial^2_{yy}), \tag{15.5}$$

where $w'(z) = \frac{dw(z)}{dz}$. Define also $z(w)$, the inverse function of $w(z)$. If we are lucky enough and found the desired conformal mapping $w(z)$, then, in the universal covering space, our initial topological problem looks formally extremely simple: we have just to solve the diffusion-like equation in time t with the diffusion coefficient D:

$$\partial_t P(w,t) - D|z'(w)|^{-2}(\partial^2_{uu} + \partial^2_{vv})P(w,t) = \delta(w - w_0)\delta(t), \tag{15.6}$$

without any topological constraints since all information about the topology is encoded now in the boundary conditions of the corresponding Cauchy problem in the covering space w. In the theory of stochastic processes Eq. (15.6) describes the diffusion in the "lifted" time, since it can be considered as a standard diffusion in the new metric-dependent time t, where $\partial_t = |z'(w)|^2\partial_t$. However the simplicity of Eq. (15.6) in majority of cases is rather illusory: finding conformal mapping $z(w)$, and then solving Eq. (15.6) analytically, both these tasks are challenging problems. Despite these hurdles, a few nontrivial cases can be treated and solved at least asymptotically.

To demonstrate how the method of conformal mappings works, let us return to the Abelian problem and reconsider entanglement of the random walk with the single obstacle in the plane. Place the obstacle (the branch point) at the origin, make a cut along the positive part of the x-axis of the complex plane $z = x + iy$ as shown in Fig. 15.6 and perform a conformal transform with the function $w(z) = \ln z$ to the universal covering space $w = u + iv$.

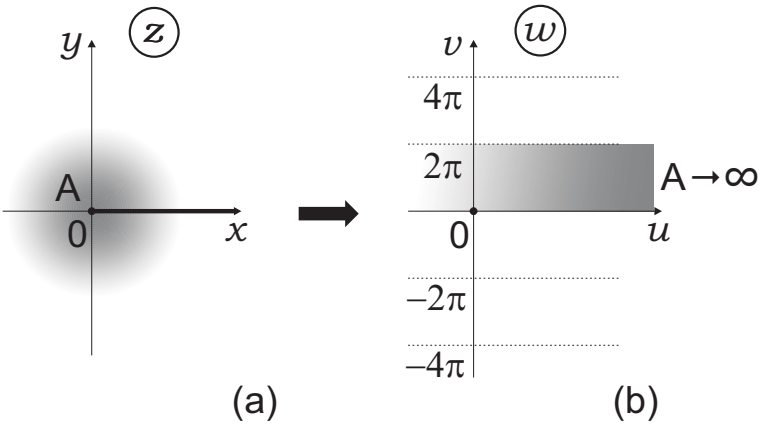


Figure 15.6: Conformal mapping of a plane with one branch point (obstacle) to the multi-sheet universal covering plane.

By elementary computations we get:

$$\begin{cases} u = \ln|z| \equiv \ln\rho, \quad v = \arg z, \\ |z'(w)|^{-2} = e^{-2u}. \end{cases} \tag{15.7}$$

If the path $\mathcal{C}$ makes n full turns around the branch point in z, it means that in the w-plane the distance between extremities of the image of $\mathcal{C}$ along the v-axis is $2\pi n$. So, closed paths crossing the cut in z are transformed into the open paths in w. The Cauchy problem in the lifted space is periodic in v and in each Riemann sheet (the strip of width 2π in the plane w) can be written as follows

$$\begin{cases} \partial_t P(u,v,t) - D\, e^{-2u}(\partial^2_{uu} + \partial^2_{vv})P(u,v,t), \\ P(u_0,v_0,t) = \delta(u-u_0)\delta(v-v_0)\delta(t), \end{cases} \tag{15.8}$$

where $\delta(...)$ is the Dirac δ-function. Making use of the substitution $\rho = e^u$, and taking into account the periodicity in v, we can rewrite Eq. (15.8) in the following form

$$\partial_t P(\rho,v,t) + D\left(\partial^2_{\rho\rho} + \frac{\partial_\rho}{\rho} + \frac{\partial^2_{vv}}{\rho^2}\right)P(\rho,v,t) = \frac{\delta(\rho-\rho_0)}{(\rho\rho_0)^{1/2}}\delta(v-v_0)\delta(t). \tag{15.9}$$

Seeking the solution of Eq. (15.9) in the form

$$P(\rho,v,t) = \sum_{m=-\infty}^{\infty} P_m e^{im(v-v_0)}, \tag{15.10}$$

we get for P_m the expression

$$P_m = \frac{1}{\pi t a^2} e^{-\frac{\rho_0^2+\rho^2}{ta^2}} I_{|\nu+m|}\left(\frac{2\rho\rho_0}{ta^2}\right), \tag{15.11}$$

where we have taken into account the expression for the diffusion coefficient in form $D = \frac{a^2}{4}$. Since

$$\sum_{m=-\infty}^{\infty} e^{im(v-v_0)} = 2\pi \sum_{n=-\infty}^{\infty} \delta(v+2\pi n - v_0), \tag{15.12}$$

we arrive at Eq. (15.3), where we should make the replacements $\theta \leftrightarrow v$ and $N \leftrightarrow t$.

Now we are in position to attack our favourite problem – finding the probability that the random walk of length t not entangled with respect to the triangular lattice of obstacles in the complex plane z, as shown in Fig. 15.7a. To solve the problem, we should construct the conformal mapping of the multi-punctured plane z to the universal covering space free of obstacles w and take the corresponding Jacobian of transformation, $|z(w)|^{-2}$. In this particular case

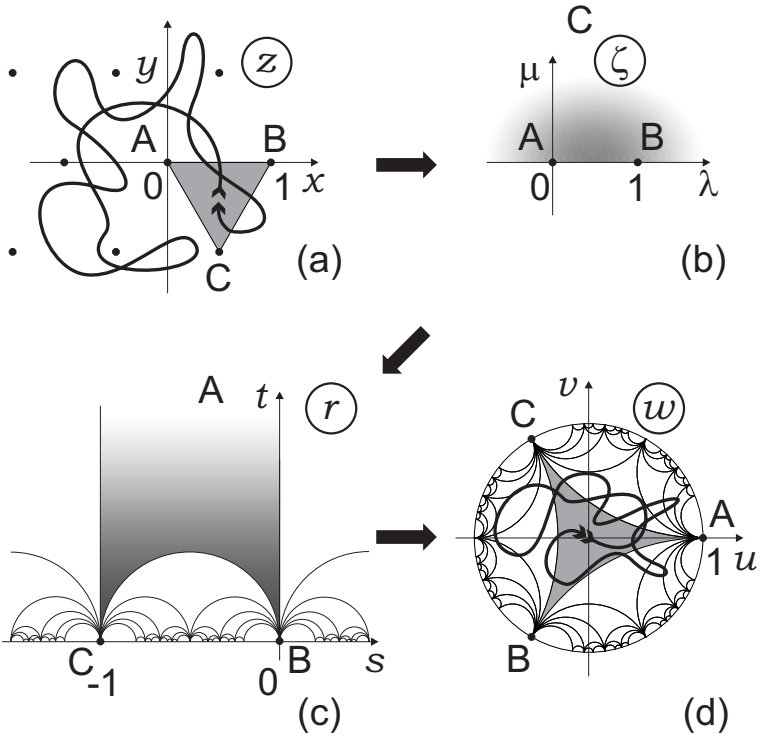


Figure 15.7: Conformal mapping $z(w)$ is realized as a composition of three mappings: $z(\zeta)$ [(a)–(b)], $\zeta(r)$ [(b)–(c)], and $r(w)$ [(c)–(d)]. Finally we have $z(\zeta(r(w)))$.

finding such a mapping is more elaborated task than for one-punctured plane, though still doable. The conformal mapping $z(w)$ of the flat equilateral triangle ABC located in z onto the zero-angled triangle ABC in w, is constructed in three sequential steps, shown in Fig. 15.7a-d.

First, we map the triangle ABC in z onto the upper half-plane ζ of auxiliary complex plane ζ with three branch points at 0, 1 and ∞ – see Fig. 15.7a-b. This mapping is realized by the function $z(\zeta)$:

$$z(\zeta) = \frac{\Gamma(\frac{2}{3})}{\Gamma^2(\frac{1}{3})} \int_0^\zeta \frac{d\xi}{\xi^{2/3}(1-\xi)^{2/3}}, \tag{15.13}$$

with the following coincidence of branching points:

$$\begin{cases} A(z=0) & \leftrightarrow \quad A(\zeta=0), \\ B(z=1) & \leftrightarrow \quad B(\zeta=1), \\ C(z=e^{-i\frac{\pi}{3}}) & \leftrightarrow \quad C(\zeta=\infty). \end{cases} \tag{15.14}$$

Second step consists in mapping the auxiliary upper half-plane $\mathrm{Im}\,\zeta > 0$ onto the circular triangle ABC with angles $\{\alpha, \alpha, 0\}$ – the fundamental domain of the Hecke group [39] in r, where we are interested in the specific case $\{\alpha, \alpha, 0\} = \{0, 0, 0\}$ – see Fig. 15.7b-c. This mapping is realized by the function $\zeta(r)$, constructed as follows [40]. Let $\zeta(r)$ be the inverse function of $r(\zeta)$ written as a quotient

$$r(\zeta) = \frac{\phi_1(\zeta)}{\phi_2(\zeta)}, \tag{15.15}$$

where $\phi_{1,2}(\zeta)$ are the fundamental solutions of the 2nd order differential equation of Picard-Fuchs type:

$$\zeta(\zeta-1)\phi''(\zeta) + \big((a+b+1)\zeta - c\big)\phi'(\zeta) + ab\phi(\zeta) = 0. \tag{15.16}$$

Following [40,41], the function $r(\zeta)$ conformally maps the generic circular triangle with angles $\{\alpha_0 = \pi|c-1|, \alpha_1 = \pi|a+b-c|, \alpha_\infty = \pi|a-b|\}$ in the upper halfplane of w onto the upper halfplane of ζ. Choosing $\alpha_\infty = 0$ and $\alpha_0 = \alpha_1 = \alpha$, we can express the parameters (a, b, c) of Eq. (15.16) in terms of α, taking into account that the triangle ABC in Fig. 15.7c is parameterized as follows $\{\alpha_0, \alpha_1, \alpha_\infty\} = \{\alpha, \alpha, 0\}$ with $a = b = \frac{\alpha}{\pi} + \frac{1}{2}, c = \frac{\alpha}{\pi} + 1$. This leads us to the following particular form of Eq. (15.16)

$$\zeta(\zeta-1)\phi''(\zeta) + \Big(\frac{\alpha}{\pi}+1\Big)\big(2\zeta-1\big)\phi'(\zeta) + \Big(\frac{\alpha}{\pi}+\frac{1}{2}\Big)^2\phi(\zeta) = 0, \tag{15.17}$$

where $\alpha = \frac{\pi}{m}$ and $m = 3, 4, ...\infty$. For $\alpha = 0$, Eq. (15.17) takes an especially simple form, known as Legendre hypergeometric equation [42,43]. The pair of possible fundamental solutions of Legendre equation are

$$\begin{aligned} \phi_1(\zeta) &= F\big(\tfrac{1}{2}, \tfrac{1}{2}, 1, \zeta\big), \\ \phi_2(\zeta) &= iF\big(\tfrac{1}{2}, \tfrac{1}{2}, 1, 1-\zeta\big), \end{aligned} \tag{15.18}$$

where $F(...)$ is the hypergeometric function. From Eq. (15.15) and Eq. (15.18) we get $r(\zeta) = \frac{\phi_1(\zeta)}{\phi_2(\zeta)}$. The inverse function $\zeta(r)$ is the so-called modular function, $k^2(r)$ (see [42–44] for details). Thus,

$$\zeta(r) \equiv k^2(r) = \frac{\theta_2^4(0, e^{i\pi r})}{\theta_3^4(0, e^{i\pi r})}, \tag{15.19}$$

where θ_2 and θ_3 are the elliptic Jacobi θ-functions [44,45],

$$\begin{aligned} \theta_2\left(\chi, e^{i\pi w}\right) &= 2e^{i\frac{\pi}{4}r}\sum_{n=0}^{\infty} e^{i\pi r n(n+1)}\cos(2n+1)\chi, \\ \theta_3\left(\chi, e^{i\pi r}\right) &= 1 + 2\sum_{n=0}^{\infty} e^{i\pi r n^2}\cos 2n\chi, \end{aligned} \tag{15.20}$$

and the correspondence of branch points in the mapping $\zeta(r)$ is as follows

$$\begin{cases} A(\zeta=0) & \leftrightarrow & A(r=\infty), \\ B(\zeta=1) & \leftrightarrow & B(r=0), \\ C(\zeta=\infty) & \leftrightarrow & C(r=-1). \end{cases} \tag{15.21}$$

Third step, realized via the function $r(w)$, consists in mapping the zero-angled triangle ABC in r into the symmetric triangle ABC located in the unit disc w – see Fig. 15.7c-d. The explicit form of the function $r(w)$ is

$$r(w) = e^{-i\pi/3}\frac{e^{2i\pi/3}-w}{1-w} - 1, \tag{15.22}$$

with the following correspondence between branching points:

$$\begin{cases} A(r=\infty) & \leftrightarrow & A(w=1), \\ B(r=0) & \leftrightarrow & B(w=e^{-2\pi i/3}), \\ C(r=-1) & \leftrightarrow & C(w=e^{2\pi i/3}). \end{cases} \tag{15.23}$$

Collecting Eqs. (15.13), (15.19), and (15.22) we arrive at the following expression for the derivative of composite function,

$$z'(\zeta(r(w))) = z'(\zeta)\,\zeta'(r)\,r'(w), \tag{15.24}$$

where $'$ stands for the derivative. We have explicitly:

$$z'(\zeta) = \frac{\Gamma(\frac{2}{3})}{\Gamma^2(\frac{1}{3})}\,\frac{\theta_3^{16/3}(0,\zeta)}{\theta_2^{8/3}(0,\zeta)\,\theta_0^{8/3}(0,\zeta)},$$

and

$$\zeta'(r)| = i\pi\frac{\theta_2^4\,\theta_0^4}{\theta_3^4}; \qquad i\frac{\pi}{4}\theta_0^4 = \frac{d}{d\zeta}\ln\left(\frac{\theta_2}{\theta_3}\right).$$

The identity

$$\theta_1'(0,e^{i\pi\zeta}) \equiv \left.\frac{d\theta_1(\chi,e^{i\pi\zeta})}{d\chi}\right|_{\chi=0} = \pi\theta_0(\chi,e^{i\pi\zeta})\,\theta_2(\chi,e^{i\pi\zeta})\,\theta_3(\chi,e^{i\pi\zeta}), \tag{15.25}$$

enables us to write

$$|z'(r)|^2 = h^2\left|\theta_1'\left(0,e^{i\pi r}\right)\right|^{8/3}, \tag{15.26}$$

where $h = \left(\frac{16}{\pi}\right)^{1/3}\frac{\Gamma(\frac{2}{3})}{\Gamma^2(\frac{1}{3})} \approx 0.325$, and

$$\theta_1(\chi,e^{i\pi r}) = 2e^{i\frac{\pi}{4}r}\sum_{n=0}^{\infty}(-1)^n e^{i\pi n(n+1)r}\sin(2n+1)\chi. \tag{15.27}$$

Differentiating Eq. (15.22), we get

$$r'(w) = \frac{i\sqrt{3}}{(1-w)^2},$$

and using this expression, we obtain the final form of the Jacobian of the composite conformal transformation $J(z(\zeta(r(w))))$:

$$J(z(w)) = |z'(w)|^2 = 3h^2 \frac{|\eta(r(w))|^8}{|1-w|^4}, \tag{15.28}$$

where

$$\eta(r) = \left(\theta_1'(0, e^{i\pi r})\right)^{1/3},$$

is the Dedekind η-function

$$\eta(w) = e^{\pi i w/12} \prod_{n=0}^{\infty} (1 - e^{2\pi i n w}), \quad (w = u + iv), \tag{15.29}$$

and the function $r(w)$ is defined in Eq. (15.22).

Thus, we arrive at the diffusion equation in the unit disc $|w| < 1$:

$$\partial_t P(w,t) - \frac{a^2}{4} J(z(w)) \left(\partial_{uu}^2 + \partial_{vv}^2\right) P(w,t) = \delta(w - w_0)\delta(t), \tag{15.30}$$

with the function $J(z(w))$ given by Eq. (15.28). The probability to find the two-dimensional random walk unentangled with the lattice of obstacles after time t, is given by the solution of Eq. (15.30), where we should plug at the very end $w = w_0 = 0$. The probability of returning to the same point $w_0 = 0$ in the initial Riemann sheet (the gray triangle in Fig. 15.7d) ensures that the trajectory is: i) closed, and ii) "contractible" (i.e., topologically trivial with respect to the lattice of obstacles).

Exact solution of Eq. (15.30) is unknown, however its asymptotic behavior we can extract relying on modular properties of Dedekind η-function Eq. (15.29). Consider the normalized Jacobian, defined as follows:

$$f(w) = J(z(w)) \left(1 - (u^2 + v^2)\right)^2, \tag{15.31}$$

where $|z| = \sqrt{u^2 + v^2}$ and $\psi = \arg z$ are radial and angular coordinates in the Poincaré unit disc disc $|w| < 1$. In Fig. fig:07a we have shown the density plot of the function $f(w)$ within the unit disc $|w| < 1$ for $f(w) > f_0 = 0.15$. As one can see from Fig. 15.8, the function $f(w)$ has identical local maxima in all the centers of circular triangles shown in Fig. 15.7d. Thus, we conclude that the Jacobian $J(z(w))$ in the centers of circular cells (domains) coincides with the metric of the Lobachevsky plane in the Poincaré disc,

$$ds^2 = \frac{du^2 + dv^2}{\left(1 - (u^2 + v^2)\right)^2} \tag{15.32}$$

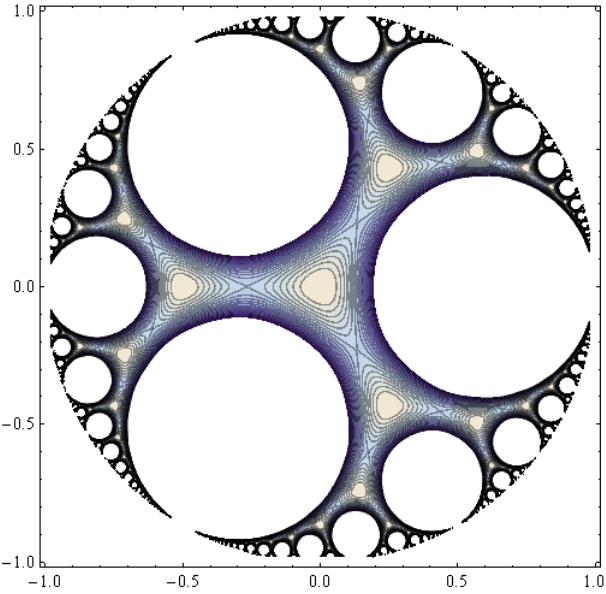


Figure 15.8: Density plot of the function $f(w)$ (see Eq. (15.31)) above the level $f_0 = 0.15$.

Note that the profile shown in Fig. 15.8 has striking similarities with the 3-branching Cayley tree and can be interpreted as a "continuous Cayley tree".

If we slightly modify our random walk model, we can exploit the connection with the Lobachevsky geometry. Namely, consider the random walk which stays in the vicinity of the center of a circular cell and then rapidly jumps to the center of the neighboring cell, stays there, then jumps again, and so forth... Such a random jumping process we can approximately describe by the diffusion in the Poincaré disc with the Lobachevsky plane metric Eq. (15.32) and the diffusion coefficient $\mathcal{D}$:

$$\partial_t P(w,t) - \mathcal{D}\big((1-(u^2+v^2)\big)^2\big(\partial^2_{uu}+\partial^2_{vv}\big)P(w,t) = \delta(u)\delta(v)\delta(t). \qquad (15.33)$$

Making use of the change of variables $(r,\psi) \to (\rho,\psi)$, where $\rho = \ln\frac{1+r}{1-r}$, we get the unrestricted random walk on the surface of the one-sheeted hyperboloid, $\mathcal{H}$ obtained by the stereographic projection from the Poincaré unit disc. Correspondingly the probability $P(\rho,t)$ reads

$$P(\rho,t) = \frac{e^{-\frac{t\mathcal{D}}{4}}}{4\pi\sqrt{2\pi(t\mathcal{D})^3}}\int_\rho^\infty \frac{\xi\exp\left(-\frac{\xi^2}{4t\mathcal{D}}\right)}{\sqrt{\cosh\xi-\cosh\rho}}d\xi. \qquad (15.34)$$

The physical meaning of the geodesic length, ρ, on $\mathcal{H}$ is straightforward: ρ is the length of the primitive path in the lattice of obstacles, i.e., the length of the shortest trajectory remaining after all topologically allowed contractions of

the random path in the lattice of obstacles. Hence, ρ can be considered a non-Abelian topological invariant, more powerful than the Gauss linking number. This invariant is not complete except one point, $\rho = 0$, where it precisely classifies the paths belonging to the trivial homotopic class in the lattice of obstacles.

Group-theoretic methods in statistics of entangled random walks

Non-Abelian entanglement of the path with two obstacles on the plane can be treated in the group-theoretic setting. Let us associate the generators g_1 and g_2 with the clockwise full turns around obstacles 1 and 2 respectively, and the inverse generators g_1^{-1} and g_2^{-1} – around the counterclockwise turns around 1 and 2, as shown in Fig. 15.9. Suppose that $g_1, g_2, g_1^{-1}, g_2^{-1}$ are the generators of the free group Γ_2 which by definition has no commutation relations. Thus, the possible contractions in the group Γ_2 are $g_1 g_1^{-1} = g_1^{-1} g_1 = g_2 g_2^{-1} = g_2^{-1} g_2 = I$, where I is the identity element.

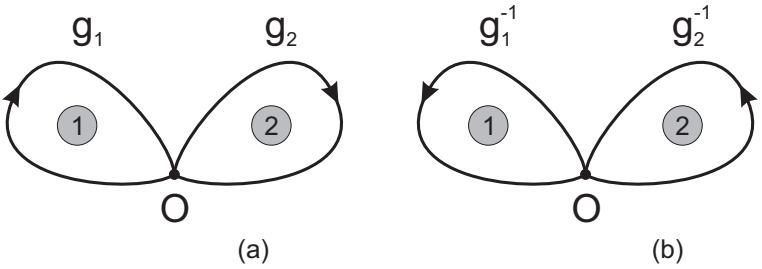


Figure 15.9: Generators of free group Γ_2.

Any closed path entangled with the obstacles 1 and 2 can be topologically presented by a word written in terms of generators $g_1, g_2, g_1^{-1}, g_2^{-1}$. For example, the Pochhammer contour shown in Fig. 15.3 reads: $g_1 g_2 g_1^{-1} g_2^{-1}$. Since the group Γ_2 is *noncommutative* (non-Abelian), we have $g_1 g_2 \neq g_2 g_1$, thus we cannot exchange the sequence of letters and replace $g_1 g_2$ by $g_2 g_1$. However, in the *commutative* (Abelian) group generated by the set $\{f_1, f_2, f_1^{-1}, f_2^{-1}\}$, we can do so, since $f_1 f_2 = f_2 f_1$, and the Pochhammer contour in the Abelian representation becomes contractible:

$$f_1 f_2 f_1^{-1} f_2^{-1} = f_2 \underbrace{f_1 f_1^{-1}}_{I} f_2^{-1} = \underbrace{f_2 f_2^{-1}}_{I} = I.$$

Suppose now that we have a set $\mathcal{S} = \{g_1, g_2, g_1^{-1}, g_2^{-1}\}$ and write random words of N letters by sequential addition of generators from the set $\mathcal{S}$. Each generator in $\mathcal{S}$ we take with the probability $p = \frac{1}{4}$. We are interested in computing the partition function $Z_N(x)$ for all N-letter random words to have the irreducible ("primitive") word of x letters. The partition function $Z_N(0)$ gives the number of N-letter words that are completely reducible (i.e., unentangled)

with the obstacles 1 and 2. The word counting problem in the free group Γ_2 can be visualized as the trajectories (built by sequential adding of letters) on the 4-branching Cayley tree shown in Fig. 15.10a. The irreducible (primitive) word is the shortest "bare" path along the tree connecting the extremities of the trajectory.

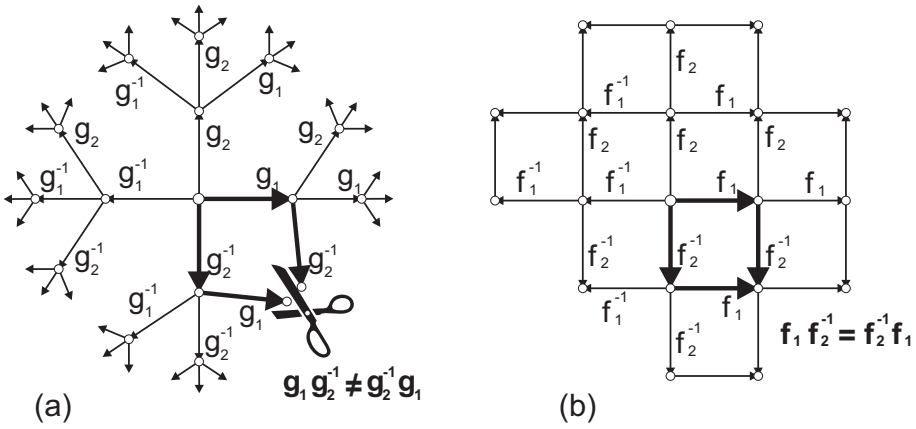


Figure 15.10: Visualization of commutation relations in free (a) and commutative (b) groups.

The partition function, $Z_N(x)$, of all N-step paths on the 4-branching Cayley tree, starting at the origin and ending at some distance x from it, satisfies the following recursion relation

$$\begin{cases} Z_{N+1}(x) &= 3Z_N(x-1) + Z_N(x+1), & x \geq 2, \\ Z_{N+1}(x) &= 4Z_N(x-1) + Z_N(x+1), & x = 1, \\ Z_{N+1}(x) &= Z_N(x+1), & x = 0, \\ Z_N(x) &= 0, & x \leq -1, \\ Z_{N=0}(x) &= \delta_{x,0}, \end{cases} \tag{15.35}$$

where x is the distance (the bare path) from the root of the Cayley graph, measured in number of generations of the tree. By making a shift $x \to x+1$, one can rewrite Eq. (15.35) as

$$\begin{cases} Z_{N+1}(x) &= 3Z_N(x-1) + Z_N(x+1) + \delta_{x,2}\, Z_N(x-1), & x \geq 1, \\ Z_N(x) &= 0, & x = 0, \\ Z_{N=0}(x) &= \delta_{x,1}, \end{cases} \tag{15.36}$$

where $\delta_{x,y}$ is the Kronecker δ–function: $\delta_{x,y} = 1$ for $x = y$ and $\delta_{x,y} = 0$ for $x \neq y$.

Eq. (15.36) can be symmetrized by the substitution

$$Z_N(x) = A^N B^x W_N(x). \tag{15.37}$$

Selecting $A = B = \sqrt{3}$, we arrive at the equation

$$\begin{cases} W_{N+1}(x) &= W_N(x-1) + W_N(x+1) + \dfrac{1}{3}\,\delta_{x,2}\,W_N(x-1), & x \geq 1 \\ W_n(x) &= 0, & x = 0 \\ W_{N=0}(x) &= \dfrac{\delta_{x,1}}{\sqrt{3}}. \end{cases} \tag{15.38}$$

Introducing the generating function

$$\mathcal{W}(s,x) = \sum_{N=0}^{\infty} W_N(x) s^n, \qquad \left(W_N(x) = \frac{1}{2\pi i} \oint \mathcal{W}(s,x) s^{-N-1}\, ds \right), \tag{15.39}$$

and its sin–Fourier transform

$$\widetilde{\mathcal{W}}(s,q) = \sum_{x=0}^{\infty} \mathcal{W}(s,x) \sin qx, \qquad \left(\mathcal{W}(s,x) = \frac{2}{\pi} \int_0^{\pi} \widetilde{\mathcal{W}}(s,q) \sin qx\, dq \right), \tag{15.40}$$

one obtains from Eq. (15.38)

$$\frac{\widetilde{\mathcal{W}}(s,q)}{s} - \frac{\sin q}{s\sqrt{p-1}} = 2\cos q\, \widetilde{\mathcal{W}}(s,q) + \frac{2}{\pi}\frac{1}{3} \sin 2q \int_0^{\pi} \widetilde{\mathcal{W}}(s,q) \sin q\, dq. \tag{15.41}$$

Solving Eq. (15.41) and performing the inverse Fourier transform, we arrive at the following explicit expression for the generating function $\mathcal{W}(s,x)$:

$$\begin{aligned} \mathcal{W}(s,x) &= \frac{2}{\pi} \int_0^{\infty} \widetilde{\mathcal{W}}(s,q) \sin qx\, dq \\ &= \frac{1}{s\sqrt{3}} \left(\frac{1-\sqrt{1-4s^2}}{2s} \right)^x \left(1 + \frac{2\left(1-\sqrt{1-4s^2}\right)}{12s^2 - 1\left(1-\sqrt{1-4s^2}\right)^2} \right). \end{aligned} \tag{15.42}$$

Since, by definition, $Z_N(x) = A^N B^x W_N(x)$ (see Eq. (15.37)), we can write down the relation between the generating functions of $Z_N(x)$ and of $W_N(x)$:

$$\mathcal{Z}(\lambda, x) = \sum_{N=0}^{\infty} Z_N(x) \lambda^N = \sum_{N=0}^{\infty} A^N B^x W_N(x) \lambda^N = B^x \mathcal{W}(\lambda A, x). \tag{15.43}$$

Thus,

$$\mathcal{Z}(\lambda, x) = 3^{x/2} \mathcal{W}(\lambda\sqrt{3}, x), \tag{15.44}$$

where $\mathcal{W}(\lambda\sqrt{3}, x)$ is given by Eq. (15.42) and we should substitute $\lambda\sqrt{3}$ for s. The partition function, $\mathcal{Z}(\lambda, x)$, of the random walk ensemble reads

$$\mathcal{Z}(\lambda, x) = \frac{6\lambda\left(\dfrac{1-\sqrt{1-4\lambda^2(p-1)}}{2\lambda}\right)^x}{18\lambda^2 - \left(1-\sqrt{1-12\lambda^2}\right)}. \tag{15.45}$$

For the grand partition function of all trajectories returning to the origin, $\mathcal{Z}(\lambda) \equiv \mathcal{Z}(\lambda, x = 0)$, we get the following expression

$$\mathcal{Z}(\lambda) = \frac{6\lambda}{18\lambda^2 - \left(1-\sqrt{1-12\lambda^2}\right)}. \tag{15.46}$$

To extract the asymptotic behavior of the partition function Z_N (the number of trajectories returning to the origin after N steps, one should perform the inverse transform similar to Eq. (15.39)

$$Z_N = \frac{1}{2\pi i}\oint \mathcal{Z}(\lambda)\lambda^{-N-1}d\lambda \sim \frac{\left(2\sqrt{3}\right)^N}{N^{3/2}}. \tag{15.47}$$

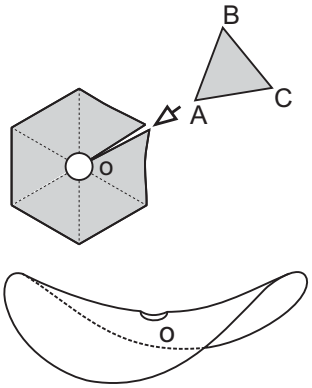
As it should be, the probability to return to the origin on a 4-branching Cayley tree, $Z_N/(4^N)$ is exponentially small.

Let us note some striking topological similarity between the Cayley tree structure of the noncommutative group Γ_2 shown in Fig. 15.9 and the metric structure of the modular group, visualized in Fig. 15.7. This similarity is not occasional. Having the graph of the group Γ_2, we can ask the question in which Riemann surface the graph of the group Γ_2 can be isometrically embedded. The answer is that the Cayley tree is the graph of isometries (one of many) of the Lobachevsky plane (the Riemann surface of constant negative curvature). This is schematically depicted in Fig. 15.11 where the chip (the saddle) is the example of the surface with constant negative Gaussian curvature. Contrary to that, the commutative group $\{f_1, f_2, f_1^{-1}, f_2^{-1}\}/[f_1 f_2 = f_2 f_1]$ isometrically covers the planar square lattice – see Fig. 15.10b.

Conditional Brownian bridges in Hyperbolic spaces

The result formulated in this Section is the central point, connecting statistics of random walks in Hyperbolic spaces and the topology of knotted random walks. There are a few different incarnations of one and the same question concerning the conditional return probability of the symmetric random walk in the Hyperbolic geometry:

(i) For the problem of the conditional paths counting on the Cayley tree, we are interested in the following question. Let $Z_N(x)$ be the number of N-step path on the Cayley tree starting from the origin and ending at some distance x measured in number of the tree generations ("coordinational spheres") from the root point. Let the paths starting from the tree root, reach the distance



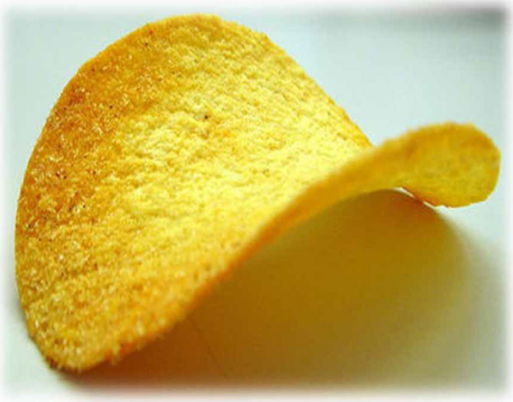


Figure 15.11: Isometric embedding of the free group into the Lobachevsky plane (Riemann surface of constant negative curvature).

x after M steps, and then return to the origin at the very last step, N. The corresponding conditional distribution, $P(x, M, N)$ we can compute as follows

$$P(x, M, N) = \frac{Z_M(x) Z_{N-M}(x)}{Z_M(0) V(x)}. \tag{15.48}$$

The expression Eq. (15.48) means that the entire N-step Brownian bridge (the path returning to the origin) consists of two independent parts: the M-step part of the path form the root point to *some* point located at the distance x, and $N - M$-step part of the path from the root point again to *some* point located at the distance x. Now we have to ensure that the ends of these two M- and $N - M$-parts *coincide* at the point $\mathbf{x}$. The factor $V(x)$ in the denominator of Eq. (15.48) is the number of different points located at the distance x from the root of the tree, $V(x) = 4 \times 3^{x-1}$ for the 4-branching Cayley tree. Thus, $Z_M(x) V^{-1}(x)$ is the probability that the M-step part ends in some *specific* point $\mathbf{x}$. Substituting expressions for $Z_M(x)$, $Z_{N-M}(x)$, $Z_N(0)$ in Eq. (15.48), we arrive at the following asymptotic form of $P(x, M, N)$ at $x \gg 1$:

$$P(x, M, N) \sim \sqrt{\frac{N}{2\pi M(N-M)}} e^{x^2 \left(\frac{1}{2M} + \frac{1}{2(N-M)} \right)}. \tag{15.49}$$

Computing $\langle x^2 \rangle$ with the function $P(x, M, N)$, we get

$$\bar{x} = \sqrt{\langle x^2 \rangle} = \frac{\sum_{x=0}^{\infty} x^2 P(x, M, N)}{\sum_{x=0}^{\infty}} = \sqrt{\frac{M(N-M)}{N}}. \tag{15.50}$$

As one sees, for any $M = cN$ ($c = \text{const}$, $0 < c < 1$), the average distance $\bar{x}$ has the behavior typical for an ordinary random walk,

$$\bar{x} = \sqrt{a(1-a)} \sqrt{N}. \tag{15.51}$$

Thus, the typical behavior of intermediate points of the Brownian bridge on the Cayley tree is statistically the same as on the one-dimensional lattice, i.e., the drift on the tree which occurs due to asymmetry of the random walk (the probability to go *from* the root is larger than the probability to come *back* to the root) is completely compensated. This is the key point for existence of topologically nontrivial crumpled globule structure, which will be discussed in the following Section.

(ii) Consider now the asymptotics of the conditional Brownian bridge in the Lobachevsky plane. Construct the desired conditional probability, $W(x, M, N)$ as follows

$$W(\rho, \tau, t) = \frac{P(\rho, \tau)P(\rho, t-\tau)}{P(0,t)} v(\rho), \tag{15.52}$$

where $P(\rho, t)$ defines the probability that the random path after time t ends in the *specific* point located at the distance ρ of the Lobachevsky plane, and $v(\rho) = \sinh \rho$ is the circumference of circle of radius ρ in the Lobachevsky plane.

The diffusion equation for the density $P(\mathbf{q}, t)$ of the free random walk on a Riemann manifold is governed by the Beltrami-Laplace operator:

$$\partial_t P(\mathbf{q}, t) = \mathcal{D} \frac{1}{\sqrt{g}} \frac{\partial}{\partial q_i} \left(\sqrt{g} \left(g^{-1} \right)_{ik} \frac{\partial}{\partial q_k} \right) P(\mathbf{q}, t), \tag{15.53}$$

where

$$\begin{aligned} P(\mathbf{q}, t=0) &= \delta(\mathbf{q}). \\ \int \sqrt{g} P(\mathbf{q}, t) d\mathbf{q} &= 1, \end{aligned} \tag{15.54}$$

and g_{ik} is the metric tensor of the manifold; $g = \det g_{ik}$. For the Lobachevsky plane one has

$$||g_{ik}|| = \left|\left| \begin{array}{cc} 1 & 0 \\ 0 & \sinh^2 \rho \end{array} \right|\right|, \tag{15.55}$$

where ρ stands for the geodesics length in the Lobachevsky plane. The corresponding diffusion equation now reads

$$\partial_t P(\rho, \varphi, t) = \mathcal{D} \left(\partial^2_{\rho\rho} + \coth \rho \partial_\rho + \frac{1}{\sinh^2 \rho} \partial^2 \varphi\varphi \right) P_p(\rho, \varphi, t). \tag{15.56}$$

The radially symmetric solution of Eq.(15.56) is

$$\begin{aligned} P(\rho, t) &= \frac{e^{-\frac{t\mathcal{D}}{4}}}{4\pi\sqrt{2\pi(t\mathcal{D})^3}} \int_\rho^\infty \frac{\xi \exp\left(-\frac{\xi^2}{4t\mathcal{D}}\right)}{\sqrt{\cosh \xi - \cosh \rho}} d\xi \\ &\simeq \frac{e^{-\frac{t\mathcal{D}}{4}}}{4\pi t\mathcal{D}} \left(\frac{\rho}{\sinh \rho} \right)^{1/2} \exp\left(-\frac{\rho^2}{4t\mathcal{D}} \right), \end{aligned} \tag{15.57}$$

(compare to Eq. (15.42)). Substituting Eq. (15.57) into Eq. (15.52), we get for the conditional probability $W(\rho,\tau,t)$ the following asymptotic expression

$$W(\rho,\tau,t)=\frac{N}{4\pi\mathcal{D}\tau(t-\tau)}\rho\,\exp\left\{-\frac{\rho^2}{4\mathcal{D}}\left(\frac{1}{\tau}+\frac{1}{t-\tau}\right)\right\}. \tag{15.58}$$

Hence we again reproduce the Gaussian distribution function with zero mean.

(iii) Eqs. (15.49) and (15.58) describing the conditional distributions of Brownian bridges on the Cayley tree and on the Riemann surface of constant negative curvature, have direct application to the conditional distributions of Lyapunov exponents for products of non-commutative matrices. Consider for specificity the random walk on the group $SL(2,R)$. Namely, we multiply sequentially N random matrices $M_j=\begin{pmatrix} a_j & b_j \\ c_j & d_j \end{pmatrix}\in SL(2,R)$, whose entries $\{a_i,b_j,c_j,d_j\}$ are randomly distributed for any $j=1,...,N$ in some finite support subject to the relation $a_jd_j-b_jc_j=1$. Thus we have a product of random matrices

$$Q(N)=\begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}\begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix}...\begin{pmatrix} a_M & b_M \\ c_M & d_M \end{pmatrix}...\begin{pmatrix} a_N & b_N \\ c_N & d_N \end{pmatrix}. \tag{15.59}$$

The asymptotic $N\to\infty$ behavior of the largest eigenvalue, Λ_N of the typical (i.e., averaged over different samples) value of $Q(N)$ is ensured by the Fürstenberg theorem [46], which states that

$$\Lambda_N\sim e^{\delta_1 N} \tag{15.60}$$

where δ_1 is the group– and measure–specific, though N-independent, "Lyapunov exponent".

Motivated by examples (i) and (ii), consider the conditional Brownian bridge in the space of matrices, i.e., consider such products $Q(N)$ that are equal to the unit matrix and ask about the typical behavior of the largest eigenvalue of first M as shown below:

$$\overbrace{\begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}\begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix}...\begin{pmatrix} a_M & b_M \\ c_M & d_M \end{pmatrix}}^{\Lambda_M=?}...\begin{pmatrix} a_N & b_N \\ c_N & d_N \end{pmatrix}=\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{15.61}$$

The answer to this question is as follows: $\Lambda_M=e^{\delta_2\sqrt{\frac{M(N-M)}{N}}}$ i.e., for $M=cN$ $(0<c<1)$, one has

$$\Lambda_M=e^{\delta_3\sqrt{N}}, \tag{15.62}$$

where we have absorbed the constants in $\delta_3=\delta_2\sqrt{a(1-a)}$. The proof of the corresponding theorem using the method of large deviations, can be found in [35], however the result Eq. (15.62), which holds for random walks on any noncommutative group, is easy to understand qualitatively. It is sufficient to recall that the Lobachevsky plane H can be identified with the group

$SL(2,R)/SO(2)$. Thus, the Brownian bridge on the group $SL(2,R)/SO(2)$, can be viewed either as the conditional random walk governed by the Beltrami-Laplace operator, or as the product of random matrices.

We arrive at the following conclusion. The Brownian bridge condition for random walks in the space of constant negative curvature makes the curved space effectively flat and turns the corresponding conditional distribution for the intermediate time moment of random walks to the Gaussian distribution with zero mean. This result is very general and can be applied to random walks on various noncommutative groups, such as modular group, $SL(n,R)$, braid groups B_n, etc. This result is crucial in our further discussions of the crumpled state of collapsed unknotted polymer chain.

15.2.3 Crumpled globule: Topological correlations in collapsed unknotted rings

In 1988 we have theoretically predicted the new condensed state of a ring unentangled and unknotted macromolecule in a poor solvent. We named this state "the crumpled globule" and studied its unusual fractal properties [4]. That time our arguments were rather hand-waving and more solid understanding came essentially later, around 2005 [31,36]. As we shall show, the most striking physical arguments of the estimation of the degree of entanglement of a part of a long unknotted random trajectory confined in a small box are provided by the statistics of conditional Brownian bridges in the space of constant negative curvature.

First of all one has to define the topological state of a part of a ring polymer chain. As we discussed in the Introduction, mathematically rigorous definition of the topological state exists for closed or infinite paths only. However, everyday's experience tells us that open but sufficiently long rope can be knotted. Hence, it is desirable to introduce a concept of a *quasiknot* available for topological description of open paths. For the first time the idea of quasiknots in a polymer context had been formulated by I.M. Lifshits and A.Yu. Grosberg [47]. They argued that the topological state of a linear polymer chain in a collapsed (globular) state is defined much better than topological state of a random coil. Actually, the distance between the ends of the chain in a globule is of order $R \sim aN^{1/3}$, where a is a size of a monomer and N is a number of monomers in a chain. Taking into account that R is sufficiently smaller than the contour length N and that the density fluctuations in the globular state are negligible, we may define the topological state of a path in a globule as a topological state of composite trajectory consisting of a chain itself and a segment connecting its ends. This composite structure can be regarded as a quasiknot for an open chain in a collapsed state. Later we shall repeatedly use this definition.

The influence of topological constraints on statistical properties of polymers, namely, the random knotting probability, in confined geometries has been numerically considered in Ref. [48], while the paper Ref. [49] has been devoted

to the determination of the equilibrium entanglement complexity of polymer chains in melts. In the works Ref. [36] we take a further step, considering the topological state of a part of ring unknotted polymer chain in a confined geometry, where the compact configuration of the polymer is modelled by dense lattice knots. The knot is called "dense" if its projection onto the plane completely fills the rectangular lattice M of size $L_v \times L_h$ as shown in Fig. 15.12a, resembling the celtic knot in Fig. 15.1. The lattice M is filled densely by a single thread, which crosses itself in each vertex of the lattice in two different ways: "up" or "down". The topology of a lattice diagram is defined by the up-down passages, and by the prescribed boundary conditions. The "woven carpet" shown in Fig. 15.12a corresponds to a trivial knot. To avoid any possible confusion, we apply our model to the polymer ring located in a thin slit between two horizontal plates as shown in Fig. 15.12b. It is evident that the ring chain in a thin slit becomes a quasi two-dimensional system. Our lattice model is oversimplified (even for the polymer chain in a thin slit) because it does not take into account the spatial fluctuations of a knotted polymer chain. However, we expect that our model properly describes the condensed (globular) structure of a polymer ring because the chain fluctuations in the globule are essentially suppressed and the chain has reliable thermodynamic structure with a constant density [47].

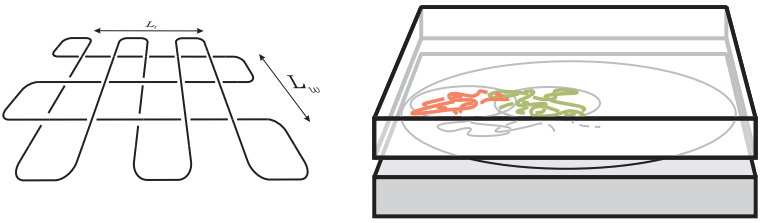


Figure 15.12: (a) Random woven carpet corresponding to the trivial knot; (b) Dense knot confined in a thin slit.

We are interested in the following statistical-topological question inspired by the conditional Brownian bridge ideology. Define at each intersection of vertical and horizontal threads (i.e., in each "lattice vertex") k the random variable ε_k, taking values

$$\varepsilon_k = \begin{cases} +1 & \text{for "up" crossing,} \\ -1 & \text{for "down" crossing.} \end{cases} \tag{15.63}$$

The set of independently generated quenched random variables $\{\varepsilon_1, ..., \varepsilon_N\}$ in all vertices of the lattice diagram, together with the boundary conditions, define the knot topology.

Suppose that we consider such a sub-ensemble of crossings $\{\varepsilon_1, ..., \varepsilon_N\}$ that corresponds to the trivial entire ("parent") knot. Let us cut a part of a parent trivial knot and close open ends of the threads as it is shown in Fig. 15.13.

This way we get the well defined "daughter" quasiknot. We are interested in the typical topological state of daughter quasiknots under the condition that the parent knot is trivial. I think, the reader can feel in this formulation the flavor of conditional Brownian bridges...

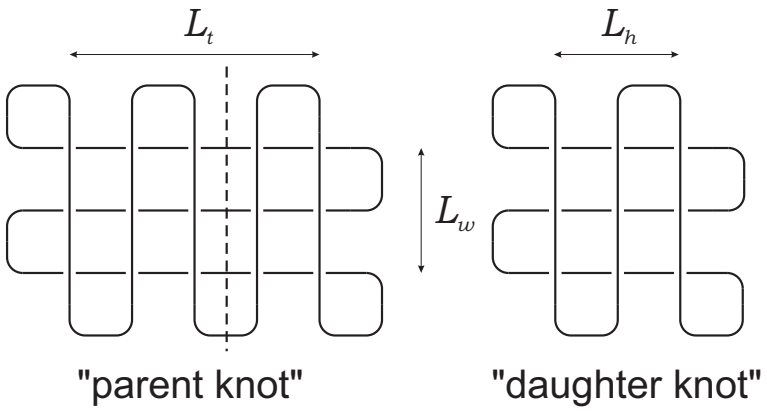


Figure 15.13: (a) Trivial "parent" knot; (b) "Daughter" knot obtained by cutting a part of the parent knot.

The averaged knot complexity, n, understood as the minimal number of crossings on the knot diagram, for the unconditional random knotting of knot diagram behaves as $n \sim N$ where N is the initial size of the lattice knot. By the semi-analytic and semi-numeric arguments we have shown in [36] that the typical conditional complexity, n^*, of the daughter knot of size M behaves as $n^* \sim \sqrt{\frac{M(N-M)}{N}}$ and for $M = cN$ $(0 < c < 1)$, $N \gg 1$, has the asymptotic behavior

$$n^* \sim \sqrt{a(1-a)}\sqrt{N}. \tag{15.64}$$

Thus, each macroscopic part of a dense lattice trivial knot is weakly knotted (compared to the unconditional random knotting).

In 2015 we have performed in Ref. [50] extensive Monte-Carlo simulations for self-avoiding polymer chains in confined geometry in 3D space. The role of topological constraints in the *equilibrium* state of a single compact and unknotted polymer remains unknown. Previous studies [4, 5] have put forward a concept of the *crumpled globule* as the equilibrium state of a compact and unknotted polymer. In the crumpled globule, the subchains were suggested to be space-filling and unknotted. Recent computational studies examined the role of topological constraints in the non-equilibrium (or quasi-equilibrium) polymer states that emerge upon polymer collapse [6, 51–54]. This non-equilibrium state, often referred to as the *fractal globule* [6, 55], can indeed possess some properties of the conjectured equilibrium crumpled globule. The properties of the fractal globule, its stability [56], and its connection to the equilibrium state are yet to be understood.

Elucidating the role of topological constraints in equilibrium and non-equilibrium polymer systems is important for understanding the organization of chromosomes. Long before experimental data on chromosome organization became available [6], the crumpled globule was suggested as a state of long DNA molecules inside a cell [5]. Recent progress in microscopy [57] and genomics [58] provided new data on chromosome organization that appear to share several features with topologically constrained polymer systems [6,59,60]. For example, segregation of chromosomes into territories resembles segregation of space-filling rings [61–63], while features of intra-chromosomal organization revealed by Hi-C technique are consistent with a non-equilibrium fractal globule emerging upon polymer collapse [6,55,64] or upon polymer decondensation [59]. These findings suggest that topological constraints can play important roles in the formation of chromosomal architecture [65].

In Ref. [50] we have examined the role of topological constraints in the equilibrium state of a compact polymer. We performed the equilibrium Monte Carlo simulations of a confined unentangled polymer ring with and without topological constraints. Without topological constraints, a polymer forms a classical equilibrium globule with a high degree of knotting [2,66,67]. A polymer is kept in the globular state by impenetrable boundaries, rather than pairwise energy interactions, allowing fast equilibration at a high volume density.

In the work of Ref. [50] we found that topological states of closed subchains (loops) are drastically different in the two types of globules and reflect the topological state of the whole polymer. Namely, loops of the unknotted globule are only weakly knotted and mostly unconcatenated. We also found that spatial characteristics of small knotted and unknotted globules are very similar, with differences starting to appear only for sufficiently large globules. Subchains of these large unknotted globules become asymptotically compact ($R_G(s) \sim s^{1/3}$), forming crumples. Analysis of the fractal dimension of surfaces of loops suggests that crumples form excessive contacts and slightly interpenetrate each other. Overall, in the the asymptotic limit (for very long chains) we have supported the conjectured crumpled globule concept [4]. However, the results of Ref. [50] also demonstrate that the internal organization of the unknotted globule at equilibrium differs from an idealized hierarchy of self-similar isolated compact crumples.

In the simulations a single homopolymer ring with excluded volume interactions was modelled on a cubic lattice and confined into a cubic container at a volume density 0.5. The Monte Carlo method with non-local moves (see [50] and references therein) allowed us to study chains up to $N = 256\,000$. If monomers were prohibited to occupy the same site, this Monte Carlo move set naturally constrains topology, and the polymer remains unknotted. The topological state of a loop was characterized by $\varkappa$, the logarithm of the Alexander polynomial evaluated at -1.1 [61, 66, 67]. To ensure equilibration, we estimated the scaling of the equilibration time with N for $N \leq 32\,000$, extrapolated it to large N, and ran simulations of longer chains, $N = 108\,000$ and $256\,000$, to exceed the estimated equilibration time. We also made sure that chains with topolog-

ical constraints remain completely unknotted through the simulations, while polymers with relaxed topological constraints become highly entangled.

To understand the role of topological correlations, we asked how the topological state of the whole polymer influences the topological properties of its subchains. Because a topological state can be rigorously defined only for a closed contour, we focused our analysis on loops, i.e., subchains with two ends occupying neighboring lattice sites. Fig. 15.14a presents the average knot complexity $\langle \varkappa(s) \rangle$ for loops of length s for both types of globules. We found that loops of the knotted globule were highly knotted, with the knot complexity rising sharply with s. Loops of the unknotted globule, on the contrary, were weakly knotted, and their complexity increased slowly with s.

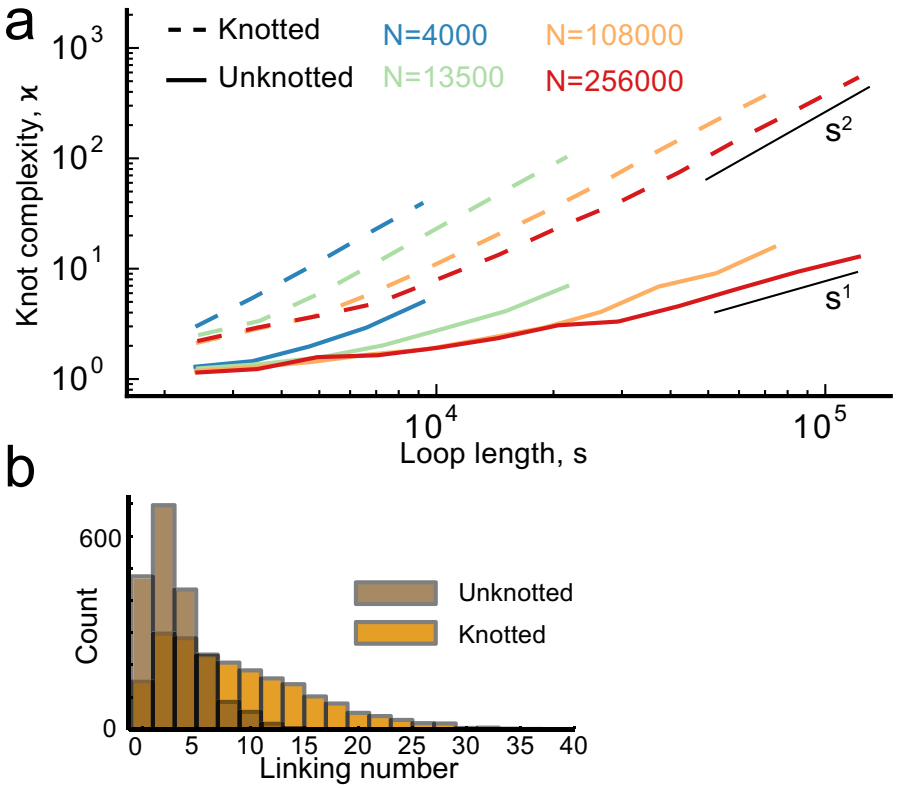


Figure 15.14: Topological properties of polymer loops in the knotted and unknotted globules. (a) Knot complexity of polymer loops as a function of their length, s, for chains of different length N (shown by colors) in knotted (dashed) and unknotted (solid) globules. (b) Distribution of the linking numbers for non-overlapping loops of length $s = 9000$ to 11 000 in 32 000-long globules.

This striking difference in the topological states of loops for globally knotted and unknotted chains is a manifestation of the general statistical behavior

of so-called matrix–valued Brownian Bridges (BB) [36]. The knot complexity $\varkappa$ of loops in the topologically unconstrained globule is expected to grow as $\varkappa(s) \sim s^2$. In contrast, due to the global topological constraint imposed on the unknotted globule, the knot complexity of its loops grows slower, $\varkappa(s) \sim s$, which follows from the statistical behavior of BB in spaces of constant negative curvature (see Fig. 15.15a, and Refs. [31, 36] for details).

Another topological property of loops of a globule is the degree of concatenation between the loops. We computed the linking number for pairs of non-overlapping loops in the knotted and crumpled (unknotted) globules, and found that loops in the unknotted globule are much less concatenated than loops in the knotted globule. Taken together, these results show that the topological state of the whole (parent) chain propagates to the daughter loops. While loops of the unknotted globule are still slightly linked and knotted, their degree of entanglement is much lower than for the loops in the topologically relaxed knotted globule.

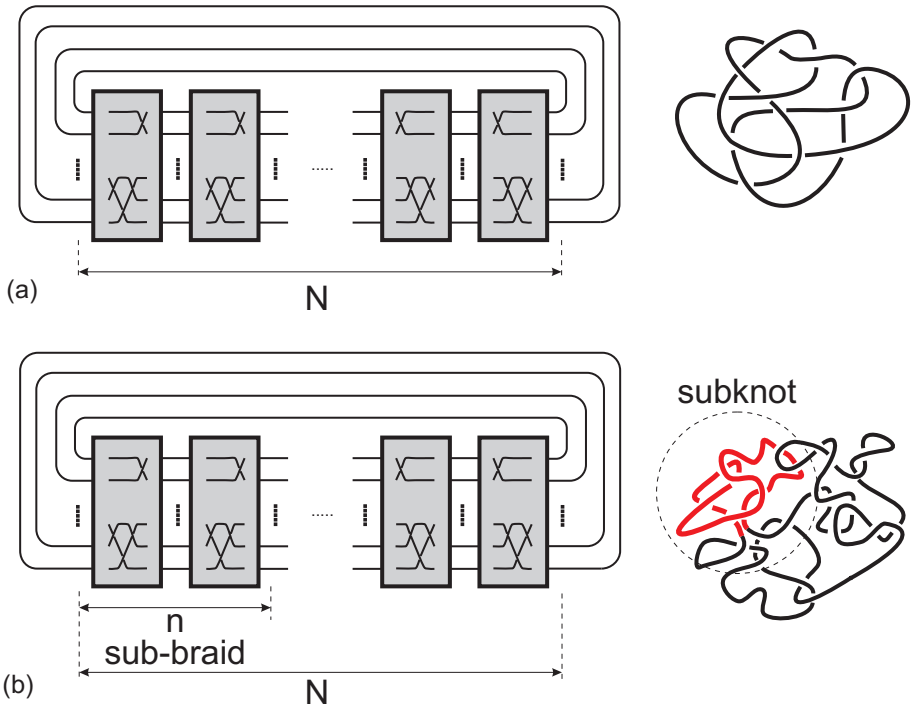


Figure 15.15: Schematic representation of knots by braids: a) unconditional random distribution of black boxes produces a very complex knot; b) conditional distribution implies the whole knot to be trivial, which imposed strong constraints on complexity of any subpart of the braid.

Our topological problem to determine the complexity of a subloop in a globally trivial collapsed polymer chain allows natural interpretation in terms

of Brownian bridges. Suppose the following imaginative experiment. Consider the phase space Ω of all topological states of densely packed knots on the lattice. Select from Ω the subset ω of trivial knots. To simplify the setting, consider a knot represented by a braid, as shown in Fig. 15.15, where the braid is depicted by a sequence of uncorrelated "black boxes" (each black box contains some number of up- and down-crossings. If crossings in all black boxes are identically and uniformly distributed, then the boxes are statistically similar. Cut a part of each braid in the subset ω, close open tails and investigate the topological properties of resulting knots. Just such a situation has been qualitatively studied in Ref. [31, 36]. The crumpled globule hypothesis states the following: if the whole densely packed lattice knot is trivial, then the topological state of each of its "daughter" knot is almost trivial. We have shown that the computation of the knot complexity in the braid representation depicted in Fig. 15.15 can be interpreted as the computation of the highest eigenvalue of the product of noncommutative matrices designated by the black boxes.

To proceed, consider first the typical (unconditional) complexity of a knot represented by a sequence of N independent black boxes. This question is similar to the growth of the logarithm of the largest eigenvalue, Λ, of the product of N independent identically distributed noncommutative random matrices. According to the Fürstenberg theorem [46], in the limit $N \gg 1$ one has

$$\ln \Lambda(N) \sim \gamma_1 N, \tag{15.65}$$

where $\gamma_1 = \text{const}$ is the so-called Lyapunov exponent (compare with Eq. (15.60)). Being rephrased for knots, this result means that the average knot complexity, $\varkappa$, understood as a minimal number of crossings, M, necessary to represent a given knot by the compact knot diagram, extensively grows with M, i.e., $\varkappa \sim M$. In the ordinary globule, for subchains of length $N^{2/3} < s < N$, the typical number of crossing, M, on the knot diagram grows as $M \sim s^2$, leading to the scaling behavior for the knot complexity $\varkappa$:

$$\varkappa \sim s^2. \tag{15.66}$$

This is perfectly consistent with the well known fact: the probability of spontaneous unknotting of a polymer with open ends in a globular phase is exponentially small. Following the standard scheme [12, 66, 67], we characterize the knot complexity, $\varkappa$, by the logarithm of the Alexander polynomial, $\ln[\mathrm{Al}(t = -1.1)\mathrm{Al}(t = -1/1.1)]$, i.e., we set $\varkappa = \ln[\mathrm{Al}(t = -1.1)\mathrm{Al}(t = -1/1.1)]$. As seen from Fig. 15.14, the conjectured dependence $\ln \mathrm{Al}(t = -1.1) \sim s^2$ is perfectly satisfied for ordinary (knotted) globule.

Consider now the *conditional* distribution on the products of identically distributed black boxes. We demand the product of matrices represented by black boxes to be a unit matrix (topologically trivial). The question of interest concerns the typical behavior of $\ln \Lambda^*(M, N)$, where $\Lambda^*(M, N)$ is the largest eigenvalue of the sub-chain of first M matrices in the chain of N ones. The answer to this question is known [35]: if $n = cN$ ($0 < c < 1$ and $N \gg 1$), then

$$\ln \lambda^*(n = cN, N) \sim \sqrt{c(1-c)}\sqrt{N} = \gamma_2(c)\sqrt{N}, \tag{15.67}$$

where $\gamma_2(c)$ absorbs all constants independent on N. Translated to the knot language, the condition for a product of N matrices to be completely reducible, means that the "parent" knot is trivial. Under this condition we are interested in the typical complexity $\varkappa^*$ of any "daughter" sub-knot represented by first $n = cN$ black boxes.

Applying the Eq. (15.67) to the knot diagram of the unknotted globule, we conclude that the typical conditional complexity, $\varkappa^*$ expressed in the minimal number of crossings of any finite sub-chain of a trivial parent knot, grows as

$$\varkappa^* \sim \sqrt{s^2} \sim s, \tag{15.68}$$

with the subchain size, s. Comparing Eqs. (15.68) and (15.66), we conclude that subchains of length s in the trivial knot are much less entangled/knotted than subchains of same lengths in the unconditional structure, i.e., when the constraint for a parent knot to be trivial is relaxed. Indeed, this result is perfectly supported by Fig. 15.14 which show linear growth of $\tilde{\varkappa} = \ln[\mathrm{Al}(t = -1.1)\,\mathrm{Al}(t = -1/1.1)]$ with s for the unknotted globule, while quadratic growth for the knotted globule.

15.3 Conclusion

15.3.1 The King is dead, long live The King!

The very concept of the crumpled (fractal) globule as of possible thermodynamic equilibrium state of an unknotted ring polymer confined in a small volume, appeared in 1988 in a joint work by A. Grosberg, S. Nechaev and E. Shakhnovich [4]. Soon after, in 1993, A. Grosberg, Y. Rabin, S. Havlin, and A. Neer published a paper where they proposed the crumpled globule model to be a possible condensed state of DNA packing in a chromosome [5]. Then, over decades, the interest to the crumpled globule was moderate: it was considered as an interesting, though sophisticated artificial exercise. The attempts to find the crumpled structure in direct numeric simulations, or in real experiments on proteins or DNAs were not too convincing.

Still, a few interesting exceptions, which fuelled some discussions around the crumpled globule, should be mentioned: (i) the observation of the two-stage dynamics of collapse of the macromolecule after abrupt changing of the solvent quality, found in light scattering experiments by B. Chu and Q. Ying [53]; (ii) the experiments on compatibility enhancement in mixtures of ring and linear chains [68], the construction of the quantitative theory of a collapse of N–isopropilacrylamide gel in a poor water [69]; (iii) the experiments on superelasticity of polymer gels prepared in diluted solutions [70]; (iv) the indications of observation of the crumpled globule in numerical simulations [71, 72]. The breakthrough in the interest to the crumpled globule happen after the brilliant experimental work of the MIT-Harvard team in 2009 [6]. Immediately after, the concept of crumpled (fractal) globule became the candidate for the new

paradigm explaining many realistic features the DNA packing and functioning in a human genome.

Analysis of chromatin folding in human genome based on a genome-wide chromosome conformation capture method (Hi-C) [6, 73] provides a comprehensive information on spatial contacts between genomic parts and imposes essential restrictions on available 3D genome structures. The experimental Hi-C maps obtained for various organisms and tissues [6, 58, 74–78] display very rich structure in a broad interval of scales. The researchers pay attention to the average contact probability, $\mathcal{P}(s)$, between two units of genome separated by a genomic distance, s, which decays in typical Hi-C maps approximately as $\mathcal{P}(s) \sim 1/s$ (see [6]).

The crumpled globule is a state of a polymer chain which in a wide range of scales is self-similar and almost unknotted, forming a fractal space-filling-like structure. Both these properties, self-similarity and absence of knots, are essential for genome folding: fractal organization makes genome tightly packed in a broad range of scales, while the lack of knots ensures easy and independent opening and closing of genomic domains, necessary for transcription [5, 55]. In a three-dimensional space such a tight packing results in a *space-filling* with the fractal dimension $D_f = D = 3$. The Hi-C contact probability, $P_{i,j}$, between two genomic units, i and j in a N-unit chain, depends on a combination of structural and energetic factors. Simple mean-field arguments (see, for example, [6]) demonstrate that in a fractal globule with $D_f = 3$ the *average* contact probability, $\mathcal{P}(s) = (N-s)^{-1} \sum_{i=0}^{N-s} P_{i,i+s}$, between two units separated by the genomic distance $s = |i-j|$, decays as $\mathcal{P}(s) \sim s^{-1}$. It should be noted that recent numeric simulations [50, 61], and more sophisticated arguments beyond the mean-field approximation [60, 79], point out that the contact probability decays as $\mathcal{P}(s) \sim s^{-\gamma}$ with $\gamma \simeq 1.05 - 1.09$.

Despite the crumpled globule being our "favourite child", I should clearly state that it does not explain exhaustively all details of the chromatin folding and definitely should be combined with more refined models and concepts. Theoretical models of chromatin packing in the nucleus, which can possibly explain the observed behavior of intra-chromosome Hi-C contact maps, split roughly into two groups. The first group of works relies on specific interactions within the chromatin, like loop or bridge formation, [80–86] and these authors do not believe in crumpled globule, while the second group aims to explain the chromatin structure in terms of large-scale topological interactions [5, 6, 50, 55, 59, 60, 79, 87, 88] based on the crumpled model of the polymer globule [4]. For example, in [89] we combined the assumption that chromatin can be considered as a heteropolymer chain with a quenched primary sequence [90], with the general hierarchical fractal globule folding mechanism. With this conjecture we were able to reproduce the large-scale chromosome compartmentalization, not assumed explicitly from the very beginning. To show the compatibility of the hierarchical folding of a crumpled globule with the fine structure of experimentally observed Hi-C maps, we suggested in Ref. [89] a

simple toy model based on the crumpled globule folding principles together with an account of quenched disorder in primary sequence.

To summarize, my feeling of the current state of the art called "the crumpled globule" is formulated in the title of this Section. All together we have come a long way from the nonperturbative description of topological constraints in collapsed polymer phase to real biophysical applications, we have understood on this way constructive connection of statistics of polymer entanglements with Brownian bridges in the non-Euclidean geometry, we have got new results for statistics of random braids, and finally we have explained some features of DNA packing in chromosomes. Keeping our eyes open, we clearly see that the appearance of new experimental results demonstrates that our initial topological arguments were too crude and too naive. However they gave birth to a new understanding of the role of topology in genomics and have led to new ideas and methods which constitute the modern STATISTICAL TOPOLOGY OF POLYMERS.

15.3.2 Where to go

I think we are only at the beginning of a highway, where the statistics of random walks is intertwined with the geometric group theory, algebraic topology and integrable systems in mathematical physics, as well as has various incarnations in physics dealing with the crumpled globule concept. Let me name some but a few such directions.

I would like to mention random walks on braid groups, growth of random heaps, viewed as random sequential ballistic deposition (random "Tetris game"). Introducing the concept of the "locally free group" as an approximant of the braid group, one can solve exactly the word problem in the locally free group and obtain analytically the bilateral estimates (from above and from below) for the growth of the volume of the braid group B_n for arbitrary n [91–93]. Interesting new results beyond the mean-field approximation for entanglement of threads in random braids have been obtained recently in Ref. [94].

Sequential ballistic deposition (BD) with the next-nearest-neighbor interactions in a N-column box can viewed as a time-ordered product of $N \times N$-matrices consisting of a single sl_2-block which has a random position along the diagonal. One can interpret the uniform BD growth as the diffusion in the symmetric space H_N. In particular, the distribution of the maximal height of a growing heap can be connected to the distribution of the maximal distance for the diffusion process in H_N, where the coordinates of H_N can be interpreted as the coordinates of particles of the one-dimensional Toda chain. The group-theoretic structure of the system and links to some random matrix models was discussed in Ref. [95].

As concerns the impact of crumpled globule concept in physics, we have demonstrated that folding and unfolding of a crumpled polymer globule can be viewed as a cascade of equilibrium phase transitions in a hierarchical system, similar to the Dyson hierarchical spin model. Studying the relaxation properties

of the elastic network of contacts in a crumpled globule, we showed that the dynamic properties of hierarchically folded polymer chains in globular phase are similar to those of natural molecular machines (like myosin, for example). We discuss the potential ways of implementations of such artificial molecular machine in computer and real experiments, paying attention to the conditions necessary for stabilization of crumples under the fractal globule formation in the polymer chain collapse [64, 96].

The ability of crumpled globule to act as molecular machine definitely provokes a new angle on an eternal problem of the origin of life, related to overcoming the "error threshold" in producing and selecting complex molecular structures during the prebiotic evolution [97]. This permits us to put forward a conjecture about a possibility for the crumpled globule to be a sort of the "primary molecular machine" naturally formed under the prebiotic conditions. The primary crumpled globule molecular machine (CGMM) made by polymers (not necessarily of biological nature), could perform some specific functions typical for true biological molecular machines. The diversity of CGMM may be concerned mainly with the attracting manifold, in which the CGMM action is performed. This allows for functional variability without altering the structural archetype. Certainly, the idea that crumpled globule could be the prebiotic molecular machine needs experimental verification. However, the results [64, 98] provide a rather optimistic view on the evolutionary scenarios in which the primary molecular machines, themselves, are taken out of the biomolecular context. In this paradigm, the beginning of biological evolution is associated with the spontaneous appearance of complex functional systems of primary "artificial" CGMM capable of performing collective reproduction and autonomous behavior, which then are replaced in evolution by more effective biomolecular systems. On this optimistic note I would like to end the story.

These notes are based on several lectures at the SERC School on Topology and Condensed Matter Physics (organized in 2015 by RKM Vivekananda University at S.N Bose National Center for Basic Sciences, Calcutta, India). I would thank Somen Bhattacharjee for kind invitation, opportunity to explore some wonderful places in India and strong push to arrange lectures as a written text. The topics discussed above summarize the subjects of millions conversations over many years with my friends and colleagues, Alexander Grosberg, Anatoly Vershik, Vladik Avetisov, Leonid Mirny, Michael Tamm. Particularly I would like to thank Maxim Frank-Kamenetskii and Alexander Vologodsky, whose nice review in 1981 fuelled my interest to polymer topology, and to Alexey Khokhlov with whom we got first results beyond the Abelian theory of polymer entanglements in 1985. The importance of conditional Brownian bridge concept in statistics of non-commutative random walks was recognized in joint work with Yakov Sinai in 1991. Especially I would like to highlight the role of Alexander Grosberg whose ironic and deep comments and ideas tease and support me for more than a quarter of century.

References

[1] L.H. Kauffman, Topology, **26** 395 (1987); L.H. Kauffman, AMS Contemp. Math. Series, **78** 263 (1989); L.H. Kauffman, *Knots and Physics* (Singapore: WSPC, 1991)

[2] A.Yu. Grosberg, A.R. Khokhlov, *Statistical physics of macromolecules* (New York: AIP Press, 1994)

[3] A.Yu. Grosberg, S. Nechaev, J. Phys. A: Math. Gen., **25** 4659 (1992)

[4] A.Yu. Grosberg, S.K. Nechaev, E.I. Shakhnovich, J.Phys. (Paris), **49** 2095 (1988)

[5] A.Yu. Grosberg, A. Feigel, Y. Rabin, Phys. Rev. E, **54**, 6618 (1996)

[6] E. Lieberman-Aiden, N.L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B.R. Lajoie, P.J. Sabo, M.O. Dorschner, *et al*, Science **326** 289 (2009)

[7] S.F. Edwards, Proc. Roy. Soc., **91** 513 (1967)

[8] F. Wiegel, *Introduction to Path-Integrals Methods in Physics and Polymer Science* (Singapore: WSPC, 1986)

[9] M. Yor, J. Appl. Prob., **29** 202 (1992); Math. Finance, **3** 231 (1993)

[10] A. Grosberg, H. Frisch, J. Phys. A: Math. Gen., **37** 3071 (2004)

[11] E. Bogomolny *Scattering on two Aharonov-Bohm vortices*, arXiv:1508.04616

[12] A.V. Vologodskii, M.D. Frank-Kamenetskii, Usp. Fiz. Nauk, **134** 641 (1981) (in Russian); A.V. Vologodskii, A.V. Lukashin, M.D. Frank-Kamenetskii, V.V. Anshelevich, Zh. Exp. Teor. Fiz., **66** 2153 (1974); A.V. Vologodskii, A.V. Lukashin, M.D. Frank-Kamenetskii, Zh. Exp. Teor. Fiz., **67** 1875 (1974); M.D. Frank-Kamenetskii, A.V. Lukashin, A.V. Vologodskii, Nature, **258** 398 (1975)

[13] V.F.R. Jones, Ann. Math., **126**, 335 (1987); Bull. Am. Math. Soc., **12** 103 (1985)

[14] W.B.R. Likorish, Bull. London Math. Soc., **20** 558 (1988); W.B.R. Lickorish, *An Introduction to Knot Theory*, Springer Serirs: Graduate texts in Mathematics (Springer: 1997)

[15] J. Birman, *Knots, Links and Mapping Class Groups*, Ann. Math. Studies, **82** (Princeton Univ. Press, 1976)

[16] *Integrable Models and Strings*, Lect. Not. Phys., **436** (Springer: Heidelberg, 1994); M. Wadati, T.K. Deguchi, Y. Akutsu, Phys. Rep., **180** 247 (1989)

[17] L.H. Kauffman, H. Saleur, Comm. Math. Phys., **141** 293 (1991)

[18] V.A. Vassiliev, *Complements of Discriminants of Smooth Maps: Topology and Applications* in Math. Monographs, **98** (Trans. AMS, 1992); D. Bar-Natan, Topology, **34** 423 (1995)

[19] M. Khovanov, *A categorification of of the Jones Polynomial*, arXiv:math.QA/9908171; D. Bar-Natan, J. Knot Theory Ramifications, **16** 243 (2007)

[20] P. Tait, Trans. Royal Soc. Edinburgh **28** 145 (1877)

[21] P.Bangert, M.Berger, R. Prandi, J. Phys. (A): Math. Gen., **35** 43 (2002)

[22] S.R. Quake, Phys. Rev. Lett., **73** 3317 (1994)

[23] Y. Sheng, P. Lai, Phys. Rev. E, **63** 021506 (2001)

[24] E. Janse van Rensburg, D. Sumners, E. Wasserman, S. Whittington, J. Phys. A: Math. Gen, **25** 6557 (1992)

[25] C. Soteros, D. Sumners, S. Whittington, Math. Proc. Camb. Phil. Soc., **111**, 75 (1992)

[26] M. Freedman, Z. He, Z. Wang, Ann. Math. **139** 1 (1994)

[27] A. Kholodenko, D. Rolfsen, J. Phys. (A): **29** 5677 (1996)

[28] O. Karpenkov, A.B. Sossinsky, Russ. J. Math. Phys., **18** 306 (2011), arXiv:1106.3414

[29] *Ideal Knots*, eds. A. Stasiak, V. Katrich, L.H. Kauffman, *Series of Knots and Everything*, **19** (WSPC: Singapore, 1998)

[30] A.Yu. Grosberg, S. Nechaev, Europhys. Lett., **20** 603 (1992)

[31] O.A. Vasilyev, S.K. Nechaev, JETP, **93** 1119 (2001); O.A. Vasilyev, S.K. Nechaev, Theor. Math. Phys., **134** 142 (2003)

[32] P.G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell Univ. Press: Ithaca, 1979)

[33] E. Helfand, D. Pearson, J. Chem. Phys., **79** 2054 (1983); M. Rubinstein, E. Helfand, J. Chem. Phys., **82** 2477 (1985)

[34] A. Khokhlov, S. Nechaev, Phys. Lett. (A) **112** 156 (1985)

[35] S. Nechaev, Ya.G. Sinai, Bol. Soc. Bras. Mat., **21** 121 (1991)

[36] S. Nechaev, O. Vasilyev, J. Knot Theory Ramifications, **14** 243 (2005); S. Nechaev, O. Vasilyev, *Thermodynamics and topology of disordered knots: Correlations in trivial lattice knot diagrams*, in Physical and Numerical Models in Knot Theory, chapter 22, 421, *Series on Knots and Everything*, (WSPC: Singapore, 2005)

[37] S.K. Nechaev, A.N. Semenov, M.K. Koleva, Physica A, **140** 506 (1987)

[38] S.K. Nechaev, J. Phys. A: Math. Gen., **21** 3659 (1988)

[39] Li-Chien Shen, On Hecke groups, Schwarzian triangle functions and a class of hyper-elliptic functions, Ramanujan J. 39, 609 (2016)

[40] W. Koppenfels, F. Stallman, *Praxis der conformen abbildung*, (Berlin: Springer, 1959)

[41] C. Carathéodory, *Functiontheorie II*, (Basel, 1950)

[42] V.V.Golubev, *Lectures on Analytic Theory of Differential Equations* (Moscow: GITTL, 1950)

[43] E. Hille, *Ordinary Differential Equations in the Complex Domain*, (J. Willey & Sons: 1976)

[44] K. Chandrassekharan, *Elliptic Functions* (Berlin: Springer, 1985)

[45] M. Mumford, *Tata Lectures on Theta, I, II*, Progress in Mathematics **28**, **34** (Boston, MA: Birkhauser, 1983)

[46] H. Furstenberg, Trans. Am. Math. Soc. **198** 377 (1963)

[47] I.M. Lifshitz, A.Yu. Grosberg, JETP, **65** 2399 (1973)

[48] M. Tesi, E. Janse van Rensburg, E. Orlandini, S. Whittington, J. Phys. A: Math. Gen., **27** 347 (1994)

[49] E. Orlandini, M. Tesi, S. Whittington, J. Phys. (A): Math. Gen., **33** L-181 (2000)

[50] M. Imakaev, K. Tchourine, S. Nechaev, L. Mirny, Soft Matter 11 (2015), 665

[51] V. G. Rostiashvili, N.-K. Lee, T. A. Vilgis, J. Chem. Phys., **118** 937 (2002)

[52] C. F. Abrams, N.-K. Lee, S. Obukhov, Europhysics Letters, **59** 391 (2002)

[53] B. Chu, Q. Ying, A. Y. Grosberg, Macromolecules, **28** 180 (1995)

[54] A. Chertovich, P. Kos, J. Chem. Phys., **141** 134903 (2014)

[55] L. A. Mirny, Chromosome Research, **19** 37 (2011)

[56] R. D. Schram, G. T. Barkema, and H. Schiessel, J. Chem. Phys., **138** 224901 (2013)

[57] Y. Markaki, M. Gunkel, L. Schermelleh, S. Beichmanis, J. Neumann, M. Heidemann, H. Leonhardt, D. Eick, C. Cremer, T. Cremer, Cold Spring Harbor Symp. Quant. Biol., **75** 475 (2010)

[58] J. Dekker, M.A. Marti-Renom, L. A. Mirny, Nat. Rev. Genet., **14**, 390 (2013)

[59] A. Rosa and R. Everaers, PLoS Comput. Biol., **4** e1000153 (2008)

[60] A. Y. Grosberg, Polym. Sci., Ser. C, **54** 1 (2012);

[61] T. Vettorel, A. Y. Grosberg, K. Kremer, Phys. Biol., **6** 025013 (2009); J. D. Halverson, W. B. Lee, G. S. Grest, A. Y. Grosberg, K. Kremer, J. Chem. Phys., **134** 204904 (2011)

[62] J. Smrek, A.Y. Grosberg, Understanding the dynamics of rings in the melt in terms of annealed tree model, arXiv:1409.1483

[63] J. Dorier, A. Stasiak, Nucleic Acids Res., **37** 6316 (2009)

[64] V. A. Avetisov, V. Ivanov, D. Meshkov S. Nechaev, JETP Lett., **98** 242 (2013); V.A. Avetisov, V.A. Ivanov, D.A. Meshkov, S.K. Nechaev, Biophysical Journal, **107** 2361 (2014)

[65] J. D. Halverson, J. Smrek, K. Kremer, A. Y. Grosberg, Rep. Prog. Phys., **77** 022601 (2014)

[66] P. Virnau, L.A. Mirny, M. Kardar, PLoS Comput. Biol., **2** e122 (2006)

[67] G. Kolesov, P. Virnau, M. Kardar and L. A. Mirny, Nucleic Acids Res., **35** W-425 (2007)

[68] A. Khokhlov, S. Nechaev, J. de Physique II, **6** 1547 (1996)

[69] A.Yu. Grosberg, S.K. Nechaev, Macromolecules, **24** 2789 (1991)

[70] K. Urayama, S. Kohjiya, Polymer **38** 955 (1997)

[71] J. Ma, J.E. Straub, E.I. Shakhnovich, J. Chem. Phys., **103** 2615 (1995)

[72] R. Lua, A. L. Borovinskiy and A. Y. Grosberg, Polymer, **45** 717 (2004)

[73] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, Science **295** 1306 (2002)

[74] J.E. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J.S. Liu, and B. Ren, Nature **485** 376 (2012)

[75] T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, G. Cavalli, Cell **148** 458 (2012)

[76] Y. Zhang, R.P. McCord, Y.-J. Ho, B.R. Lajoie, D.G. Hildebrand, A.C. Simon, M.S.Becker, F.W. Alt,J. Dekker, Cell **148** 908 (2012)

[77] S. Sofueva, E. Yaffe, W.-C. Chan, D. Georgopoulou, M.V. Rudan, H. Mira-Bontenbal, S.M. Pollard, G.P. Schroth, A. Tanay, S. Hadjur, The EMBO Journal, advance online publication (2013) doi:10.1038/emboj.2013.237

[78] T.B.K. Le, M.V. Imakaev, L.A. Mirny, M.T. Laub, Science **342** 731 (2013)

[79] A.Yu. Grosberg, Soft Matter, **10**, 560 (2014)

[80] R.K. Sachs, G. van der Engh, B. Trask, H. Yokota, and J.E. Hearst, Proc. Nat. Acad. sci., **92**, 2710 (1995)

[81] C. Münkel, and J. Langowski, Phys. Rev. E, **57**, 5888 (1998)

[82] J. Ostashevsky, Mol. Biol. of the Cell, **9** 3031 (1998)

[83] J. Mateos-Langerak, M. Bohn, W. de Leeuw, O. Giromus, E. M. M. Manders, P. J. Verschure, M. H. G. Indemans, H. J. Gierman, D. W. Heerman, R. van Driel, and S. Goetze, Proc. Nat. Acad. Sci., **106**, 3812 (2009)

[84] B.V.S. Iyer, and G. Arya, Phys. Rev. E, **86**, 011911 (2012)

[85] M. Barbieri, M. Chotalia, J. Fraser, L.-M. Lavitas, J. Dostie, A. Pombo, and M. Nicodemi, Proc. Nat. Acad. Sci., **109**, 16173 (2012)

[86] C.C. Fritsch, J. Longowski, Chromosome Res., **19**, 63 (2011)

[87] A. Rosa, R. Everaers, Phys. Rev. Letters, **112**, 118302 (2014)

[88] M. Tamm, L. Nazarov, A. Gavrilov, and A. Chertovich, Phys. Rev. Lett., **114** 178102 (2015)

[89] V.A. Avetisov, L. Nazarov, S.K. Nechaev, M.V. Tamm, Soft Matter, **11** 1019 (2015)

[90] G.J. Filion, J.G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L.D. Ward, W. Brugman, I. de Castro Genebra de Jesus, R.M. Kerkhoven, H.J. Bussemaker, B. van Steensel, Cell, **143**, 212 (2010)

[91] J. Desbois, S. Nechaev, **88** 201 (1997); J. Phys. A: Math. Gen., **31** 2767 (1998); A. Comtet, S. Nechaev, J. Phys. A: Math. Gen., **31** 5609 (1998)

[92] A.M. Vershik, S. Nechaev, R. Bikbov, Statistical properties of locally free groups with application to braid groups and growth of heaps, Comm. Math. Phys., 212 (2000), 469-501

[93] N. Haug, S. Nechaev, M. Tamm, J. Stat. Mech. P-10013 (2014)

[94] P. Serna, G. Bunin, A. Nahum, Phys. Rev. Lett., **115** 228303 (2015)

[95] A. Gorsky, S. Nechaev, R. Santachiara, G. Schehr, Nuclear Physics B **862** [FS] 167 (2012)

[96] G. Bunin, M. Kardar, Coalescence Model for Crumpled Globules Formed in Polymer Collapse, Phys. Rev. Lett., **115** 088303 (2015)

[97] G. Palyi, C. Zucci, L. Caglioti (Eds.) *Fundamentals of Life* (Elsevier: Paris, 2002); Yu.N. Zhuravlev, V.A. Avetisov, Biogeosciences, **3** 281 (2008); Yu.N. Zhuravlev, V.A. Avetisov, *Hierarchical scale-free presentation of biological realm - Its Origin and Evolution* in *Biosphere Origin and Evolution*, N. Dobretsov, N. Kolchanov A. Rozanov, G. Zavarzin (Eds.), (Springer: New York, 2008), 69; S. Wright, *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*, Proceedings of the Sixth International Congress on Genetics 355 (1932); V. A. Avetisov, V.I. Goldanskii, PNAS **93** 11435 (1996)

[98] V.A. Avetisov, S.K. Nechaev, Geochemistry International, **52** 1252 (2014)

16

Introduction to abelian and non-abelian anyons

Sumathi Rao

In this set of lectures, we will start with a brief pedagogical introduction to abelian anyons and their properties. This will essentially cover the background material with an introduction to basic concepts in anyon physics, fractional statistics, braid groups, and abelian anyons. The next topic that we will study is a specific exactly solvable model, called the toric code model, whose excitations have (mutual) anyon statistics. Then we will go on to discuss non-abelian anyons, where we will use the one dimensional Kitaev model as a prototypical example to produce Majorana modes at the edge. We will then explicitly derive the non-abelian unitary matrices under exchange of these Majorana modes.

16.1 Introduction

The first question that one needs to answer is why we are interested in anyons [1]. Well, they are new kinds of excitations which go beyond the usual fermionic or bosonic modes of excitations, so in that sense they are like new toys to play with! But it is not just that they are theoretical constructs - in fact, quasi-particle excitations have been seen in the fractional quantum Hall (FQH) systems, which seem to obey these new kind of statistics [2]. Also, in the last decade or so, it has been realized that if particles obeying non-abelian statistics could be created, they would play an extremely important role in quantum computation [3]. So in the current scenario, it is clear that understanding the basic notion of exchange statistics is extremely important.

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_16

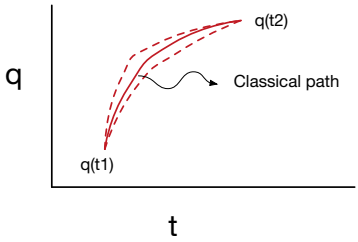


Figure 16.1: Sum over paths from $q(t1)$ to $q(t2)$

So if we want to start by explaining [4–6] what an 'anyon' is, to someone who may be hearing the word for the first time, we can tell them that just as fermions are particles obeying Fermi-Dirac statistics and bosons are particles obeying Bose-Einstein statistics, anyons are particles obeying 'any' statistics. Clearly, they are not as ubiquitous as bosons and fermion, else they would have been just as familiar to everyone as bosons and fermions. But as we shall see later in this lecture, even theoretically, anyons can only occur in two dimensions, whereas the world is three dimensional. So it is only in planar systems, or in systems where the motion in the third dimension is essentially frozen, that excitations can be anyonic.

Hence, although the theoretical possibility of anyons was studied as early as 1977 [1], it shot into prominence only in the late eighties and early nineties, when not only the excitations in the FQH systems were found to be anyonic, for a while, there was also speculation that anyons could explain the unusual features of high temperature superconductivity [7].

The easiest way to understand the notion of phases and statistics under exchange of particles in quantum mechanics is to think about how particles move around each other and from this point of view, the easiest way of understanding the quantum motion of these particles is via path integrals. Here, we will assume that you have some familiarity with the idea of path integrals, although not many details will be required. To recollect it, we just mention the following few things. In quantum mechanics, the probability amplitude to go from one space-time point to another is given by

$$A = \sum_{paths} e^{i\mathcal{S}} \ , \tag{16.1}$$

where $\mathcal{S} = \int \mathcal{L} dt$ is the action for the particular trajectory or path. In other words, quantum mechanically, we need to include all possible paths between the initial and final points of the trajectory see Fig.16.1). But most of these paths will interfere destructively with each other and hence will not contribute to the probability amplitude. The only exception is the classical path and the paths close to it, which interfere constructively with the classical path - i.e.,

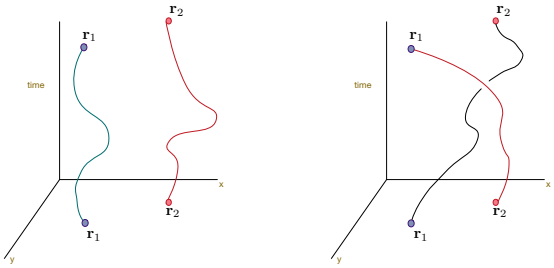


Figure 16.2: Direct and exchange paths

the most important contribution will be from the classical path. That is all the information that we need here.

With this introduction, we come to the details of what we will study here. In Sec.(II), we will explain the basic notions of anyon physics and why they can exist only in two spatial dimensions. We will analyze a simple physical model of an anyon and use it to understand the quantum mechanics of two anyons and see that even in the absence of any interactions, it needs to be studied as an interacting theory, with the interactions arising due to the anyonic exchange statistics. Then, in Sec.(III), we will study the exactly solvable toric code model as an example of a system with anyonic excitations. Finally, in Sec.(IV), we will discuss non-abelian statistics, where again, we will explain many features of non-abelian anyons using the one-dimensional Kitaev model as a typical example.

16.2 Abelian anyons

16.2.1 Basic concepts of anyon physics

The term 'exchange statistics' refers to the phase picked up by a wave-function when two identical particles are exchanged. But this definition is slightly ambiguous. Does statistics refer to the phase picked up by the wave-function when all the quantum numbers of the particles are exchanged (i.e., under permutation of the particles) or the actual phase that is obtained when two particles are adiabatically transported giving rise to the exchange? In three dimensions, these two definitions are equivalent but not in two dimensions. In quantum mechanics, we deal with interference of paths of particles and hence, it is the second definition which is more relevant, and we will show how it can be different from the first definition.

Let us first consider the statistics under exchange of two particles in three dimensions. By the path integral prescription, the amplitude for a system of

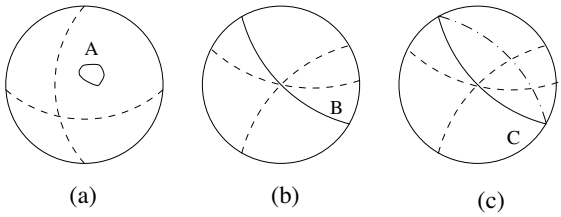


Figure 16.3: Paths in three dimensional configuration space with fixed radius

particles that moves from $(\mathbf{r}_1(t_1), \mathbf{r}_2(t_1))$ to $(\mathbf{r}_1{}'(t_2), \mathbf{r}_2{}'(t_2))$ is given by

$$A = \sum_{paths} e^{i \int_{t_1}^{t_2} dt \mathcal{L}[\mathbf{r}_1(t), \mathbf{r}_2(t)]} \ . \tag{16.2}$$

If the two particles are identical, then there are two classes of paths (see Fig.16.2). How do we see this? If we use the convention that we always refer to the position of the first particle first and the second particle second, then we see that the final configuration remains the same whether we have $(\mathbf{r}_1(t_2), \mathbf{r}_2(t_2))$ or $(\mathbf{r}_2(t_2), \mathbf{r}_1(t_2))$ because the two particles are identical. Even though the particles are exchanged in one path and not in the other, the final configuration is the same. In terms of the centre of mass ($\mathbf{R} = (\mathbf{r}_1 + \mathbf{r}_2)/2$) and relative coordinates ($\mathbf{r} = \mathbf{r}_1 - \mathbf{r}_2$), we see that the centre of mass motion is the same for both the paths, but the relative coordinate changes for both the paths. Also, since the CM motion moves both the particles together, it is independent of any possible phase under exchange. For convenience in visualizing the configuration space, let us keep $|\mathbf{r}|$ fixed and non-zero, i.e, the two particles do not intersect. Then, the vector $\mathbf{r}$ takes values on the surface of the sphere.

Now let us draw the paths in configuration space as shown in Fig.16.3. It is clear that the paths can only move along the surface of the sphere as the two particles move, since $|\mathbf{r}|$ is fixed. But once they get back to their original positions or get exchanged, since they are indistinguishable, the path is closed. In other words, closed paths on the surface of the sphere are formed by the particles coming back to their original positions (no exchange) or going to the antipodal point ($\mathbf{r} \to -\mathbf{r}$ or getting exchanged). But if we exchange the particles another time, then $\mathbf{r}$ comes back to itself, after having gone around the sphere once. In terms of diagrams, this is shown in Fig.16.13. Since we have eliminated coincident points, the wave-function is non-singular and well-defined at all points in the configuration space, and consequently on the surface of the sphere. So the phase picked up by the wave-function is also well-defined and does not change under continuous deformations of the path. Let us consider the possible phases of the wave-function when the motion of the particles is along each of the three paths, - A (no exchange), B (single exchange) and C (two exchanges) - depicted in Figs. 16.3(a),(b) and (c). Path A is a closed path

which does not involve any exchange and can clearly be shrunk to a point. Hence, the wave-function cannot pick up any phase other than unity. Path B involves the exchange of the two particles and goes from a point on the sphere to its diametrically opposite point. Since the two end points are fixed, this path cannot be shrunk to a single point. So this exchange can have a non-trivial phase in the wave-function . However, path C which forms a closed loop on the surface and involves two exchanges can be continuously shrunk to a point by imagining the path to be a physical string looped around a sphere. So this again cannot pick up any phase. Let η be the phase picked up under a single exchange. Since two exchanges are equivalent to no exchange, $\eta^2 = +1 \Longrightarrow \eta = \pm 1$. Hence, the only statistics possible in three dimensions are Fermi statistics or Bose statistics.

With slightly more mathematical rigour, one can say that the configuration space of relative coordinates is given by $(\mathbb{R}_3 - origin)/\mathbb{Z}_2$. Here $\mathbb{R}_3$ is just the three dimensional Euclidean space spanned by the relative coordinate $\mathbf{r}$. We subtract out the origin because we have assumed that paths do not cross (which is true for all particles other than bosons, because of hard core repulsion and for bosons, is not relevant anyway, because the exchange phase is unity). The division by $\mathbb{Z}_2$ is because of the identification of $\mathbf{r}$ with $-\mathbf{r}$, which is because the particles are indistinguishable. To study the phase picked up by the wave-function of a particle as it goes around another particle, we need to classify all paths in this configuration space. The claim, from the pictorial analysis above, is that there are just two classes of paths. Mathematically, this is expressed in terms of the first homotopy group Π_1 of the space, which is the group of inequivalent paths (paths not deformable to each other), passing through a given point in the space, with group multiplication being defined as traversing paths in succession and group inverse as traversing a path in the opposite direction. Thus

$$\Pi_1(\mathbb{R}_3 - origin)/\mathbb{Z}_2) = \Pi_1(\mathbb{R}P_2) = Z_2, \tag{16.3}$$

where $\mathbb{R}P_2$ stands for real projective space and is the notation for the surface of the sphere with diametrically opposite points identified and $\mathbb{Z}_2 = (1, -1)$ is a group of just two elements.

Now that we have determined that there are two classes of paths in three dimensions, in terms of path integrals, the amplitude can be written as

$$A[\mathbf{r}_1(t_1), \mathbf{r}_2(t_1)) \rightarrow (\mathbf{r}_1{}'(t_2), \mathbf{r}_2{}'(t_2)] = \sum_{direct\ paths} e^{i\mathcal{S}} + \sum_{exchange\ paths} e^{i\mathcal{S}} \ . \tag{16.4}$$

The direct paths involve all closed paths which end at the same point and the exchange paths involve all paths which end on antipodal points, (which are also closed paths). In terms of the path integral, we can also introduce a phase between the two classes of paths and write

$$A[\mathbf{r}_1(t_1), \mathbf{r}_2(t_1)) \rightarrow (\mathbf{r}_1{}'(t_2), \mathbf{r}_2{}'(t_2))] = \sum_{direct\ paths} e^{i\mathcal{S}} + e^{i\phi} \sum_{exchange\ paths} e^{i\mathcal{S}} \ . \tag{16.5}$$

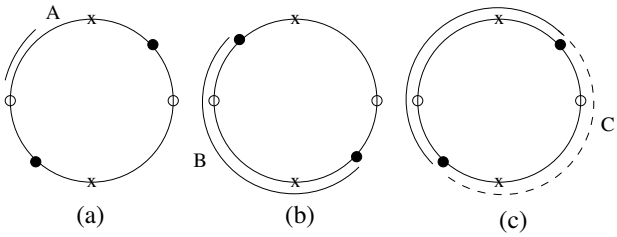


Figure 16.4: Paths in two dimensional configuration space with fixed radius

Since have already seen that exchanging the particle twice leads again to the direct path, it is clear that $e^{2i\phi} = 1$, which implies that ϕ can only be $0, \pi$ giving rise, as before, to bosons and fermions.

What changes in two dimensions? The point is that the topology of the configuration space is now different. In two spatial dimensions, the configuration space of the relative coordinates is given by $(\mathbb{R}_2 - origin)/\mathbb{Z}_2$. Just as before, for ease of visualization, we shall keep the magnitude of the relative coordinate fixed, so that configuration space can be represented by a circle, and since the particles are indistinguishable, diametrically opposite points are identified (see Figs.16.4(a),(b),(c)). Here, however, several closed paths are possible. The path A that involves no exchanges can obviously be shrunk to a point, since it only moves along the circle and back. But path B that exchanges the two particles is non-contractible since the end-points are fixed. But even path C, where both the solid and dashed line are followed in the clock-wise direction (or anti-clockwise direction) cannot be contracted to a point. This is easily understood by visualizing the paths as physical strings looping around a cylinder. Thus, if η is the phase under single exchange, η^2 is the phase under two exchanges, η^3 is the phase under three exchanges and so on. All we can say is that since the modulus of the wave-function remains unchanged under exchange, η has to be a phase - $\eta = e^{i\theta}$. This explains why we can get 'any' statistics in two dimensions.

The distinction between the paths in two and three dimensions can also be seen as follows. In three dimensions, the loop that is formed by taking a particle all around another particle (two exchanges) can be lifted off the plane and shrunk to a point as shown in Fig.16.5. This is not possible if the motion is restricted to a plane, as long as we disallow configurations where two particles are at the same point (removal of the origin).

The mathematical crux of the distinction between configuration spaces in two and three dimensions, is that the removal of the origin in two dimensional space, makes the space multiply connected (unlike in three dimensional space, where removal of the origin keeps it singly connected). So it is possible to define paths that wind around the origin. Mathematically, we can say that

$$\Pi_1((\mathbb{R}_2 - \text{origin})/\mathbb{Z}_2) = \Pi_1(\mathbb{R}P_1) = \mathbb{Z}, \tag{16.6}$$

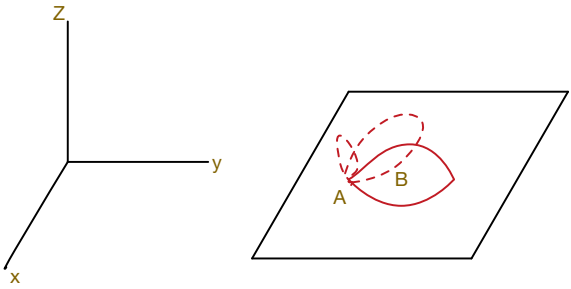


Figure 16.5: Path of A around B being lifted off the surface and shrunk to a point

where $\mathbb{Z}$ is the group of integers under addition. $\mathbb{R}P_1$ is just the notation for the circumference of a circle with diametrically opposite points identified. The different paths are labelled by integer winding numbers.

So in terms of path integrals, we now see that paths starting and ending at the same positions (upto exchanges due to indistinguishability) can be divided into an infinite number of classes, all distinct. So we can write the total amplitude as

$$A = \sum_{direct\ paths} e^{i\mathcal{S}} + e^{i\phi} \sum_{single\ exchange} e^{i\mathcal{S}} + e^{2i\phi} \sum_{two\ exchanges} e^{i\mathcal{S}} + \ldots, \quad (16.7)$$

where $\phi = 0, \pi$ give the usual bosons and fermions, but since in general, $e^{in\phi} \neq 1$ for any n, ϕ can be anything and as we said earlier, 'any' statistics are possible in two dimensions.

Note that if we do want to understand exchange of particles in the Hamiltonian formulation without invoking path integral ideas, we need to pin down the particles by using a confining potential - *i.e.*, by putting them in a box -

$$H = \sum_i \frac{\mathbf{p}_i^2}{2m} + \sum_i V_{\text{box}}(\mathbf{x}_i - \mathbf{R}_i). \quad (16.8)$$

The particles can now be moved around by changing $\mathbf{R}_i$ as a function of time. Since the particles are identical, exchanges are equivalent to closed paths, and do not depend on the geometry of the paths $\mathbf{R}_i(t)$ involved. So the statistics of the particles under exchange can be found by computing the Berry phase when the particles are exchanged. However, in this review, we shall basically use the path integral formalism.

16.2.2 Anyons obey braid group statistics

The distinction between the phase of the wave-function when the quantum numbers of the particles are exchanged and the phase obtained under adiabatic transport of particles should now be clear. Under the former definition,

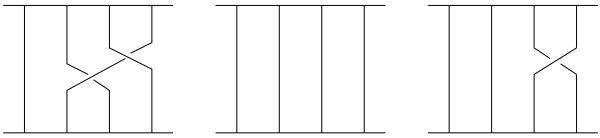


Figure 16.6: Elements of the braid group $\mathbb{B}_4$

the phase η^2 after two exchanges is always unity, whereas the phase under the latter definition has many more possibilities at least in two dimensions. Mathematically, the first definition classifies particle under the permutation group $\mathbb{P}_N$, whereas the second one classifies particles under the braid group $\mathbb{B}_N$. The permutation group $\mathbb{P}_N$ is the group formed by all possible permutations of N objects with group multiplication defined as successive permutations and group inverse as undoing the permutation. It is clear that permuting two objects twice brings the system back to the original configuration. Thus particles that transform as representations of the permutation group can only be bosons or fermions.

On the other hand, when we adiabatically exchange two particles, we can visualize the process as paths in space-time with time being the vertical axis and space being the horizontal axis as shown in Fig.16.2. The particles can circle around each other and form closed paths by coming back to their original positions (upto permutations of the positions). The adiabatic exchange of particles classifies particles under the braid group. As we saw earlier, even under adiabatic exchange, in three spatial dimensions, we only have fermions or bosons, whereas in two dimensions, there are many other possibilities. Formally, the braid group $\mathbb{B}_N$ is the group of inequivalent paths that occur when adiabatically transporting N particles. Since they represent a configuration of N particles, at some particular time (say $t = 0$), evolving to a configuration of N particles at some later time $t = T$, the world lines cannot cross each other or form knots around each other or loop back. At each time, we want to have only N particles. Each history or set of trajectories of the N particles becomes a braid. For example, in Fig.16.6, we show an example of some elements of the braid group $\mathbb{B}_4$, which is the braid group of 4 particles. Exchanges of neighbouring particles (by some counting rule, since the particles are in two dimensional space) form the generators of the group. For instance, the generators of the group $\mathbb{B}_4$ are given in Fig.16.7 and are denoted as σ_j, $j = 1, 2, 3$. σ_j describes exchange of j^{th} particle with $(j+1)^{th}$ particle in a counter-clockwise direction (by definition), so that the clockwise exchange is denoted by $(\sigma_j)^{-1}$. The identity element is given by σ_0 where there is no exchange, and group inverse by the clockwise exchange $(\sigma_j)^{-1}$ as shown in Fig.16.8. Group multiplication is defined as following one trajectory by another in time as shown in Fig.16.9. Note that we have put crosses on the time-lines which are identified (are at equal times)

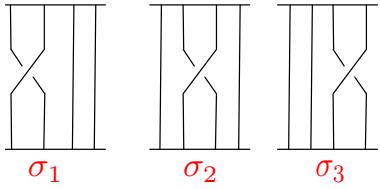


Figure 16.7: The three generators of the braid group $\mathbb{B}_4$

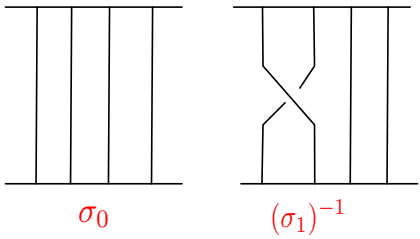


Figure 16.8: The identity and the inverse of the generator σ_1.

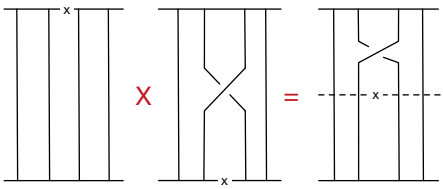


Figure 16.9: Group multiplication.

in the figure. It is now easy to check that $(\sigma_j)(\sigma_j)^{-1} = \sigma_0$ as shown in Fig.16.10 (without the crosses). It is also easy to see that $(\sigma_1)^n \neq \sigma_0$ for any n, which is the reason that 'any' statistics are allowed in two dimensions. (See Fig.16.11).

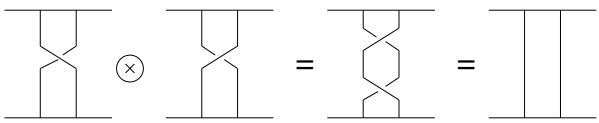


Figure 16.10: Product of σ_1 and $(\sigma_1)^{-1}$ giving rise to identity.

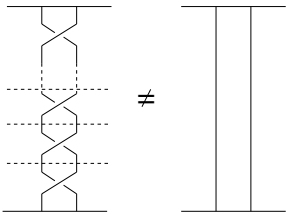


Figure 16.11: $(\sigma_1)^n \neq \sigma_0$

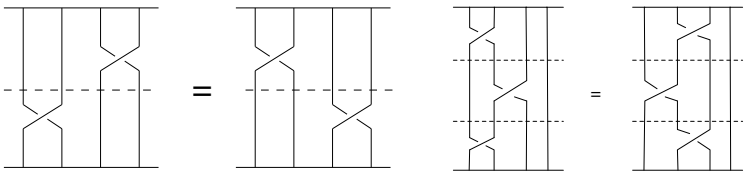


Figure 16.12: Yang-Baxter relations.

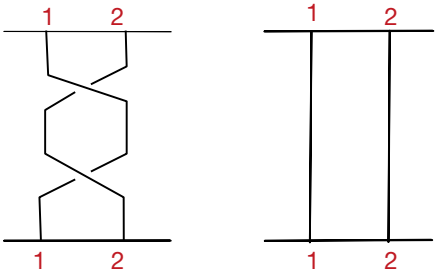


Figure 16.13: Different elements of the braid group, but same element of the permutation group.

We will end this subsection by mentioning the two defining relations satisfied by the generators of the braid group.

$$\begin{aligned} \sigma_i \sigma_j &= \sigma_j \sigma_i, \quad |i-j| \geq 2, \\ \sigma_j \sigma_{j+1} \sigma_j &= \sigma_{j+1} \sigma_j \sigma_{j+1}. \end{aligned} \tag{16.9}$$

The second one is called the Yang-Baxter relation. Both these relations can be easily checked pictorially (as we show in Fig.16.12 for the generators of $\mathbb{B}_4$).

It should be clear by now that the braid group leads to a much finer classification than the permutation group. For instance, the two elements shown in Fig.16.13 are different elements of the braid group, but the same element of the permutation group. So the quantum theory of anyons has the quantum

states of the anyons transforming as unitary representations of the braid group. Abelian anyons form one-dimensional representations of the braid group. There are an infinite number of such representations, because under exchange, the phase that is picked up is $e^{i\theta}$ and θ can take any value. $\theta = 0$ and $\theta = \pi$ represent bosons and fermions respectively. We will discuss non-abelian representations in the last section.

16.2.3 Spin of an anyon

Let us start with spin in the familiar three dimensional world. We know that spin is an intrinsic angular momentum quantum number that labels different particles. The three spatial components of the spin obey the commutation relations given by

$$[S_i, S_j] = \epsilon_{ijk} S_k. \tag{16.10}$$

We shall show that these commutation relations constrain the spin to be either integer or half-integer. Let $|s, m\rangle$ be the state with $S^2|s, m\rangle = s(s+1)|s, m\rangle$ and $S_z|s, m\rangle = m|s, m\rangle$. By applying the raising operator, we may create the state

$$S^+|s, m\rangle = [s(s+1) - m(m+1)]^{1/2}|s, m+1\rangle = |s, m'\rangle. \tag{16.11}$$

Requiring this state to have positive norm for all m, leads to $m < s$. Thus, it is clear that for some integer $m' = m+$integer, $m' > s$ unless $s = m'$, or

$$s - m = \text{integer}. \tag{16.12}$$

Similarly by insisting that $S^-|s, m\rangle$ have a positive norm, we get $s(s+1) - m(m-1) \geq 0$, which implies that $m \geq -s$ for all m. Once again, to avoid $m < -s$, we need to set

$$m - (-s) = \text{integer}. \tag{16.13}$$

Adding the equations in Eqs.16.12 and 16.13, we get

$$2s = \text{integer} \Longrightarrow s = \text{integer}/2\ . \tag{16.14}$$

Thus, just from the commutation relations, we can prove that the particles in three dimensions have either integer or half-integer spin.

However, in two dimensions, there exists only one axis of rotation, perpendicular to the plane of the two dimensions. Hence, here spin only refers to S_3 which has no commutation relations to satisfy, and hence it can be anything!

16.2.4 Physical model of an anyon

Now let us construct a simple physical model of an anyon [8]. Imagine a spinless particle of charge q orbiting around a thin solenoid along the z-axis at a distance $\mathbf{r}$ as shown in Fig16.14. When there is no current through the solenoid, the orbital angular momentum of the charged particle is quantized as an integer

- l_z = integer. When a current is turned on, the particle feels an electric field that can easily be computed using

$$\int (\nabla \times \mathbf{E}) d^2\mathbf{r} = \int B d^2\mathbf{r} = -\frac{\partial \phi}{\partial t}, \tag{16.15}$$

where ϕ is the total flux through the solenoid. This is just the Aharonov-Bohm effect. Hence,

$$\int \mathbf{E} \cdot d\mathbf{l} = 2\pi |\mathbf{r}| E_\theta = -\dot{\phi}, \quad \text{leading to} \quad \mathbf{E} = -\frac{\dot{\phi}}{2\pi |\mathbf{r}|} (\hat{z} \times \mathbf{r}) \ . \tag{16.16}$$

Thus, the angular momentum of the charge particle changes with the rate of change given being proportional to the torque - i.e.,

$$\dot{l}_z = \mathbf{r} \times \mathbf{F} = \mathbf{r} \times q\mathbf{E} = -\frac{q\dot{\phi}}{2\pi}, \quad \text{leading to} \quad \Delta l_z = -\frac{q\phi}{2\pi}. \tag{16.17}$$

Thus Δl_z is the change in the angular momentum due to the flux in the solenoid. In the limit where the solenoid becomes very narrow and the distance between the charged particle and the solenoid is shrunk to zero, the system may be considered as a single composite object - a charge-fluxtube composite. In fact in a planar system, there is no extension in the z direction. So this essentially point-like composite object with fractional angular momentum can be considered as a model of an anyon. Note that we have denoted this angular momentum as the change in the angular momentum due to the flux. So if we start with the original charge to be spinless, then the spin of the composite particle is given by $l_z = s_z = q\pi/2\pi$. This is also sometimes referred to as a topological spin and is intrinsic to the anyon. However, this is a little too naive. In an anyon, the charge and the flux it carries are related - the charge gets turned on along with the flux. This implies that the q in Eq.16.17 is time-dependent, and $q(t) = c\phi(t)$ for some constant c. Hence, we find

$$\Delta l_z = \frac{c\phi^2}{4\pi} = \frac{q\phi}{4\pi}, \tag{16.18}$$

so that the angular momentum of a charge-flux composite with charge proportional to flux is less than what we originally computed by a factor of 1/2. In the next subsection, we shall see that it has the right statistics, and complete the identification of the charge-fluxtube composite as an anyon.

16.2.5 Two anyon quantum mechanics

We shall now study the quantum mechanics of two anyons in order to determine its statistics using the simple physical picture of the anyon that we developed in the last subsection. The Hamiltonian for the system is given by

$$H = \frac{(\mathbf{p}_1 - q\mathbf{a}_1)^2}{2m} + \frac{(\mathbf{p}_2 - q\mathbf{a}_2)^2}{2m}, \tag{16.19}$$

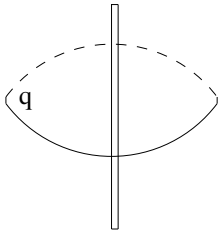


Figure 16.14: Physical model of an anyon.

with

$$\mathbf{a}_1 = \frac{\phi}{2\pi}\frac{\hat{z}\times(\mathbf{r}_1-\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|^2}, \quad \text{and} \quad \mathbf{a}_2 = \frac{\phi}{2\pi}\frac{\hat{z}\times(\mathbf{r}_2-\mathbf{r}_1)}{|\mathbf{r}_1-\mathbf{r}_2|^2}, \tag{16.20}$$

where $\mathbf{a}_1$ and $\mathbf{a}_2$ are the vector potentials at the positions of the composites (anyons) 1 and 2 due to the fluxes in composites (anyons) 2 and 1 respectively. Let us now work in the centre of mass (CM) and relative (rel) coordinates - i.e., we define respectively

$$\mathbf{R} = \frac{\mathbf{r}_1+\mathbf{r}_2}{2} \Rightarrow \mathbf{P} = \mathbf{p}_1+\mathbf{p}_2, \quad \text{and} \quad \mathbf{r} = \mathbf{r}_1-\mathbf{r}_2 \Rightarrow \mathbf{p} = \frac{\mathbf{p}_1-\mathbf{p}_2}{2}\ . \tag{16.21}$$

In terms of these coordinates, the Hamiltonian can be recast as

$$H = \frac{\mathbf{P}^2}{4m} + \frac{(\mathbf{p}-q\mathbf{a})^2}{m} \quad \text{with} \quad \mathbf{a}_{rel} = \frac{\phi}{2\pi}\frac{\hat{z}\times\mathbf{r}}{|\mathbf{r}|^2}\ . \tag{16.22}$$

Thus the CM motion which translates both the particles rigidly and is independent of the statistics is free. The relative motion, on the other hand, which is sensitive to whether the particles are bosons, fermion or anyons, reduces to the problem of a single particle of mass $m/2$ orbiting around a flux ϕ at a distance $\mathbf{r}$. Since the composites have been formed of a bosonic charge orbiting around a bosonic flux, the wave -function of the two composite system has to be symmetric under exchange and the boundary condition is given by

$$\psi(\mathbf{r}_1,\mathbf{r}_2) = \psi(\mathbf{r}_2,\mathbf{r}_1) \Longrightarrow \psi_{rel}(\mathbf{r}) = \psi_{rel}(-\mathbf{r}) \Longrightarrow \psi_{rel}(r,\theta+\pi) = \psi_{rel}(r,\theta), \tag{16.23}$$

where ψ_{rel} is the wave-function of the relative piece of the Hamiltonian and $\mathbf{r} = (r,\theta)$ in cylindrical coordinates.

Now, let us perform a (singular) gauge transformation so that

$$\mathbf{a}_{rel} \longrightarrow \mathbf{a}'_{rel} = \mathbf{a}_{rel} - \nabla\Lambda(r,\theta), \quad \text{where} \quad \Lambda(r,\theta) = \frac{\phi}{2\pi}\theta\ . \tag{16.24}$$

This gauge transformation is singular because θ is a periodic angular coordinate with period 2π and is not single-valued. In the primed gauge,

$$
\begin{aligned}
a'_{rel,\theta} &= a_{rel,\theta} - \frac{1}{r}\frac{\partial \Lambda}{\partial \theta} = \frac{\phi}{2\pi r} - \frac{\phi}{2\pi r} = 0, \\
\text{and} \quad a'_{rel,r} &= a_{rel,r} - \frac{\partial}{\partial r}\Lambda = 0 - 0 = 0,
\end{aligned} \tag{16.25}
$$

i.e., the gauge potential vanishes completely and the Hamiltonian just reduces to

$$
H = \frac{\mathbf{P}^2}{2m} + \frac{\mathbf{p}^2}{2m}, \tag{16.26}
$$

which is just the Hamiltonian of two free particles. However, in the primed gauge, the wave function of the relative Hamiltonian has also changed. It is now given by

$$
\psi'_{rel}(r, \theta + \pi) = e^{-iq\phi/2}\psi_{rel}(r, \theta) \ , \tag{16.27}
$$

-i.e., the particles obey anyonic statistics. So the problem of two free anyons is equivalent to the problem of two interacting charge flux composites described by the Hamiltonian in Eq.16.19. The quantum mechanical problem can be solved - we need to go back to the boson gauge (since we do not know how to solve quantum mechanics problems with non-trivial statistics) and write the Hamiltonian in terms of centre of mass (CM) and relative coordinates and note that the CM motion becomes free and the relative motion acquires an extra $q\phi/2m$ factor in the angular momentum term, thus adding to the centrifugal barrier. The radial part of the relative motion can then be identified as a Bessel equation. Thus the two anyon wave-function can be written as

$$
\psi(\mathbf{R}, \mathbf{r}) = \psi_{\rm CM}(\mathbf{R})\psi_{\rm rel}(\mathbf{r}) = e^{i\mathbf{P}\cdot\mathbf{R}}e^{i(l+q\phi/2m)\theta}J_{|l+q\phi/2m|}(kr), \tag{16.28}
$$

where $\mathbf{r} = (r, \theta)$. The two particle wave-function can be recast in terms of the original single particle coordinates - *i.e.*, $\psi((\mathbf{R}, \mathbf{r}) = \psi(\mathbf{r}_1, \mathbf{r}_2)$. However, unless $q\phi/2m$ is either integer or half-integer, the two particle wave-function cannot be factorized into a product of two suitable one-particle wave-functions. Also, the energy levels of the two anyon system cannot be obtained as sums of one anyon energy levels. This is easier to see with discrete energy levels and so we will next solve the problem of two anyons in a harmonic oscillator potential.

The Hamiltonian for two anyons in a harmonic oscillator potential is given by

$$
H = \frac{\mathbf{p}_1^2}{2m} + \frac{\mathbf{p}_2^2}{2m} + \frac{1}{2}m\omega^2\mathbf{r}_1^2 + \frac{1}{2}m\omega^2\mathbf{r}_2^2 \ . \tag{16.29}
$$

The problem can be separated into CM and relative coordinates in terms of which the Hamiltonian is given by

$$
H = \frac{\mathbf{P}^2}{4m} + \frac{\mathbf{p}^2}{m} + m\omega^2\mathbf{R}^2 + \frac{1}{4}m\omega^2\mathbf{r}^2 \ . \tag{16.30}
$$

The problem can now be solved in terms of the cylindrical (R, Θ) and (r, θ) coordinates. The CM motion is independent of the statistics of the particles

and one can simply find the energy levels as

$$E_{CM} = \omega(n + |L| + 1) \ . \tag{16.31}$$

The Hamiltonian for the relative motion can be solved in the same way, except that because of the phase under exchange, we need to go to the boson gauge, where there is a dependence on the statistical gauge field - *i.e.*, we have

$$H_{rel}\psi_{rel} = [\frac{(\mathbf{p} - q\mathbf{a}_{rel})^2}{m} + \frac{1}{4}m\omega^2\mathbf{r}^2]\psi_{rel} = E_{rel}\psi_{rel} \ . \tag{16.32}$$

In terms of the (r, θ) coordinates, we now find that the energy levels are given by

$$E_{rel} = \omega(n + |l + \alpha/\pi| + 1). \tag{16.33}$$

Note that $\alpha = 0$ and $\alpha = \pi$ give the usual energy levels for bosons and fermions respectively. Otherwise, they are given by

$$\begin{aligned} E_j &= (2j + 1 + \alpha/\pi), \quad \text{degeneracy factor} = j + 1, \\ E_j &= (2j + 1 - \alpha/\pi), \quad \text{degeneracy factor} = j \ . \end{aligned} \tag{16.34}$$

Clearly, the levels are not equally spaced and the total energy of the two anyon system given by

$$E_{2anyons} = E_{CM} + E_{rel} = (2j + p + 2 \pm \alpha/\pi)\omega, \quad p, j = \text{integers} \tag{16.35}$$

is not a sum of the one particle levels $E = (n+1)\omega$, with $n =$ integer. Similarly, the two anyon wave function is also not a simple product of one anyon wave functions. We find that

$$\begin{aligned} \psi(\mathbf{R}, \mathbf{r}) &= e^{-m\omega(R^2+r^2/4)}r^{\alpha/\pi}e^{i\alpha\theta/\pi} \\ \implies \psi(\mathbf{r}_1, \mathbf{r}_2) &\propto e^{-m\omega(r_1^2+r_2^2)/2}(\mathbf{r}_1 - \mathbf{r}_2)^{\alpha/\pi} \end{aligned} \tag{16.36}$$

which does not factorize into a product of two single particle wave-functions except when $\alpha = 0, \pi$. This is why even a system of free anyons needs to be tackled as an interacting problem. For more details, see Ref. [4].

16.2.6 Many anyon systems

Finally, we briefly mention what happens when we have many anyons. As we have seen above, even the two free anyon system is an interacting system because of the statistical interactions. Hence, any many anyon system needs to be treated as an interacting system, where each particle has long-range statistical interactions with each of the other particles.

Here, we will just mention one important concept of many anyon systems, which is that of fusion rules. A system which has anyons must have many types of anyons. If we have an anyon with statistics parameter θ, then we can combine two such anyons or make a bound state of two such particles. What would be

the statistics of the bound state? You may naively think that it should be 2θ. But that is not correct. One can see this by thinking of the anyon as a charge-flux composite in two ways. (1) In terms of angular momentum, we can add up the individual spins of the two anyons: $\theta/2\pi + \theta/2\pi = \theta/\pi$. But we also need to include the orbital angular momentum of the anyon pair. Normally, the orbital angular momentum is an integer because the wave-function of the two particle system is single-valued, if you take one particle around the other. But here, the counter-clockwise transport of one anyon around another in a full circle leads to the phase $e^{-2i\theta} = e^{i2\pi L}$ where L is the orbital angular momentum. Now adding the spin and orbital angular momentum, we find that the total angular momentum is $\theta/\pi + \theta/\pi = 2\theta/\pi = 4\theta/2\pi$ which gives the statistics parameter as 4θ. (2) Here, we note that a single anyon exchange leads to a phase of $e^{i\theta}$. So when a two anyon molecule exchanges with another 2 anyon molecule, there are $2^2 = 4$ exchanges and hence the phase acquired will be $e^{4i\theta}$, which agrees with the total spin of the bound state as well.

Now, we can generalize this to say that more particles can be bound together to form new bound states or new types of particles. This is called fusion. The statistics parameter when n such particles are bound together is given by $e^{in^2\theta}$. One can think of this new particle as an n-charge-n-flux composite. The formation of a different type of anyon by bringing together two anyons is called fusion. If we bring together an anyon and anti-anyon with opposite statistics parameter (θ and $-\theta$), the result has statistics zero, which is equivalent to having no particles. The system with no particles (called the vacuum) is often denoted by the identity I. It is also called a trivial particle.

For abelian anyons, it is clear that if we bring together 2 anyons with statistics parameter θ, they give rise to an anyon with statistics parameter 4θ and in general n such particles give rise to anyons with statistics parameter $n^2\theta$. But for non-abelian anyons, this is no longer true. There is no unique way of combining anyons to form new anyons and one can have different outcomes by bringing them together (just like two spin1/2 particles can be brought together to form spin 0 or spin1 particles). These are called fusion channels. The probability of the different outcomes is specified by a set of numbers which give rise to the fusion rules.

In the next section, instead of studying anyons abstractly as we have done in this section, we shall study an explicit lattice model, whose excitations turn out to be abelian anyons. We will come back to this later when we study non-abelian anyons.

16.3 Toric code model as an example of abelian anyons

Let us begin this section, by first answering two questions - what is the toric code and why do we want to study it. The toric code is actually a spin model defined on a two dimensional lattice. It was engineered by Kitaev [9] to be exactly solvable and to have low energy excitations that are anyonic. The reason that this model became so important was because it was shown by Kitaev that

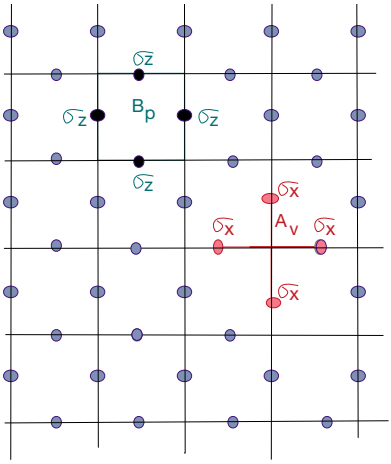


Figure 16.15: The Kitaev toric code model on a square lattice. Spins are placed on the links. The operator A_v is the product of the x-components of the spin of the four links that cross at the vertex and the operator B_p is the product of the z-component of the spins around the perimeter of the square.

these anyons could be used, in principle, to perform fault tolerant quantum computation – fault tolerant because information could be stored in the fusion properties of anyons, which could not be destroyed by local perturbations. This model is thus, the prototypical concrete lattice model for many of the more abstract ideas of quantum computation.

16.3.1 Toric code on a square lattice

In general, the toric code model can be defined on any lattice, but here we will work on the simplest model which is an exactly solvable spin 1/2 model on a two dimensional square lattice of $N \times N$ points [10, 11]. The spins are placed on the edges or links of an open lattice as shown in Fig.16.15 and it is easy to check that we have twice as many spins as the number of lattice points - $2N^2$. The components of the spins on different links commute with one another. On a given link, the spins satisfy the usual anti-commutation relations $\{\sigma_\alpha, \sigma_\beta\} = 2\delta_{\alpha\beta}$ where $\alpha, \beta = x, y, z$. The Hamiltonian for the model is given by [10]

$$H = -J_e \sum_v A_v - J_m \sum_p B_p \tag{16.37}$$

where

$$A_v = \Pi_{j \in v} \sigma_j^x \quad \text{and} \quad B_p = \Pi_{j \in p} \sigma_j^z \tag{16.38}$$

are the vertex operator that acts on the four spins surrounding the vertex and the plaquette operator involving the four spins around the plaquette, respec-

tively, as shown in Fig.16.15. Note that this is not a very common or physical looking Hamiltonian since it has four spin interactions, and has explicitly been engineered for a purpose! As can be easily seen from the diagram, every spin is a part of two vertex and two plaquette operators. The eigenvalues of both A_v and B_p are ± 1. However, the product of the eigenvalues over all the plaquettes or all the vertices is always unity - i.e.,

$$\Pi_{j\in V} A_V = \Pi_{j\in P} B_p = +1, \tag{16.39}$$

where V = total number of vertices in the model is N^2 and P = total number of plaquettes in the model is also N^2.

Note that the vertex operators contain the x-component of the spins and the plaquette operators contain the z-component of the spins. Hence, it is easy to check that the vertex operators and the plaquette operators commute among themselves

$$[A_v, A_{v'}] = [B_p, B_{p'}] = 0. \tag{16.40}$$

But it is also true that $[A_v, B_p] = 0$. This is trivially true if they do not have any spins in common, since spins on different sites commute. But in case, they do have a spin in common, they will always have two spins in common. For example, from Fig.16.16, it is clear that the plaquette B_{α_1} shares spins with the four vertex operators $A_{\beta_1}, ...A_{\beta_4}$. But with each of them, it shares 2 spins. For instance, we have

$$[B_{\alpha_1}, A_{\beta_1}] = \sigma_1^z\sigma_2^z\sigma_3^z\sigma_4^z\sigma_1^x\sigma_2^x\sigma_5^x\sigma_6^x - \sigma_1^x\sigma_2^x\sigma_5^x\sigma_6^x\sigma_1^z\sigma_2^z\sigma_3^z\sigma_4^z, \tag{16.41}$$

which is zero because $\sigma_1^z\sigma_1^x = -\sigma_1^x\sigma_1^z$ and $\sigma_2^z\sigma_2^x = -\sigma_2^x\sigma_2^z$. So the two negative signs cancel each other. The same thing goes through for the other three vertex operators as well. So the bottomline is that all the terms in the Hamiltonian commute with each other, and commute with the Hamiltonian. So all the A_v and B_p operators can be simultaneously diagonalized and their values can be used to label the states.

The next step is to find the ground state of the Hamiltonian. This means that the energy of all the terms in the Hamiltonian have be minimized, which, in turn means that each of the A_v and B_p terms have to be maximized. Let us work in the σ_z diagonal basis. The eigenvalues of σ_j^z are $s_j = \pm 1$. Let us also define $\omega_p(s)$, the product of the eigenvalues around the plaquette p (where s stands for the configuration of $\{s_j\}$). This can also be just $+1$ or -1. We will call this the flux through the plaquette. A configuration where $\omega_p = -1$ is called a vortex (or flux) configuration. (See Fig. 16.17 for examples.) Now suppose that we have only B_p terms in the Hamiltonian, i.e., $H = -J_m \sum_p B_p$. Here, again the only possible eigenvalues for the operator B_p are $+1$ and -1 and the configurations are all given in Fig. 16.17. We note that there are eight configurations with 'no flux' ($\omega_p = +1$) and eight configurations with non-zero flux (vortex configurations with $\omega_p = -1$) as shown in Fig.16.17. As to why these are called flux configurations, it turns out that this toric code model turns out to be the same as the Z_2 gauge theory on a lattice, and the flux

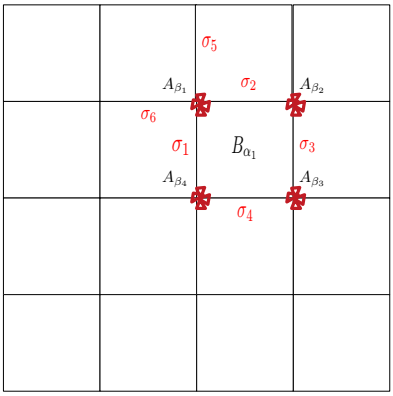


Figure 16.16: Commutation relations of the operators on different links and same links. See text for details.

is that of the Z_2 gauge field. But the study of this connection is beyond the scope of these lectures and for us here, the word flux can be thought of as just nomenclature. The ground state is now clearly given by any linear combination of configurations which have no vortices -i.e, the ground state is given by

$$|\psi\rangle = \sum_{\{s,\omega_p(s)=+1\forall p\}} C_s|s\rangle, \tag{16.42}$$

where C_s is arbitrary. All we know about the ground state is that it has no vortices and

$$B_p|\psi\rangle = |\psi\rangle \ \ \forall \ B_p, \tag{16.43}$$

because the eigenvalue of B_p acting on $|s\rangle$ is always $+|s\rangle$. For $2N^2$ spins, the ground state degeneracy would be 2^{N^2} (because pairs have to be either $\uparrow$ or $\downarrow$). In other words, of the total number of configurations 2^{2N^2}, 2^{N^2} would have $\omega_p = +1$ (ground state configuration) and 2^{N^2} would have $\omega_p = -1$ (vortex configuration).

Now let us add back the vertex terms to the Hamiltonian. The A_v acts on $|s\rangle$ by flipping spins, since $\sigma_x\sigma_z\sigma_x^{-1} = -\sigma_z$, but in any plaquette, it will always flip two spins. So it will keep configurations with $\omega_p = 1$ in configurations with $\omega_p = 1$. Hence, operation of A_v on $|s\rangle$ will only take it to some other $|s'\rangle$, which will also belong to the same set of vortex free configurations with $\omega_p(s) = 1$. But since A_v can act on any of the lattice points, A_v can be an eigen-operator for $|\psi\rangle$ only if $C_s = +1$ for all s. So now, we define

$$|\Psi_0\rangle = \sum_{\{s,\omega_p(s)=+1\forall p\}} |s\rangle, \tag{16.44}$$

as the ground state with all A_v and B_p acting on it with eigenvalue 1. (We could have worked in a σ_x diagonal basis, and defined a state $|s'\rangle$ with 'no (x)

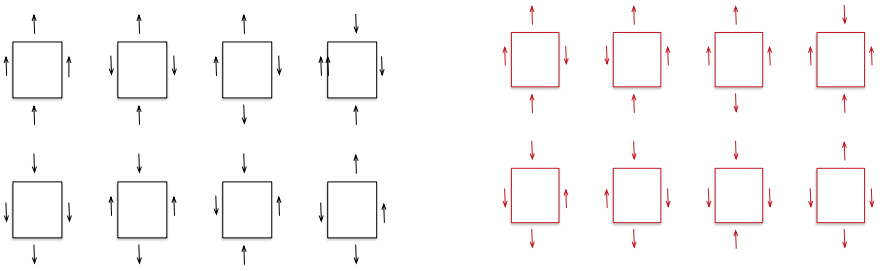


Figure 16.17: The configurations of spins on each plaquette on the left have $\omega_p(s) = 1$ and the configurations on the right have $\omega_p(s) = 1$

vortices' in the dual lattice configuration so that A_v acting on the state would give $+1$ for all v and repeated the same argument. So it is clear that this ground state is a sum over all spin configurations that have neither x nor z 'vorticity'.) It has an energy $E = -2N^2 J_e - 2N^2 J_m$. Any excitation over this ground state must have non-zero vorticity which would imply that at least one of the spins would have to flip in either the z or x directions. In that case, 2 of the B_p's or 2 of the A_v's would have eigenvalues -1 and hence, the energy of the excitation would be either $4J_m$ or $4J_e$, but we will come back to these excitations later.

So we have found the ground state of an interacting Heisenberg spin model in two dimensions exactly, essentially because the Hamiltonian has been constructed to be exactly solvable. The ground state can also be written as

$$|\Psi_0\rangle = N \prod_v (1 + A_v)|\xi\rangle, \tag{16.45}$$

where $|\xi\rangle$ is some reference state. For example, we can take $|\xi\rangle$ to be the state with all spins pointing $\uparrow$. The easiest way to check that this is the ground state is to check that for all A_v and B_p, we get the eigenvalue $+1$, when they act on this state. Let us check this.

$$A_{v'}|\Psi_0\rangle = A_{v'} N \prod_v (1 + A_v)|\xi\rangle. \tag{16.46}$$

First consider the terms where $v \neq v'$. Then $A_{v'}A_v = A_v A_{v'}$. But for $v = v'$, $A_{v'}(1 + A_{v'}) = A_{v'} + 1$, since $A_{v'}^2 = 1$. Hence

$$A_{v'}|\Psi_0\rangle = +1|\Psi_0\rangle. \tag{16.47}$$

Furthermore, we already know that A_v acting on any state does not change its z-vorticity since it always flips two spins. Since the reference state has vorticity $= +1$ and $B_p = +1$ on all states with vorticity $=+1$, we also have

$$B_p|\Psi_0\rangle = +1|\Psi_0\rangle \ . \tag{16.48}$$

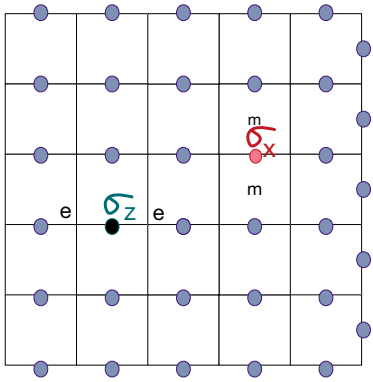


Figure 16.18: Flipping the z-component of the spin by σ_x creates 2 monopoles, in adjacent plaquettes whereas flipping the x-component of the spin by σ_z creates 2 charges at adjacent vertices.

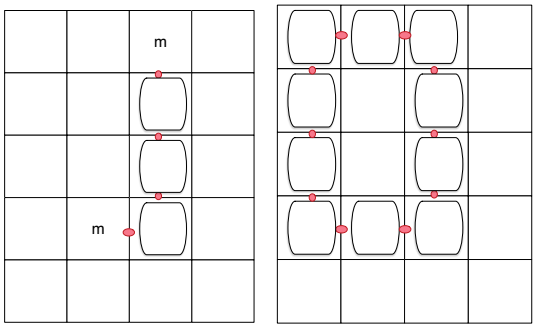


Figure 16.19: (a) A pair of monopoles created by flipping the z-component of the spin by σ_x along a string on the dual lattice. (b) Annihilating the monopoles by closing the string. This configuration again has zero vorticity and commutes with the Hamiltonian. (Note that the spins that do not change are not shown in this diagram and in most further diagrams, where they are obvious, in the interest of not cluttering the diagrams.)

This is essentially a unique ground state on a plane or with open boundary conditions. We shall see later what happens when we have periodic boundary conditions which is equivalent to putting the model on a torus. But before that, let us see how to create excitations over the ground state.

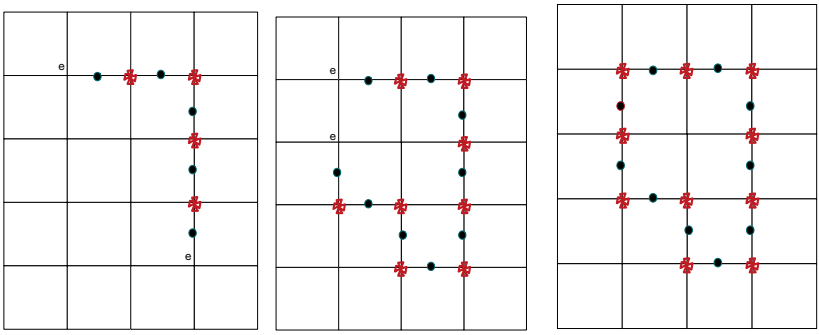


Figure 16.20: (a) A pair of electric charges created by flipping the x-component of the spin by σ_z along a string. Note that the vertices in between the spins all have spins flipped on two links connected to them. (b) Moving one of the charges by flipping more spins and increasing the length of the string (c) Annihilating the charges by closing the string. This configuration has zero vorticity and commutes with the Hamiltonian.

16.3.2 Excitations over the ground state and fusion rules

There are two types of excitations that we can create on the ground state - one by applying σ_x and the other by applying σ_z. Another way of thinking about excitations is to note that both A_v and B_p acting on the ground state have eigenvalues $+1$. One can make excitations if any of these values for some vertex or some plaquette becomes -1. First, let us try to make the vorticity in a given plaquette become -1. To do this, we need to flip the z-component of the spin on one link, which can be done by applying σ_x^i on the i^{th} link. So $\sigma_x^i|\Psi\rangle$ flips $\uparrow$ to $\downarrow$ on the i^{th} spin. However, this gives the value for B_p to be -1 on two plaquettes, since the link is common to two plaquettes. The energy of this excitation is clearly $2J_m + 2J_m = 4J_m$ since the sign change of a single plaquette costs $2J_m$. These excitations are shown in Fig.16.18 where 2 e-excitations have been created on neighbouring vertices and 2 m-excitations in neighbouring plaquettes.

Note that the the pairs of excitations can be moved away from one another at no cost in energy. This is most easily seen diagrammatically. The string $\Pi_{j\in t'}\sigma_x^j$ between the two magnetic excitations (or monopoles) changes the z-component on all the links between the end points as shown in Fig.16.19(a), but essentially all the intermediate plaquettes have $B_p = +1$. If we consider drawing a line between the two monopoles, note that this line (or contour or string) is defined on the dual lattice.

Similarly, $A_v = -1$ if the x-component of the spin on one of the links gets flipped. This can be done by applying σ_z to the ground state. But here again, the change of the x-component of the spin on a link affects two vertices, and hence creates a pair of electric excitations with energy $2J_e + 2J_e = 4J_e$. Once

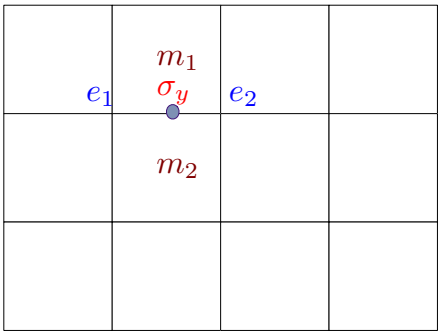


Figure 16.21: A pair of monopoles and a pair of electric charges created by flipping both the x-component and z-component of the spin by σ_y. This is equivalent to a pair of ϵ excitations.

again, the two members of the pair of excitations can be moved away from one another at no cost in energy as is shown in Figs.16.20(a,b) by applying a string of operators, $\Pi_{j\in t}\sigma_z^j$ between the two electric excitations. Here if we draw a line between the e-excitations, the line or string is defined on the lattice. The string changes the eigenvalues of the x components of all the spins belonging to the string. However, except at the end points of the string, every vertex will have two spins flipped and will continue to have $A_v = +1$. It is only at the two ends of the string, that the vertices will have $A_v = -1$. For instance in Fig.16.20(a), we have flipped the x-component of 5 spins. However, we have created only two electric excitations.

We can even make closed string loops that create, move and annihilate the electric or magnetic charges as shown in Figs.16.19(b) and 16.20(c). These loops have z or x vorticity $=+1$ - i.e., they commute with all A_v and B_p. In other words, they commute with the Hamiltonian and can be thought of as symmetries, and the existence of these symmetries is the reason for the exact solvability of the model. The ground state can also be thought of as a superpositions of all these loops.

Are there any other kind of excitations in the toric code? One might think that by applying σ_y on the ground state, we may get new excitations. In fact, it turns out that there is a new excitation, which is a composite of the electric and magnetic excitations, which is called the ϵ excitation -

$$|\epsilon\rangle = \sigma_z^i \sigma_x^i |\Psi_0\rangle = i\sigma_y^i |\psi_0\rangle. \tag{16.49}$$

Essentially, σ_y^i acting on a spin flips both its σ_x and σ_z components. So by applying it on a link as shown in Fig.16.21, it creates a pair of electric and magnetic excitations, or a pair of $\epsilon = e \times m$ excitations. There are no further excitations that can be created. So the particle content of the model is given by I (no particles), m-particles, e-particles and $\epsilon = e \times m$-particles.

Now let us see what happens if we apply two σ operators on the same plaquette. We know that σ_x^j applied on a state creates a pair of excitations (m- particles) on two adjoining plaquettes. But if we apply σ_x^i twice, we do not get two pairs of excitations, because $(\sigma_x^j)^2 = I$ and I operating on a state does not give rise to any excitation. So one cannot have more than one m-particles in each plaquette. This leads us to what are called fusion rules. We find that

$$e \times e = I, \tag{16.50}$$
$$m \times m = I, \tag{16.51}$$
$$\epsilon \times \epsilon = I\ . \tag{16.52}$$

Further we had already seen that the first line of the following set of equations are true and it is not hard to check that the others are true as well -

$$e \times m = \epsilon, \tag{16.53}$$
$$e \times \epsilon = m, \tag{16.54}$$
$$m \times \epsilon = e\ . \tag{16.55}$$

We had earlier seen that we can define string operators to move particles away from one another and even annihilate them by forming closed loops. We said that all these closed loops formed by creating, moving and annihilating particles, cost no energy and commute with the Hamiltonian. They are products of A_v's or B_p's and can be thought of as trivial symmetries of the Hamiltonian because they map the Hamiltonian onto itself. The ground state is unique and a linear combination of all these vortex-free states.

16.3.3 Toric code on a torus and topological degeneracy

But a new element is introduced if we have periodic boundary conditions or equivalently, consider the toric code model on a torus. In this case, there are two other independent operators, that we can define, which are not the products of the A_v and B_p operators of the Hamiltonian, and which can take the values $+1$ and -1. We can write them as

$$W_{1\gamma_1} = \prod_{j\in\gamma_1} \sigma_z^j, \tag{16.56}$$
$$W_{2\gamma_2} = \prod_{j\in\gamma_2} \sigma_z^j, \tag{16.57}$$

where γ_1 and γ_2 are paths which go from one edge of the torus to the other in the two orthogonal directions - for definiteness, let us assume that the loop γ_1 is in the vertical direction and the the loop γ_2 is in the horizontal direction, as shown in Fig.16.22. They form non-contractible loops. The paths can be moved around by multiplying the W_i's with B_p's, (because they do not change anything since they just give $+1$ on the ground state as shown in Fig.16.23). But there is precisely one non-contractible loop in each direction, which we can

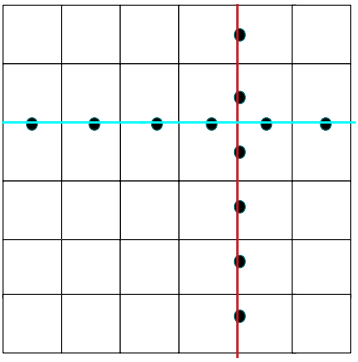


Figure 16.22: Operators $W_{1\gamma_1}$ and $W_{2\gamma_2}$ which are defined by taking the product of the z component of the spins, σ_z^j, along the paths γ_1 (red) going from top to bottom and γ_2 (cyan) going from left to right

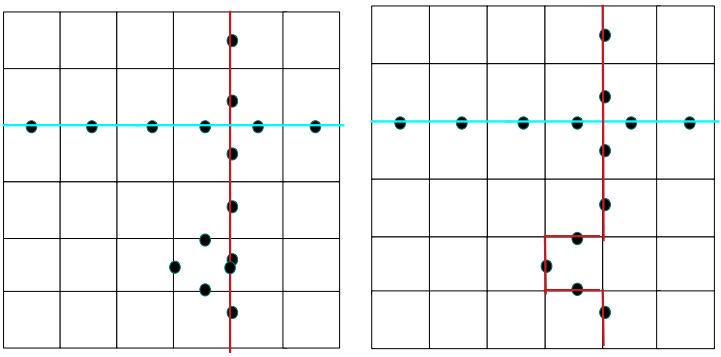


Figure 16.23: The operator $W_{1\gamma_1}$ can changed by multiplying the operator by B_p. This essentially changes the path γ_1, which is now no longer the straight vertical path.

take to be the shortest path. It is easy to see that $W_1^2 = W_2^2 = 1$ which says that W_1, W_2 have eigenvalues ± 1. They are symmetries, because they commute with the Hamiltonian. They also commute with one another and hence, they give rise to a four-fold degeneracy of the ground state

$$|W_1, W_2\rangle_{gs} = |1,1\rangle, \quad |1,-1\rangle, \quad |-1,1\rangle, \quad |-1,-1\rangle, \tag{16.58}$$

because each of the W_i can take values ± 1.

Now let us show that this degeneracy is topological and there is no local operator that can cause transitions between these four ground states.

We shall first see if there are any operators that connect the degenerate states in the ground state manifold. We have already defined W_1 and W_2. Let

us now define also

$$W_{3\gamma_1} = \prod_{j\in\gamma_1} \sigma_x^j, \tag{16.59}$$

$$W_{4\gamma_2} = \prod_{j\in\gamma_2} \sigma_x^j, \tag{16.60}$$

where $W_3^2 = W_4^2 = 1$. These operators also commute with the Hamiltonian, but they do not increase the degeneracy, because as we shall see below, they do not commute with the earlier operators. There are only 4 mutually commuting operators that commute with the Hamiltonian. Clearly, $[W_1, W_2] = 0$ and $[W_3, W_4] = 0$. It is also easy to see that if both the loops are in the vertical or horizontal direction, they will commute, since they can always be displaced, i.e., we see that $[W_1, W_3] = 0$ and $[W_2, W_4] = 0$. However, this is not true when we consider W_1 and W_4. With even the simplest choice of path, they must have at least one spin in common, since one of the paths is vertical and the other horizontal. (More complicated paths will also always give odd number of common spins). So if this common spin is at location 0, then it is the anticommutator of the two operators which vanishes - $\{W_1, W_4\} = \{\sigma^z_{(0)}, \sigma^x_{(0)}\} = 0$ (and similarly $\{W_2, W_3\} = 0$). So it is as if we can define $W_1 = \sigma_1^z, W_2 = \sigma_2^z, W_3 = \sigma_2^x, W_4 = \sigma_1^x$, so that the W_i's can be represented as Pauli matrices. Thus, W_3 and W_4 can change the eigenvalues of W_1 and W_2 by acting on the spins one at a time.

Now how do we confirm that the ground state degeneracy is a topological degeneracy and is topologically protected? The idea is that no local operator allows transitions between the different ground states. Suppose Ω is a local operator -i.e., it is of the form

$$\Omega \equiv \sigma_i^\alpha \sigma_j^\beta \sigma_k^\delta \ldots, \tag{16.61}$$

where the links i, j, k are nearby in the sense that the maximum distance is small compared to the thermodynamic limit N. Then, we can always ensure that

$$[\Omega, W_i] = 0. \tag{16.62}$$

This means that Ω commutes with both σ_i^x and σ_i^z - i.e., with both Pauli matrices at a given site. This means that it has to be proportional to the identity. So it cannot directly cause any transition between different ground states. It can only lead to transitions if it can cause indirect transitions which will take it out of the ground state manifold. This would cost an energy E_0 which is the energetic distance to the next state. Moreover, since the operator is local, it would have to be applied N times since the $W_{3,4}$ operators change the state of 1 link at a time. So if the relevant matrix element of Ω is ω, we see that the total transition amplitude is of $O((\omega/E_0)^N)$ which goes to zero in the thermodynamic limit as long as $\omega/E_0 < 1$. In other words, no local operators can cause transitions between the degenerate states. One needs a non-local

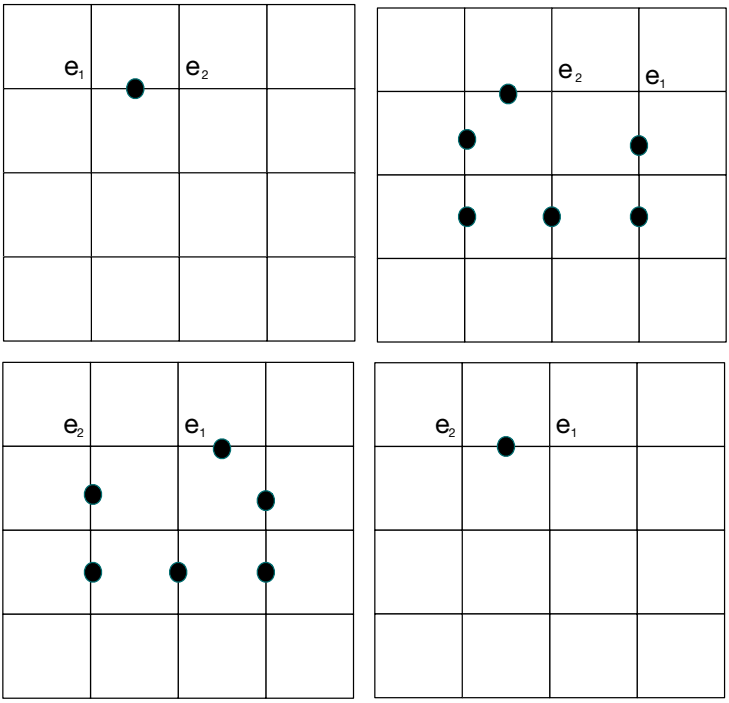


Figure 16.24: The exchange of the charges e_1 and e_2 by operating with the string operator $\Pi_j \sigma_j^z$. In the last step, we have simply removed a closed string loop of vorticity zero.

measurement to distinguish between the four degenerate ground states. This is why the ground state is said to have topological order, and this is what makes it relevant for quantum computation.

16.3.4 Statistics and braiding properties of the excitations

Finally, we want to understand the statistics and braiding properties of the excitations. Let us first look at the statistics of e-particles. $\sigma_i^z|\zeta\rangle$, where $|\zeta\rangle$ is the ground state, creates two e-particles in the adjacent vertices to the edge i. A string $\Pi_j \sigma_j^z$ can separate the two excitations. As shown in Fig.16.24, the two particles can even be exchanged by applying the string operators. But since all σ^z's commute with one another, whichever way we move them, we do not get any phase. Thus, the e-excitations are bosons. Similarly, it is easy to argue that all m-excitations are also bosons.

But now let us consider the mutual statistics between e and m particles. We first create pairs of e and m excitations at sites i and j by applying $\sigma_i^z \sigma_j^x|\zeta\rangle$. We then separate the excitations by applying string operators as shown in

Fig.16.25(a) and then we move an e-particle all around an m-particle, as shown in the series of figures in Fig.16.25(b), Fig.16.25(c) and Fig.16.25(d). It is clear that there is one site at which the σ^x has to be taken beyond a σ^z spin. This anti-commutation gives rise to a minus sign. The closed loop can be removed and the net effect of the process is that

$$\sigma_i^z \sigma_j^x |\zeta\rangle \rightarrow -\sigma_i^x \sigma_j^x |\zeta\rangle \ . \tag{16.63}$$

Now let R_{em} be the operator that exchanges the e and m particles. We have found that the wavefunction for the creation of the two excitations,

$$\psi(\mathbf{r}_e, \mathbf{r}_m) \rightarrow R_{em}^2 \psi(\mathbf{r}_e, \mathbf{r}_m) = -\psi(\mathbf{r}_e, \mathbf{r}_m), \tag{16.64}$$

since taking one particle completely around another is equivalent to two exchanges. Hence $R_{em} = e^{\pm i\pi/2} = \pm i$, which means that the e and m particles have mutual anyon statistics.

We have already seen that $R_{ee} = R_{mm} = 1$. The e and m particles are bosons under exchange. What about the ϵ particles? It is clear that if we take one ϵ particle completely around another, it is equivalent to taking an electron-monopole pair completely around another electron-monopole pair (since $\sigma_y |g.s\rangle \sim \sigma_x \sigma_z |g.s\rangle$). In this case, the phase we expect to get is $+1$ since there will be two negative signs coming from anti commuting a σ_x through a σ_z and vice-versa, so two anti-commutations altogether. But this is not enough to tell us whether a single exchange gives a $+1$ or a -1. But the fact that taking the m particle around the e particle gives rise to a -1 can be interpreted as getting a negative sign when the ϵ particle is rotated through 2π. This is the signature of a fermion. It is a 'spinor' and requires a rotation through 4π to get back to itself. Hence, the ϵ particle is a fermion and we conclude that $R_{\epsilon\epsilon} = -1$.

So now we have all the fusion and braiding rules for the excitations of the toric code. The particle content of the model is given by I, e, m, ϵ. The fusion rules are

$$\begin{aligned} &e \times e = I, \quad m \times m = I, \quad \epsilon \times \epsilon = I, \\ &e \times m = \epsilon, \quad e \times \epsilon = m, \quad m \times \epsilon = e. \end{aligned} \tag{16.65}$$

and the braiding rules are

$$R_{ee} = R_{mm} = 1, \quad R_{\epsilon\epsilon} = -1, \quad R_{em} = i. \tag{16.66}$$

Hence, this model has mutual (abelian) anyonic statistics. We will leave this model here and now go on to study a model which has non-abelian anyon statistics.

16.4 Non-abelian anyons

In this section, we will study a model [12] which has non-abelian anyon excitations. Let me start with a brief explanation of why non-abelian anyons [13, 14]

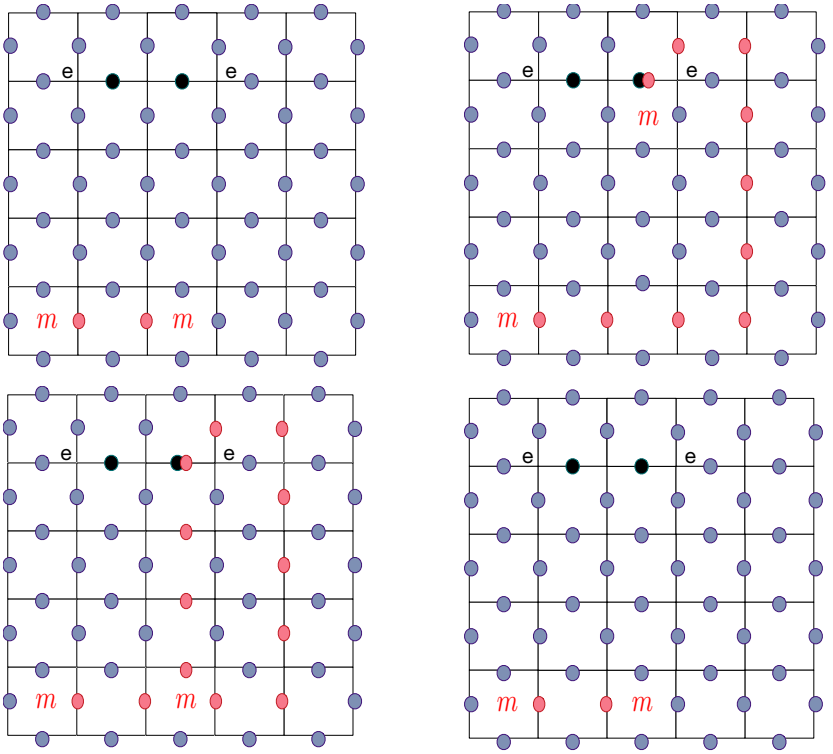


Figure 16.25: The sequence of steps in taking an m charge around an e charge by operating with the string operator $\Pi_j \sigma_j^x$. In the second step, we note that a σ_x has to go past a σ_z, leading to a negative sign. In going from the third to the last figure, a closed loop of σ_x operators has been removed, returning the configuration to the original configuration in the first figure.

are of interest today, other than being an exotic form of exchange statistics. The reason is that they are expected to be relevant to quantum computation [15]. Quantum computation requires the possibility of storing quantum information. This needs a 'protected' portion of the Hilbert space which will not be disturbed by noise, temperature, etc, as long as the length scales are below the gap which separates the 'protected' states from the rest of the states. This protection may be due to some symmetry or even due to topology, which, in a sense acts like a robust symmetry, since it cannot be easily destroyed. This leads us to the idea of topological order, which exists if there is a degeneracy due to topology. For instance, in the earlier section, we studied the toric code, which had only abelian (mutual) anyons and no degeneracy on the plane. But the toric code on a torus has a four-fold degeneracy which was topological and could be used for quantum computation.

In general, in models with non-abelian anyons, we will find that the ground state is degenerate, and if the ground state is separated from all other states by a gap, then the ground state has topological order and can be used for quantum computation.

With this motivation, let us now see [17] what one means by non-abelian excitations. Suppose we have N degenerate states represented by ψ_α , $\alpha = 1 \dots N$, representing N particles. The idea is that if we now exchange particles 1 and 2, it can now do more than just give a phase. It can rotate the state to another wave-function in the same degenerate space. In other words, the column vector $\psi = (\psi_1 \psi_2 \dots \psi_N)^T \to \psi' = U\psi$ where $U \equiv U_{\alpha\beta}$ is an $N \times N$ unitary matrix. If we now exchange two other particles, say 2 and 3, it could lead to $\psi \to \psi' = V\psi$ with V another unitary $N \times N$ matrix. If U and V do not commute, which, in general, they will not, the particles are said to have non-abelian statistics. Clearly, to have non-abelian statistics, we need to have at least 2 degenerate states, since otherwise, U and V are just phases and commute and we have abelian anyons.

One could think about generalizing the physical model of any anyon - a charge orbiting around a flux - to the non-abelian case. In this case, the non-abelian charge would be a vector $|q_i\rangle = \begin{pmatrix} q_1 & q_2 & . & . & q_N \end{pmatrix}$ moving around a non-abelian flux and returning to its original position, but in the process, instead of just acquiring phase factors,

$$|q_i\rangle \to |q_i'\rangle = \sum_j U_{ij} |q_j\rangle \ , \tag{16.67}$$

where U_{ij} is the non-abelian flux matrix. But unlike the abelian flux which was path independent, the non-abelian counterpart U is path dependent. It depends on where the path begins and ends as well as the contour. If we think of U as a matrix belonging to the gauge group $U(N)$ or $SU(N)$, it means that the transport of charge around a flux is gauge dependent because U depends on the choice of gauge. It is only the eigenvalues of U (also called conjugacy class of the flux in group G) which is gauge independent. So for the non-abelian anyons, the physical picture does not help in simplifying or understanding the model and it is better to deal with the more abstract picture.

The main ingredients for a theory of non-abelian anyons are the following –

(1) We need a list of types of particles in the model.
(2) We need fusion rules – rules for fusing two constituents into one and also for splitting a particle into two constituents, which is its inverse.
(3) We need rules for braiding two particles (equivalently exchanging two particles).

So let us now start with an abstract model. We first need a list of particles with their charges. Note that here by charges, we mean topological charges. In a condensed matter system, one can have quasi-particle excitations which are local or which are topological. For instance, in the toric code model that we studied in the last section, a spin operator could be applied locally (at a

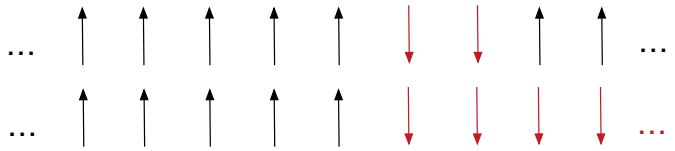


Figure 16.26: The first denotes a local excitation, since it can be removed by flipping two spins locally, whereas the second one is a topological quasi-particle since it is protected by the boundary conditions which are fixed.

given position) to create a pair of excitations (e-type or m type particles by applying σ_x or σ_z). So a pair of excitations is a local or trivial type of particle or equivalent to the identity. But individually, the e or m particles carry a Z_2 charge which is topological. Another example is that of spin flip excitations in an Ising model as shown in Figs. 16.26(a) and 16.26(b). A single spin-flip (or even a few spin flips) is a local excitation and is equivalent to the identity as far as particle type is considered. But a domain wall is a topological excitation that is protected by the boundary conditions and cannot be removed by local perturbations. So let us list the topological particles in our system as $a, b, c, \ldots$.

Next, we need to specify the fusion rules for these particles. The fusion algebra is defined as

$$a \times b = \sum_c N_{ab}^c c, \tag{16.68}$$

where $N_{ab}^c \geq 1$. This equation simply means that if $N_{ab}^c = 0$,then the particle c is not obtained and if $N_{ab}^c = 1$, then the particle c is obtained as a fusion product and if $N_{ab}^c \rangle 1$, then c can be obtained in N_{ab}^c ways. Here, a, b, c are just labels of the different kinds of particles. For abelian anyons, an anyon with statistics parameter α_1 will fuse with an anyon with statistics parameter α_2 to yield a specific anyon with statistics parameter α_3 and $N_{\alpha_1\alpha_2}^{\alpha_3} = 1$ and is otherwise zero. But this is not true for non-abelian anyons. The fusing of two anyons could lead to different types of particles with different probabilities. In that sense, fusion of particles is like a measurement. Given two abelian anyons a and b, their fusion is well-defined and leads to a unique answer. But for non-abelian anyons, it does not lead to a single answer. The integers N_{ab}^c define the probabilities of the outcome. The simplest non-abelian model will have $N_{ab}^c \neq 0$ for at least two distinct values of c. The Hilbert space of two (or more) dimensions formed from these distinct outcomes of fusing two particles is known as the topological Hilbert space of the pair of non-abelian anyons.

Now, let us consider what happens when we have many non-abelian anyons. For abelian anyons, each time you bring in a new particle, the fusion rules give you only one possible outcome. But for non-abelian anyons, since even two anyons can have more than one outcome, a third anyon can fuse with either of the outcomes and again give rise to more possible outcomes. So one can get many fusion paths, since fusion is not unique. Since sometimes,

different paths can lead to the same outcomes, there are consistency conditions that need to be satisfied called pentagon equations. Once we also include braiding matrices in the game, there are also consistency conditions called hexagon equations which need to be satisfied.

However, instead of going further ahead with the abstract analysis, we will now change gears and study a concrete model with non-abelian excitations. This is the Kitaev one-dimensional toy model with unpaired Majorana fermions. We will explicitly show that these Majorana fermions obey non-abelian statistics under exchange.

16.4.1 Kitaev model in one dimension

The Hamiltonian for the Kitaev model in one dimension is given by [12]

$$H = -\mu \sum_{x=1}^{N} n_x - \sum_{x=1}^{N-1} (t c_x^\dagger c_{x+1} + \Delta c_x c_{x+1} + h.c.), \tag{16.69}$$

where c_x represents spinless fermions on site x, t is the amplitude of hopping to nearest neighbour sites, and Δ is the superconducting parameter and denotes p-wave pairing - p wave because the electrons are of the same kind (spinless or equivalently same projection of spin) - on nearest neighbour sites are paired. μ is the chemical potential and $n_x = c_x^\dagger c_x$ is the number operator, so that $N = \sum_x n_x$.

Now, let us rewrite the Hamiltonian in terms of new operators called Majorana operators.

$$\begin{aligned} c_x &= \frac{1}{2}(\gamma_{A,x} + i\gamma_{B,x}), \\ c_x^\dagger &= \frac{1}{2}(\gamma_{A,x} - i\gamma_{B,x}) \ . \end{aligned} \tag{16.70}$$

This implies that $\gamma_{A,x} = c_x + c_x^\dagger$ and $\gamma_{B,x} = i(c_x - c_x^\dagger)$ are hermitian (self-conjugate) operators. This is the definition of Majorana operators.

Now, let us look at some properties of these Majorana modes. We can check that they are fermions, in the sense that they anti-commute. More precisely, they satisfy the algebra given by

$$\{\gamma_a, \gamma_b\} = \delta_{ab}, \quad \gamma_a^2 = \gamma_b^2 = 1 \ , \tag{16.71}$$

whereas genuine fermions satisfy

$$\{c_a, c_b^\dagger\} = \delta_{ab}, \{c_a, c_b\} = 0, \quad c_a^2 = 0 \ . \tag{16.72}$$

Pairs of Majorana fermions (γ_A and γ_B) can be combined to form genuine fermions which can form a single 2 level system, depending on whether the fermion state is occupied or unoccupied. The next step is to consider what

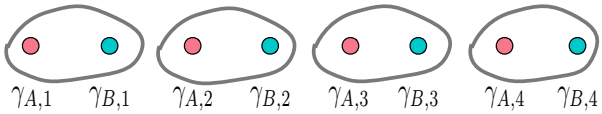


Figure 16.27: Here, the bonds are between Majorana modes at the same site. The ground state is unique and the end Majorana modes do not play any special role.

happens when we have $2N$ Majorana fermions. We can pair them up to make N ordinary fermions -

$$q_x = \frac{1}{2}(\gamma_{A,x} + i\gamma_{B,x}), \quad q_x^\dagger = \frac{1}{2}(\gamma_{A,x} - i\gamma_{B,x}), \tag{16.73}$$

(the same equations used to split the fermions into Majorana modes given in Eq.16.70) with the number operators at each site x given as $N_x = q_x^\dagger q_x = 0, 1$. This gives a 2^N dimensional Fock space.

Why is it interesting to rewrite fermions in terms of pairs of Majorana fermions? Naively, this seems to be something which can always be done, and does not lead to anything new. But if a pair of Majoranas can be spatially separated, then the fermion made from them is delocalized. It is hence, protected from local changes that affect only one of them and hence protected from decoherence. This is why Majorana modes are expected to be relevant in quantum computation.

Now, let us get back to the Kitaev model. To understand the physics in a simple way, let us consider two simple limits, where the Hamiltonian becomes particularly simple. First, consider the case when $\mu = 0$ and $t = \Delta$. Here, we get

$$H = -it \sum_{x=1}^{N-1} \gamma_{B,x}\gamma_{A,x+1} \ . \tag{16.74}$$

In the other limit, we take $\mu < 0$ and $t = \Delta = 0$ and get

$$H = -\frac{\mu}{2} \sum_{x=1}^{N} (1 + i\gamma_{B,x}\gamma_{A,x}) \ . \tag{16.75}$$

What do these two limits mean?

We first analyze the second case. Here, the fermion at each site is simply broken up into two Majorana fermions and the μ term simply couples them as shown in Fig.16.27. In this case, there is a unique ground state corresponding to the vacuum state with no fermions. Adding a fermion to the system costs an energy μ, so the system is gapped.

The first limit, on the other hand, couples Majorana modes at adjacent sites. In terms of new fermions $d_x = (\gamma_{B,x} + i\gamma_{A,x+1})/2$, the Hamiltonian can

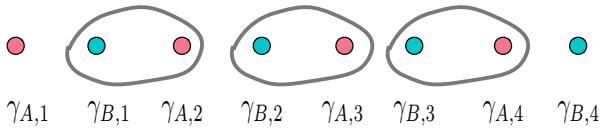


Figure 16.28: Here, the bonds are between Majorana modes on adjacent sites. There are unpaired Majorana modes at the two ends. The ground state is doubly degenerate depending on whether the fermion state formed from the unpaired Majorana modes is occupied or unoccupied.

be rewritten as

$$H = 2t \sum_{x=1}^{N-1} d_x^\dagger d_x. \tag{16.76}$$

This clearly shows that the system has a gap and the cost of adding a fermion to the system is $2t$. But the interesting point to note is that the Hamiltonian is completely independent of the two Majorana modes $\gamma_{A,1}$ and $\gamma_{B,N}$ as shown in Fig.16.28, at the two ends of the wire. These two Majorana modes can be combined to form a fermion as

$$c_M = \frac{1}{2}(\gamma_{A,1} + i\gamma_{B,N}) \ . \tag{16.77}$$

But this is a highly non-local fermion, since $\gamma_{A,1}$ and $\gamma_{B,N}$ are localized at opposite ends of the chain. Moreover, since this fermion is absent from the Hamiltonian, the energy is the same whether or not this fermion state is occupied. So the ground state is degenerate. If $|0\rangle$ is the ground state, then $c_M^\dagger|0\rangle$ is also a ground state. Note that $\gamma_a^2 = 1$ implies that there is no Pauli principle for the Majorana modes - in fact, as we saw earlier, there is no notion of occupation number for a single Majorana mode. Number operators only exist for fermions formed from pairs of Majorana modes. Depending on the occupation or not of the zero energy mode of the fermion - i.e, of c_M - there exists an odd or even number of fermions in the ground state referred to as 'fermion parity'. To change the parity, electrons have to be added or removed from the superconductor. This is unlike normal gapped superconductors, (e.g. the second limit), which have a unique ground state with even fermion parity.

In the more general case, [18] when μ, t and Δ are non-zero, the general features of the topologically trivial case with a unique ground state, and the topologically non-trivial case with the Majorana edge states persist. Why do we call them topologically trivial and non trivial in the two cases? Well, in the trivial case there are no edge states and in the non-trivial case, there are edge states. In terms of the bulk properties of the Kitaev chain, one can find the bulk quasiparticle spectrum by going to momentum space and rewriting the

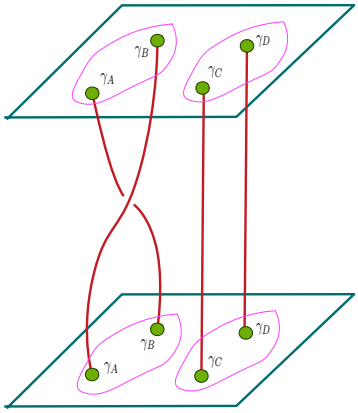


Figure 16.29: Here, we exchange the Majorana modes belonging to the same fermion. This only gives rise to a phase and hence abelian statistics.

Hamiltonian in Eq.16.69 as

$$H = \sum_k \zeta_k c_k^\dagger c_k + \sum_k (\Delta_k c_k c_{-k} + h.c.), \tag{16.78}$$

using $c_x = \frac{1}{\sqrt{N}} \sum_k e^{ikx} c_k$ and $c_x = c_{x+N}$, $\zeta_k = -2t \cos k - \mu$ and $\Delta_k = -2i\Delta \sin k$ We then find the quasi-particle spectrum by going to Nambu space and writing the Bogoliubov-de Gennes Hamiltonian as

$$H = (c_k^\dagger c_{-k}) \begin{pmatrix} \zeta_k & \Delta_k^* \\ \Delta_k & -\zeta_k \end{pmatrix} \begin{pmatrix} c_k^\dagger \\ c_{-k} \end{pmatrix}, \tag{16.79}$$

and finally we get the spectrum $E_k = \sqrt{(\zeta_k^2 + |\Delta_k|^2)}$. Hence, the model is gapped, except when $\mu = -2t$ when $k = k_F = 0$ or when $\mu = +2t$ when $k = k_F = \pm\pi$. The lines $\mu = \pm 2t$ are where the system becomes gapless. For $\mu < 2t$, the system is topological and is adiabatically connected to the first limit with Majorana edge states. For $\mu > 2t$, the system is topologically trivial and is adiabatically connected to the second limit with no edge states.

16.4.2 Statistics of the Majorana modes

Now, let us consider the statistics [19, 20] of the Majorana modes. Let us start with the simplest case where $N = 1$. In this case, there are only two Majorana modes and the braid group only has a single generator τ. As we saw in the first section, τ is the operator that exchanges the Majorana modes A and B (as shown in Fig.16.29) -

$$\gamma_A \to \gamma'_A = \tau^\dagger \gamma_A \tau = e^{i\phi} \gamma_B \ , \tag{16.80}$$

but it does not fix the phase ϕ which is arbitrary. We can choose it to be $+1$. But then the phase of

$$\gamma_B \to \gamma'_B = \tau^\dagger \gamma_B \tau = e^{i\phi}\gamma_A \tag{16.81}$$

is forced to be -1. This is because when there are only 2 Majorana modes, and the system is isolated, the fermion parity is forced to be conserved. The fermion state formed from the two Majorana modes is either occupied or unoccupied and we can check that

$$i\gamma_A\gamma_B = (1 - 2c_M^\dagger c_M). \tag{16.82}$$

So if the right hand side remains unchanged, then $i\gamma_A\gamma_B$ has to remain unchanged, which is only possible, if we choose the phases as shown above, since $\gamma_A\gamma_B = -\gamma_B\gamma_A$. Here, we can choose the exchange operator to be of the form

$$\tau = \frac{1}{\sqrt{2}}(1 + \gamma_A\gamma_B). \tag{16.83}$$

It is easy to check that τ defined this way is unitary and that it actually carries out the exchange by substituting for τ in Eq.16.81. It is also easy to check that τ can be rewritten as $\exp(\pi\gamma_A\gamma_B/4)$. If we write it in terms of the fermion number operator,

$$\tau = e^{i\pi(1-2n)/4}, \quad \text{where} \quad n = c_M^\dagger c_M. \tag{16.84}$$

Clearly, since n does not change, the statistics parameter is abelian and it cannot rotate states in the ground state manifold $(|0\rangle, c_M^\dagger|0\rangle)$.

Now, let us see what happens when $N = 2$. Here, we have 4 Majorana modes $\gamma_i, i = A \ldots D$ which can form 2 normal fermions -

$$\begin{aligned} c_1 &= \tfrac{1}{2}(\gamma_A + i\gamma_B), \quad c_1^\dagger = \tfrac{1}{2}(\gamma_A - i\gamma_B), \\ c_2 &= \tfrac{1}{2}(\gamma_C + i\gamma_D), \quad c_2^\dagger = \tfrac{1}{2}(\gamma_C - i\gamma_D). \end{aligned} \tag{16.85}$$

The degenerate states of the system are given by $|n_1, n_2\rangle = c_1^\dagger c_2^\dagger|0,0\rangle = \{|0,0\rangle, |1,0\rangle, |0,1\rangle, |1,1\rangle\}$. Operator τ_{AB} exchanges the Majoranas A and B keeping C, D unchanged and operator τ_{CD} exchanges the Majoranas C and D keeping A, B unchanged. Similarly, we can define, τ_{AC}, τ_{BD}, etc.

It is now clear that analogous to the $N = 1$ case, if we exchange the two Majorana zero modes from the same fermion, as shown in Fig.16.29, we will only get a phase - i.e.,

$$\begin{aligned} \tau_{AB}|n_1, n_2\rangle &= e^{i\pi(1-2n_1)/4}|n_1, n_2\rangle, \\ \tau_{CD}|n_1, n_2\rangle &= e^{i\pi(1-2n_2)/4}|n_1, n_2\rangle \ . \end{aligned} \tag{16.86}$$

In the first equation, n_2 comes along for a ride and in the second equation, n_1 comes along for a ride. So both these operators are abelian operators. But now,

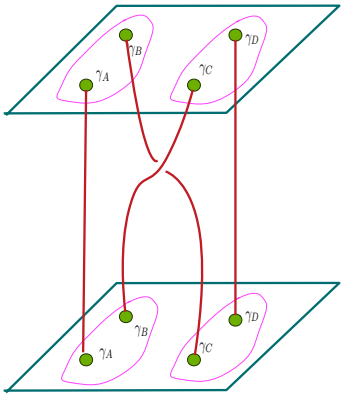


Figure 16.30: Here, we exchange the Majorana modes belonging to two different fermions. This leads to a unitary rotation in the space of degenerate states and hence, to non-abelian statistics.

let us exchange one Majorana from one of the fermions with another Majorana from the other fermion, (as shown in Fig.16.30)

$$\tau_{BC} = \frac{1}{\sqrt{2}}(1 + \gamma_B \gamma_C). \tag{16.87}$$

In terms of the fermions c_1 and c_2, this can be written as

$$\tau_{BC} = \frac{1}{\sqrt{2}}[1 - i(c_1 - c_1^\dagger)(c_2 + c_2^\dagger)]. \tag{16.88}$$

Now acting this on $|n_1, n_2\rangle$ does not lead to a phase. Instead, it leads to a rotation in the space of degenerate states given by

$$\tau_{BC}|n_1, n_2\rangle = \frac{1}{\sqrt{2}}[|n_1, n_2\rangle + i(-1)^{n_1}|1 - n_1, 1 - n_2\rangle] \ . \tag{16.89}$$

If we now consider sequential exchanges, it is clear that different exchanges will not commute with one another - the final state will depend on the order of the operations. This is what is meant by saying that the Majorana particles have non-abelian statistics under exchange.

The derivation of the non-abelian statistics is not dependent on the details of how the exchange between the particles is carried out, and hence it cannot be changed by disorder or local details. It is topologically stable.

To understand multiple non-abelian anyons, as we already mentioned, we need to understand fusion paths, since the fusion rules do not need to unique results. These fusion paths represent a basis of the degenerate ground state manifold, and are most conveniently studied in terms of conformal blocks of the appropriate conformal field theory [3]. But that is beyond the scope of these lectures and we will stop here.

16.5 Conclusion

Let me conclude by repeating the main message of these lectures - understanding the notion of anyons and non-abelian anyons is an exciting field today. The study of these excitations could lead to an understanding of concepts like decoherence and entanglement which are relevant in quantum computation. Work on non-abelian states, in general, is still in its infancy. For young researchers, hence, this should be a useful and relevant topic of study at the crossroads of condensed matter physics and quantum information. For more information and references, there are many recent reviews [2, 3, 15] available on the net.

References

[1] J. M. Leinaas and J. Myrrheim, Il Nuovo Cimento, **37**, 1 (1977).

[2] See for instance, review by A. Stern, 'Anyons and the quantum Hall effect - a pedagogical review', cond-mat/0711.4697.

[3] C. Nayak, S. H. Simon, M. Freedman and S. D. Sarma, Rev. Mod. Phys. **80** 1083 (2008).

[4] See review by S. Rao, 'An Anyon Primer', hep-th/9209066 based on Lectures delivered at the VIII SERC School in High Energy Physics, 30 Dec. '91 - 18 Jan '92, held at Physical Research Laboratory, Ahmedabad, and at the I SERC School in Statistical Mechanics, Feb '94, held at Puri, published in 'Models and Techniques of Statistical Physics', edited by S. M. Bhattacharjee (Narosa Publications).

[5] 'Anyons' Alberto Lerda, Lecture notes in Physics Monographs, Series 14, Springer-Verlag publications, 1992.

[6] 'Fractional statistics and quantum theory', A. Khare, World Scientific, 1998.

[7] See for instance, 'Fractional statistics and Anyon Superconductivity', edited by F. Wilczek, Series on directions in condensed matter physics, World Scientific Publications, October 1990.

[8] F. Wilczek, Phys. Rev. Letts. **48**, 1144 (1982); ibid, **49**, 957 (1982).

[9] A. Kitaev, 'Fault tolerant quantum computation by anyons', Ann. Phys. **321**, 2 (2003).

[10] A. Kitaev and C. Laumann, 'Topological phases and quantum computation', 0904.2771, lectures given by A. Kitaev at the 2008 Les Houches summer school 'Exact methods in low-dimensional physics and quantum computing'.

[11] M. Burello, 'Topological order and quantum computation: Toric code' , talk available on the net.

[12] A. Y. Kitaev, 'Unpaired Majorana fermions in quantum wires', Physics-Uspekhi, **44**, 131 (2001).

[13] G. Moore and N. Read, Nucl. Phys. **B360**, 362 (1991).

[14] R. S. K. Mong *et al*, Phys. Rev. **X4**, 011036 (2014).

[15] J. Alicea, 'New directions in the pursuit of Majorana fermions in solid state systems', cond-mat/1202.1293, Rep. Prog. Phys. **75**, 076501 (2012).

[16] X. G. Wen, Int. J. Mod. Phys. **B**4, 239 (1990); X. G. Wen and Q. Niu, Phys. Rev.**B** 41, 9377 (1990).

[17] J. Preskill, 'Lecture notes for Physics 219: Quantum computation' http://www.theory.caltech.edu/ preskill/ph219/topological.pdf

[18] M. Leinjse and K. Flensberg, cond-mat/1206.1736, Semicond. Sci. Technol. **27**, 124003 (2012).

[19] D. A. Ivanov, Phys. Rev. Letts. **86**, 268 (2001).

[20] E. Berg, 'Anyonic statistics : A new paradigm for non-abelian statistics', talk available on the net.

17

An introduction to Quantum Spin Liquids

Subhro Bhattacharjee

Quantum spin liquids represent phases of condensed matter that fall beyond the paradigm of Landau's symmetry based classification. They, instead, are characterized by the presence of subtle patterns of long-range many-body quantum entanglement. An ever growing list of experiments suggests that understanding such phases of matter forms a crucial step towards the development of a new and general framework of condensed matter systems. We provide an introduction to such physics in the context of quantum spin liquids that would be relevant to frustrated quantum magnets. We take two examples, (1) the two dimensional Z_2 quantum spin liquid in Kitaev's Toric code model, and (2) the three dimensional $U(1)$ quantum spin liquid in the XXZ pyrochlore system, both for spin-$\frac{1}{2}$, to explain some of the inherent properties of quantum spin liquids, as we know them. The aim is to contrast these properties with those of conventional phases like the magnetically ordered ones. These differences range from novel excitations such as mutual semions in the form of Ising electric and magnetic charges in Toric code to emergent photons and fractionalized spin-$\frac{1}{2}$ excitations in XXZ pyrochlores. These two examples have been chosen to bring out the differences clearly, without going into the general structure of emergent gauge theories in quantum spin liquids. However, all the principles introduced here have very general applications.

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_17

17.1 Introduction

The theoretical framework of modern condensed matter physics, developed over the last hundred years or so, has been widely successful in describing phases of many-body systems, both classical and quantum. Most of the technological advancements that we see around us owe their origin to our growing understanding of condensed matter phases like metals, semiconductors, superconductors, magnets, *etc.* The framework of understanding many of these phases, the conventional condensed matter theory, heavily rests on two major organizing principles that can be traced back to the pioneering works of Landau, among others. These two organizing principles are: (1) the idea that condensed matter phases can be classified on the basis of *spontaneous symmetry breaking*, and (2) the principle of *adiabatic continuity* connecting interacting systems to suitable non-interacting or mean-field descriptions [1–3].

Ever since the discovery of quantum Hall effect in 1980s [4, 5], the above principles have been repeatedly challenged in terms of their applicability in their current forms. So, in the last three decades, these ideas had to be suitably generalized and/or adapted to incorporate description of condensed matter phases that do not directly fall within the purview of the above conventional understanding. Theoretical work, often spurred by fantastic experiments and rapid developments in material sciences, have opened up to us an ever expanding array of condensed matter phases whose description require new mathematical techniques and more importantly new ideas and outlook towards studying condensed matter systems. These systems have been variously dubbed as *quantum ordered* [6] as their characterization is often based on features of quantum entanglement, which do not have a classical counterpart.

In this article, I have presented a brief review of a class of systems that have emerged as canonical examples of quantum ordered phases by providing important insights into some of their currently understood features. These systems consist of a bunch of interacting quantum spins, each sitting on the sites of some lattice in two or three spatial dimensions (throughout this article, I have not described one dimensional systems, which form an interesting class on their own). These spin systems describe the insulating phase of the so called *Mott insulators* where the electrons are localized to the lattice sites by repulsive electron-electron interactions. While the electronic charge degrees of freedom are localized, the electron spins sitting at the lattice sites interact with each other, and these interactions dictate the magnetic properties of such systems. Several experimental examples of Mott insulators are well known. While many of them order magnetically at low temperatures, a class remains in the disordered paramagnetic phase to the lowest experimentally observable temperature [7, 8]. This raises an interesting question — can one have a quantum paramagnetic ground state in such an interacting spin system which does not break any symmetry at zero temperature? Here we shall see that such a quantum paramagnet, dubbed a quantum spin liquid, is indeed possible and can have exotic properties like presence of long range quantum entanglement lead-

ing to ground state topological degeneracy, emergent photon-like gauge bosonic excitations and fractionalization of quantum numbers [6–11].

The present article grew out of a set of lectures that I gave at the SERC school on topological condensed matter in Kolkata during December 2015. This is a slightly expanded and suitably edited version of those lectures. The plan for the rest of this article is as follows. I have introduced the basic ingredients of studying many-body systems and spin Hamiltonians in Section 17.2, followed by a brief discussion of the conventional paradigm of solid state physics in Section 17.3 stressing on the ideas of spontaneous symmetry breaking. Then, instead of going into general discussions of quantum ordered phases, I have focussed on three central ideas in relation to quantum ordered phases in general and quantum spin liquids in particular. These are: (1) topological order, (2) fractionalization of quantum numbers, and (3) associated emergent gauge field excitations, as discussed in Section 17.4. I have done this in the context of two well studied microscopic spin systems — the Toric code model [12, 13] in Section 17.5 and XXZ spin models in pyrochlore lattice [14] in Section 17.6. I have used these examples to point out the general structure of the phases and the low energy theories that we encounter in the context of quantum spin liquids. Unfortunately, I have had to refer to original literature and excellent review articles for experiments on candidate quantum spin liquid materials [7,8]. Almost all of the topics covered in this article have been taken from recent research literature and I have tried to provide appropriate references. In that sense, there is no originality regarding the material content of this article except, perhaps, in the presentation and choice of topics.

17.2 Introduction to spin systems

In condensed matter physics, we are interested in the properties of systems consisting of $\sim 10^{23}$ particles. These particles can be fermions, bosons or spins (or even classical particles though here we have exclusively focussed on degrees of freedoms for which a quantum mechanical description is required) which are mutually interacting with each other and the system has certain symmetries. So where do we start to describe such systems? There are essentially three main ingredients to any condensed matter system:

1. **The degrees of freedom:** This is done by specifying the Hilbert space. One of the most important features of Hilbert spaces that is relevant for condensed matter systems, is that they have a tensor product structure, *i.e.*,

$$\mathcal{H} = \bigotimes_i \mathcal{H}_i. \tag{17.1}$$

For example, in a system of electrons (fermions) on a lattice, four states are allowed at each site. These are — (1) no electrons, $|0\rangle$, (2) an up electron, $|\uparrow\rangle$, (3) a down electron, $|\downarrow\rangle$, and (4) two electrons with opposite spins,

$|\uparrow\downarrow\rangle$. The total Hilbert space for the entire lattice is a direct product of four such states at each site.

2. **Symmetry:** The system usually has some symmetry like spin rotation, lattice translation, lattice rotations, time reversal *etc.* These symmetries act in a particular way on the degrees of freedom (states of the Hilbert space) through unitary operations. Since the Hilbert space has a tensor product structure, the symmetry acts locally on the degrees of freedom as a product of local unitary transformations.

3. **Hamiltonian:** Once the degrees of freedom and the symmetries are specified, the time evolution of these degrees of freedom, interacting with each other in a way that is consistent with the symmetries of the system, needs to be specified. This is done by specifying the Hamiltonian. A key feature of the condensed matter Hamiltonians is that they often involve a few degrees of freedom at a time in each term of the Hamiltonian.

It is extremely important to note that the above three features have the idea of an energy scale built into them. For example, as discussed in Appendix A, in spin systems, the degrees of freedom are the electron spins which, within a Hubbard model scenario, arise when the electrons get localized on the lattice sites due to coulomb repulsion between two electrons and it requires energy to overcome this repulsion (see for example Ref. [15]) which is nothing but the charge gap. Below this charge gap, the spins are the right degrees of freedom and the spin Hamiltonian is the right "effective" Hamiltonian to describe the system. Somewhat rarely, symmetries may also emerge. In this sense all the above ingredients are *effective*. There is no way within an effective model to calculate the energy scale upto which the description in terms of a given set of degrees of freedom remains valid. However, we should keep in mind that there are such energy scales often dubbed as the *ultraviolet* (UV) scale of the problem. The problem and the degrees of freedom (just like the spin problem) may look different well above and well below such energy scales. Indeed two quite different high energy systems can give rise to a very similar low energy physics, as can be understood within the framework of *renormalization group* [2]. This we already know from our knowledge of superconductors, magnets, etc. For the low energy effective model, the information of the underlying *high energy* enters through certain numbers (often dimensionfull) like critical temperature. Calculating such critical temperatures are usually very hard. However, calculating universal numbers like the linear dispersion of phase mode in superfluids, are mostly independent of the exact microscopic details.

17.2.1 Spin systems

Consider a lattice in two or three spatial dimensions (as shown in figure 17.1) where each site is occupied by a spin-$\frac{1}{2}$. Let us define spin-$\frac{1}{2}$ operators of the

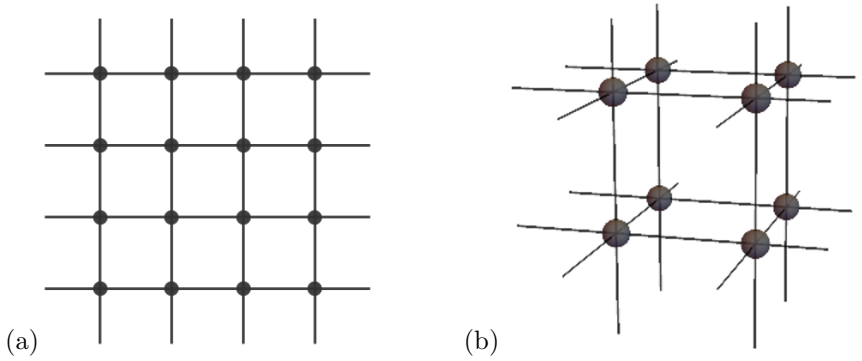


Figure 17.1: The square and the cubic lattices as examples of two and three dimensional lattices.

form

$$s_i^\alpha = \frac{1}{2}\sigma_i^\alpha, \qquad \alpha = x, y, z, \tag{17.2}$$

where i denotes the sites of the lattice and σ^α denotes Pauli matrices. These spins form the degrees of freedom and clearly the Hilbert space has a tensor product structure. Such systems usually have a list of symmetries, *e.g.* lattice symmetries, time reversal, *etc.* that form a group, the *symmetry group*. Generic spin Hamiltonians, invariant under such symmetries, can be written as

$$H = \sum_{ij} J_{ij}^{\alpha\beta} s_i^\alpha s_j^\beta + \sum_{ijkl} J_{ijkl}^{\alpha\beta\gamma\delta} s_i^\alpha s_j^\beta s_k^\gamma s_l^\delta + \cdots, \tag{17.3}$$

where i, j, k, l denote lattice sites and $\alpha, \beta, \gamma, \delta = x, y, z$ denote the spin components. Both $J_{ij}^{\alpha\beta}$ and $J_{ijkl}^{\alpha\beta\gamma\delta}$ are coupling constants. It is important to note that one can, in principle derive the above Hamiltonian from a "more microscopic" electron Hamiltonian (see for example Appendix A)in an effective low energy limit. This will give an estimate of the coupling parameters $J_{ij}^{\alpha\beta}$ and $J_{ijkl}^{\alpha\beta\gamma\delta}$ in terms of those of the electron Hamiltonian [15]. However, in a real material, since the "microscopic" electron Hamiltonian is often extremely complicated, and also due to renormalization effects, such a derivation of low energy coupling constants is only approximate. Here we would not attempt to derive the spin Hamiltonians from electrons (except in Appendix A).Instead we invoke symmetries to write down spin Hamiltonians such as in Eq. 17.3.

The question now is — what is the ground state of this system and what are the low energy excitations about this ground state that govern the outcome of most condensed matter experiments? Generally this is impossible to answer because the ground state is of the form

$$|\psi_g\rangle = \sum_n A_n |\{\sigma^z\}\rangle_n, \tag{17.4}$$

where we have chosen an Ising basis to expand the state. For N spin-$\frac{1}{2}$s, there are 2^N complex numbers A_n, and it is impossible to solve this for systems with $N \sim 10^{23}$. However, it turns out that as long as we ask questions whose answers are universal numbers, we do not need to know all such details. A suitable way to progress in the large N limit, is to approximate the exact ground state in terms of an approximate ground state, *i.e.,*

$$|\psi_g\rangle \approx |\psi_{gs}^{approx}\rangle, \tag{17.5}$$

which captures the *basic features* of the exact ground state. This, however, does not solve the problem as important *basic features* of the system are not known a-priori.

For conventional magnetic systems which order at low temperature and hence described by spontaneous symmetry breaking (see next section), the *basic features* refer to quantifying the presence of broken symmetry. We first review such systems before moving on to quantum spin liquids.

17.3 Example of Spontaneous symmetry breaking in spin systems: Magnetic order

Generally spin systems, described by Hamiltonians like the one in Eq. 17.3, that are found in nature, often undergoes a magnetic ordering transition at low temperature. For example, let us focus on a Hamiltonian of a Heisenberg antiferromagnet on a square lattice which has the particular form

$$H = J \sum_{\langle ij \rangle} \mathbf{s_i} \cdot \mathbf{s_j}, \tag{17.6}$$

where $J > 0$ and $\langle ij \rangle$ denotes sum over nearest neighbours. This particular Hamiltonian can be obtained from Eq. 17.3 by demanding spin-rotation symmetry on the first term and neglecting the rest.

The exact ground state of the above system has a very complicated form of the type given by Eq. 17.4. In the $N \to \infty$ limit, the regime of our interest, it turns out that the exact ground state can be approximated to a great accuracy by an approximate one which has the form

$$|\psi_g^{approx}\rangle = |\uparrow\downarrow \cdots \uparrow\downarrow\rangle, \tag{17.7}$$

where all the up spins belong to one sub-lattice while the down ones belong to the other. Note that, since the Hamiltonian in Eq. 17.6 has spin rotation symmetry, a state which is obtained by globally rotating the above state by any angle, but keeping intact the antiparallel nature, would equally satisfy the criterion of the approximate ground state like the one given in Eq. (17.7) above. This, therefore, is indeed an incredible simplification where we now need to specify two numbers, *i.e.*, the Euler angles denoting the orientation of one

spin to specify the entire motif. Indeed, on calculating the expectation value of any spin using the above approximate ground-state wave-function, we get

$$\langle \psi_g^{approx} | \mathbf{s}_i | \psi_g^{approx} \rangle \sim \pm \mathbf{n}(\theta, \phi) \neq 0, \tag{17.8}$$

where $\pm$ refer to the two different sublattices to which the site i may belong and θ, ϕ are the two Euler angles needed to specify the direction of $\mathbf{n}$. The particular values of θ, ϕ chosen by the system spontaneously results in breaking of the spin-rotation symmetry.

The incredible simplification obtained in the $N \to \infty$ limit is due to the fact that out of all the 2^N numbers in Eq. 17.4, only a particular set becomes important to characterize the *basic features* of the ground state in this limit which encodes the fact that the ground state spontaneously breaks the spin rotation symmetry, as given by Eq. 17.8. Indeed equations like Eq. 17.8 are general to all symmetry broken phases and form the basis of the Curie-Weiss type of mean-field theories. Expectation values like $\mathbf{n}$ quantify the amount of symmetry breaking, or. in other words, quantify the amount of (magnetic) order present in the system. Thus such fields are called order parameter fields.

Having obtained the ground state, the low energy states can be obtained by studying long wavelength fluctuations of the order parameter. In particular, for spontaneously broken continuous symmetry, such low energy modes are guaranteed to have a massless spectrum of Nambu-Goldstone bosonic modes (in the present case, also called spin-waves or magnons). Elementary low energy excitations above the antiferromagnetically ordered ground state are of the form

$$|\mathbf{k}\rangle = \sum_i e^{i\mathbf{k}\cdot\mathbf{r_i}} \tilde{s}_i^- |\psi_g^{approx}\rangle, \tag{17.9}$$

where $\tilde{s}_i^- = s_i^-(s_i^+)$ stands for sublattice of up-spins (down-spins). These Nambu-Goldstone modes have an energy of the form $\epsilon_{\mathbf{k}} \propto |\mathbf{k}|$, where the proportionality constant is a non-universal number. We must note that such modes can be easily gapped out by putting in perturbations that reduce the symmetry that is spontaneously broken to a discrete one [15, 16].

The above ideas form the starting point for order parameter based Landau-Ginzburg type of field theories which captures a large class of phases and phase transitions [1, 2]. However, there are notable exceptions in one and two spatial dimensions like the Berezinskii-Kosterlitz-Thouless phase transition [17–21], where the above framework fails due to strong thermal/quantum fluctuations which invalidates the above arguments. In these cases, fluctuations about the mean value of order-parameter overwhelm the mean value and hence the detailed structure of the fluctuations, *i.e.*, topological defects of the order parameter [20, 21], is needed to be taken into account.

This completes our introduction to the conventional magnetically ordered phases within the framework of conventional paradigm. From now on, we shall focus on examples in which the above arguments either do not apply or are needed to be overhauled to a great extent. By studying such examples in the

context of magnetism, we shall try to understand the nature of the framework, as far as we understand it today, required to describe the *quantum ordered* phases of condensed matter.

17.4 Quantum ordered phases in magnets: Quantum spin liquids

Having outlined the framework for describing condensed matter phases that can be captured within the paradigm of spontaneously symmetry breaking, we now proceed to discuss the basic features of quantum spin liquids which are examples of a quantum ordered phase. Instead of proceeding with a general discussion, we shall focus on the following central features of such quantum spin liquids:

1. Topological order as reflected in presence of long range quantum entanglement of the many-body ground state.

2. Topological ground state degeneracy and non-trivial statistics of the excitations.

3. Fractionalization of microscopic quantum numbers.

4. Emergent gauge fields, like a photon-like gapless excitation, whose gaplessness is robust to symmetry breaking perturbations.

The above features will be exemplified by discussing two well known spin systems — (1) The Toric code model [12, 13], and (2) spin-$\frac{1}{2}$ XXZ pyrochlore antiferromagnets [14] that stabilizes quantum spin-liquid ground states.

In both these systems, the ground state does not have any magnetic order. In particular, the Toric code ground state is a two dimensional $\mathbb{Z}_2$ quantum spin liquid [6, 10] with gapped bosonic and fermionic excitations. The XXZ pyrochlore antiferromagnets, on the other hand, stabilizes a three dimensional $\mathbb{U}(1)$ quantum spin liquid [6, 10] which allows gapped fractionalized $S = 1/2$ excitations and gapless excitations very similar to a photon, *i.e.*, a gauge boson, whose gaplessness (unlike the Nambu-Goldstone mode) is protected by an emergent gauge invariance at low energy.

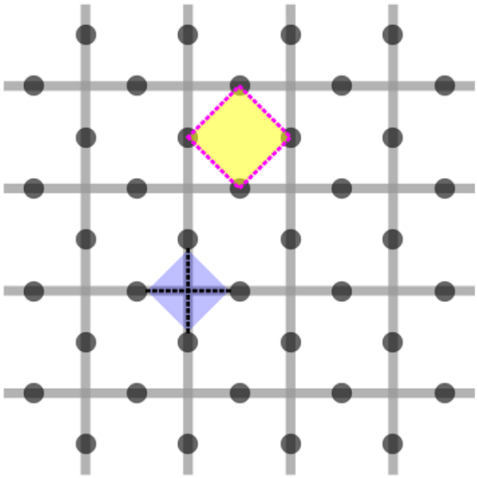


Figure 17.2: **The Toric code model** on a square lattice. The spin-$\frac{1}{2}$s live on the bonds (denoted by filled circles). The darker (lighter) simplex around a site (plaquette) denotes operator $\hat{A}_s$ ($\hat{B}_p$) for the site (plaquette) as defined in the main body of the text.

17.5 Topological Order, Excitations with non-trivial statistics and ground state long range entanglement in Toric code model

17.5.1 The Toric code model

Consider a model of spin-$\frac{1}{2}$s sitting on the bonds of a square lattice, as shown in Fig. 17.2, and interacting with a Hamiltonian [12, 13]:

$$H_{TC} = -\sum_s J_s \hat{A}_s - \sum_{\square} J_p \hat{B}_p, \tag{17.10}$$

with $J_s, J_p > 0\ \forall s, p$ where s and p denote the sites and plaquette of the square lattice respectively, as shown in figure 17.2, and

$$\hat{A}_s = \prod_{i \in s} \sigma_i^x; \qquad \hat{B}_p = \prod_{i \in p} \sigma_i^z. \tag{17.11}$$

Note that this Hamiltonian does not have spin rotation or lattice symmetries in general when the coupling constants differ from site/plaquette to site/plaquette. All symmetries like time reversal that is left can also be broken by adding small perturbations and still the following discussions will hold. We start by noting that

$$A_s^2 = B_p^2 = 1. \tag{17.12}$$

So the eigenvalues of these operators are ± 1. Also, all the operators in the Hamiltonian commute [12], *i .e.*,

$$[A_s, A_{s'}] = [B_p, B_{p'}] = [A_s, B_p] = 0. \tag{17.13}$$

17.5.2 The ground state

To obtain the ground state of the model, we notice that it must satisfy [12]

$$A_s|\psi_g\rangle = |\psi_g\rangle, \qquad \forall s, \tag{17.14}$$

and

$$B_p|\psi_g\rangle = |\psi_g\rangle, \qquad \forall p. \tag{17.15}$$

To obtain this ground state, we begin with a state with all spins in $\sigma_i^z = +1$ state and call this state

$$|\{\sigma^z = +1\}\rangle. \tag{17.16}$$

Clearly though this state satisfies Eq. 17.15, it does not satisfy the first relation in Eq. 17.14. In fact, operating $\hat{A}_s$ on this state gives

$$\hat{A}_s|\{\sigma^z = +1\}\rangle = |\{\sigma^z = +1\}'\{\sigma_i^z = -1, \forall i \in s\}\rangle. \tag{17.17}$$

This stands for a state where all spins are up, except for the four spins connected to $\hat{A}_s$ as shown in Fig. 17.3. Clearly this state also satisfies Eq. 17.15. Similarly, if one keeps on acting with a string of $\hat{A}_s$ operators, one gets a state where all spins are up except for those sitting on the loop described by the string of the $\hat{A}_s$ operators. Thus the state

$$|\psi_{gs}\rangle = \prod_s \left(\frac{1 + \hat{A}_s}{2}\right) |\{\sigma^z = +1\}\rangle, \tag{17.18}$$

gives an equal superposition of states composed of all closed loops of down spins living in a sea of up spins. It can now be easily checked that this state satisfies both Eqs. 17.14 and 17.15.

The above ground state does not break any symmetry of the Hamiltonian. The spin-spin correlation functions are all short ranged, *i.e.*, exponentially decaying function of separation. Indeed all correlation functions of local operators are exponentially decaying in this ground state [13].

So the central question that one may ask is if the above ground state "same as" a random product state of the form

$$|\uparrow\uparrow\downarrow\uparrow\cdots\uparrow\uparrow\downarrow\downarrow\uparrow\downarrow\rangle. \tag{17.19}$$

The answer to the above question is negative. However, to understand the basic difference between the ground state in Eq. 17.18 and the random product

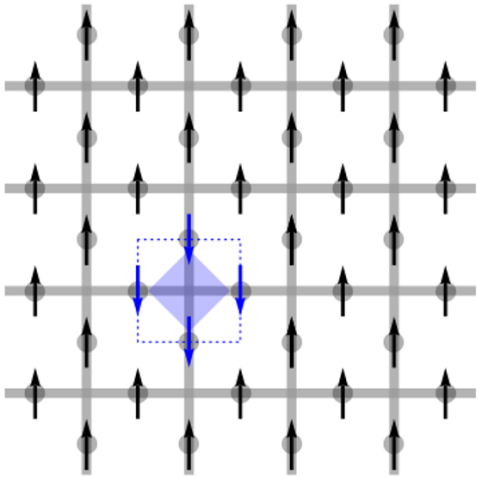


Figure 17.3: The action of $\hat{A}_s$ (the site encircled by the shaded simplex) on an all $\sigma^z = +1$ state, gives a state where all spins are up except for those on the closed dotted loop.

state in Eq. 17.19, we need to frame the question in a more quantitative form. This is what we do in the following and in the process, we shall discover a number of extremely interesting properties of the Toric code model which are characteristic of many quantum spin liquids, particularly to those known as $\mathbb{Z}_2$ quantum spin liquids.

17.5.3 The excitations

We start by examining the excitations. The excitations are obtained by flipping

$$A_s \to -1, \qquad B_p \to -1. \tag{17.20}$$

This can be done by acting with σ_i^z on a bond i for the first class of excitation, or applying σ_i^x for the second class of excitation, as shown in Fig. 17.4. Note that such excitations when created by local spin-operators, are always created in pairs. Thus, the number of sites or plaquette excitations are separately conserved modulo 2 [12]. Also, applying σ^x or σ^z twice does not create any excitation as $(\sigma^\alpha)^2 = +1$. Thus at any site or plaquette the number of excitations is an Ising variable which can either be 0 or 1.

The cost of these excitations, with respect to the ground state, is given by

$$E_s = 2(J_s + J_{s'}); \qquad E_\square = 2(J_p + J_{p'}), \tag{17.21}$$

where the first expression stands for excitations that sit on the sites, whereas the second one is for excitations which reside on plaquettes as shown in Fig. 17.4.

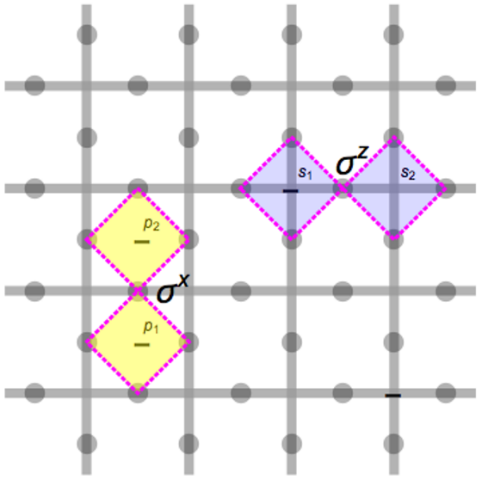


Figure 17.4: The elementary excitations in a Toric code model. The $\sigma^x(\sigma^z)$ operator creates two excitations in the adjoining plaquettes (sites) by flipping $\hat{B}_p(\hat{A}_s)$ operators defined on these plaquettes (sites).

Let us call $A_s = -1$ excitations that sites of the square lattice as electric charges. However, here the electric charge is conserved modulo 2. Hence it is really an Ising electric charge. Instead of using a single σ^z operator as shown in Fig. 17.4, if we use a string of operators

$$W_e(s_1, s_2) = \prod_{i \in s_1 \to s_2} \sigma_i^z, \tag{17.22}$$

where the product is taken over any path starting from the site s_1 and ending at site s_2, then it creates two electric charge excitations at the sites s_1 and s_2 as shown in Fig. 17.5. This is because the W_e operator commutes with all B_p's, and also all the A_s, except at the two ends s_1 and s_2. For these end operators

$$A_{s_a} W_e = -W_e A_{s_a}, \qquad (a = 1, 2). \tag{17.23}$$

So it flips A_s at the two ends. This excited state is given by

$$|e_1, e_2\rangle = W_e(s_1, s_2)|\psi_g\rangle. \tag{17.24}$$

Note that the energy cost is independent of the exact position of the string and only depends on the position of the end points. Due to this property, if two such electric charges are created, a similar operator like W_e can move them from one site to another. It is now clear that since the path of the string does not matter, interchanging the position of two electric charges give the same state, *i.e.*,

$$|e_1, e_2\rangle = |e_2, e_1\rangle. \tag{17.25}$$

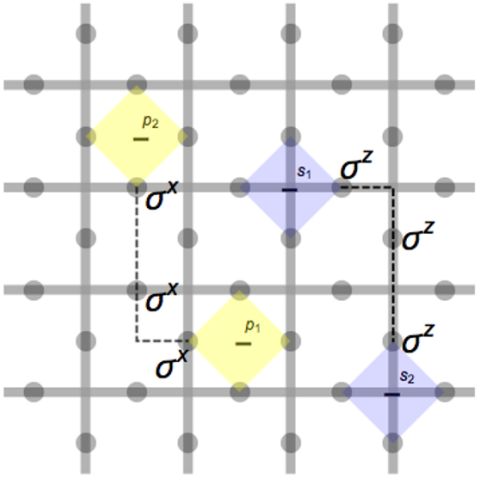


Figure 17.5

Thus these charges, that are identical quantum particles, follow bosonic statistics.

Turning to the other class of excitations ($B_p = -1$) on the plaquettes, we shall now see that in comparison to the above excitations (which we call Ising electric charges), the plaquette ones behave as Ising magnetic charges.

Similar to the above W_e operator, we now define

$$W_m(p_1, p_2) = \prod_{i \in p_1 \to p_2} \sigma_i^x, \tag{17.26}$$

an operator that defines a string in the dual lattice as shown in Fig. 17.5 with the creation of two magnetic charges at the end of the string. Again, this operator commutes with all the A_ss and also all the B_ps except at the two ends p_1 and p_2. For these end operators

$$B_{p_a} W_m = -W_m B_{p_a}, \qquad (a = 1, 2). \tag{17.27}$$

So it flips B_p at the two ends. The excited state is given by

$$|m_1, m_2\rangle = W_m(p_1, p_2)|\psi_g\rangle. \tag{17.28}$$

Just like the electric charges, the actual path does not matter. As the path operators commute with each other, the magnetic charges are also bosons.

17.5.4 Semionic mutual statistics of the electric and the magnetic charges and the bound state fermions

Consider a situation where there is an electric and a magnetic charge, where the other partner charges are sitting far off, as shown in Fig. 17.6. The magnetic

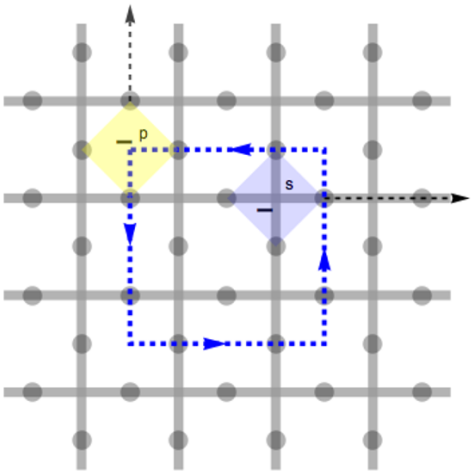


Figure 17.6: An Ising magnetic charge (on the plaquette) is moved around an Ising electric charge (on the site) as shown by the dotted loop on the dual lattice. The dashed lines on the dual lattice and the direct lattice run to infinity and their exact positions are irrelevant (see text).

charge can be moved around the electric charge by moving it through a closed loop as shown in the figure using an appropriate W_m operator which we call $W_{m,\circ}$. However, since this W_m operator necessarily has to cross the W_e string (as shown in the figure) an odd number of times (at least once), and the two strings anti-commute at the bond wherever they cross. Note that while the positions of the bonds are arbitrary, there are always odd numbers of such bond. Thus,

$$W_{m\circ}|e,m\rangle = -|e,m\rangle, \tag{17.29}$$

i.e., the state gains a negative sign. This is equivalent to saying that the electric and the magnetic charges see each other as a source of π flux. This is just an Ising version of Aharonov-Bohm phase. Such particles are called mutual semions.

One can think about forming a bound state of an electric and a magnetic charge. Extension of the above argument shows that such a bound state, (e,m), is a fermion.

17.5.5 The long range quantum entanglement in the ground state

Now that we have discussed the basic excitations, we are now in a position to answer the question that we raised earlier: what, if at all, distinguishes the ground state of the Toric code model in Eq. 17.18 from the random product ground state in Eq. 17.19?

To start of, we note that the excitations in the Toric code model have finite energy [12,13]. In other words, the excitation spectrum of Toric code has a finite energy gap. Let us now consider a Hamiltonian

$$H' = -\sum_i \Gamma_i \sigma_i^z, \tag{17.30}$$

where $\Gamma_i (\neq 0)$ is a random coupling constant. The random product state given in Eq. 17.19 is the ground state of this Hamiltonian and it is also clear that generic excitations in this system are also gapped with the minimum being given by $min(\Gamma_i)$. Now the equivalence or the non-equivalence of the two ground states in Eq. 17.18 and Eq. 17.19 can be stated as follows.

If we take the two corresponding Hamiltonians in Eq. 17.10 and 17.30, and consider a continuous one parameter family of Hamiltonians of the form

$$H_\lambda = (1-\lambda) H_{TC} + \lambda H', \tag{17.31}$$

where $\lambda \in [0,1]$, such that this interpolates between the Toric code Hamiltonian (for $\lambda = 0$) and the random field Hamiltonian (for $\lambda = 1$). Suppose that we calculate the ground state and lowest excited state for arbitrary λ and find that they are represented by $|\psi_{gs}^\lambda\rangle$ and $|\psi_1^\lambda\rangle$ respectively. Also consider that the energy gap from the ground state is given by Δ_λ. From the above argument, it is clear that both $\Delta_{\lambda=0}$ and $\Delta_{\lambda=1}$ are non-zero. Thus, if

$$\Delta_\lambda \neq 0, \qquad \forall\ \lambda \in [0,1], \tag{17.32}$$

then we can say that the Toric code ground state and the random product ground state can be connected to each other adiabatically without closing the energy gap.

On the other hand, if Δ_λ becomes zero for at least one intermediate value of λ, then the two states cannot be adiabatically connected with each other. If the above argument is true for all paths of interpolation between the Toric code and the random field Hamiltonians, then the two ground states are necessarily separated from each other by at least one "transition".

For the present example, and indeed for all quantum spin liquids, this second situation occurs, *i.e.*, they can never be adiabatically deformed into a random or any other product state. The reason why this cannot be done stems from the fact that the two ground states have different signatures of many-body quantum entanglement.

A particular measure of such many-body entanglement between different parts of a quantum many-body system is the entanglement entropy, a brief discussion of which is given in Appendix B. As discussed in this appendix, the random product state has a zero entanglement entropy whereas the Toric code ground state is long range entangled, and this prevents continuous deformation of one state into another. Such long range entanglement is a characteristic feature of quantum ordered states like the quantum spin-liquid which does not have a classical correspondence. Also, note that the above distinction of phases does not depend on the symmetry based classification of condensed matter phases discussed above.

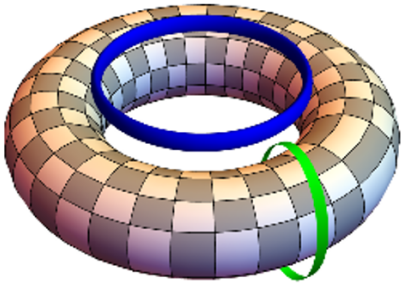


Figure 17.7: The toric code defined on a 2-torus. There are two non-contractible loops on a 2-torus which are denoted by lighter and darker loops. We denote the direction of the darker loop as $\hat{x}$ and the lighter loop as $\hat{y}$.

17.5.6 Topological Degeneracy and topological quantum numbers

A direct outcome of such long range entanglement is topological ground state degeneracy and non-trivial topological quantum numbers which are not related to any symmetry, but depends only on the topology of the manifold in which the system is embedded.

To understand these features in context of the Toric code [12], let us look back at the ground state, but this time on a 2-torus as shown in Fig. 17.7. We note that on a 2-torus, there are two non-contractible loops as shown in Fig. 17.7. For each such loop, let us define two operators as shown in Fig. 17.8.

Let us focus on the darker loop of Fig 17.7 or Fig. 17.8 which according to our notation is the $\hat{x}$ direction. Of the two lines that run in the $\hat{x}$ direction, one (the thick dark line in Fig. 17.8) passes through the bonds of the direct lattice where as the other (the dotted darker line in Fig. 17.8) passes through the bonds of the dual lattice. Let us denote these two loops as L_x^e and L_x^m respectively. Similarly for the $\hat{y}$ direction one can associate two loops as shown in fig. 17.8 denoted by L_y^e (for the thick red line in Fig. 17.8) and L_y^m (for the dotted lighter line in Fig. 17.8) respectively.

Now we define four operators for the four loops that we just defined as follows:

$$\hat{Z}_{\hat{x}} = \prod_{i \in L_x^e} \sigma_i^z; \qquad \hat{Z}_{\hat{y}} = \prod_{i \in L_y^e} \sigma_i^z; \tag{17.33}$$

$$\hat{X}_{\hat{x}} = \prod_{i \in L_x^m} \sigma_i^x; \qquad \hat{X}_{\hat{y}} = \prod_{i \in L_y^m} \sigma_i^x. \tag{17.34}$$

Clearly,

$$\hat{Z}_\alpha^2 = \hat{X}_\alpha^2 = 1, \qquad (\alpha = \hat{x}, \hat{y}). \tag{17.35}$$

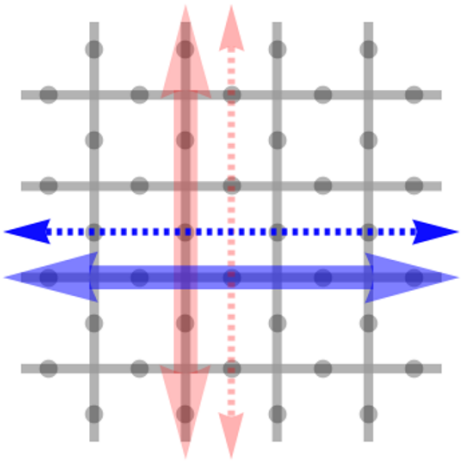


Figure 17.8: The two operators defined on the two non-contractible loops of the 2-torus (see text for details)

Thus their eigenvalues are ± 1. The algebra of the operators is given by

$$\left[\hat{X}_\alpha, \hat{X}_\beta\right] = \left[\hat{Z}_\alpha, \hat{Z}_\beta\right] = 0; \qquad \left[\hat{X}_{\hat{x}}, \hat{Z}_{\hat{x}}\right] = \left[\hat{X}_{\hat{y}}, \hat{Z}_{\hat{y}}\right] = 0, \tag{17.36}$$

and

$$\left\{\hat{X}_{\hat{x}}, \hat{Z}_{\hat{y}}\right\} = \left\{\hat{X}_{\hat{y}}, \hat{Z}_{\hat{x}}\right\} = 0, \tag{17.37}$$

Since the last two sets do not commute, they do not have simultaneous eigenstates. However one can always find the eigenstates of commuting operators $\hat{Z}_{\hat{x}}$ and $\hat{Z}_{\hat{y}}$ which can be denoted by

$$|\hat{Z}_{\hat{x}} = \pm 1, \hat{Z}_{\hat{y}} = \pm 1\rangle, \tag{17.38}$$

It is easy to check that each of the four operators defined above commute with the Toric code Hamiltonian (Eq. 17.10) on the torus, *i.e.,*

$$\left[\hat{X}_\alpha, H_{TC}\right] = \left[\hat{Z}_\alpha, H_{TC}\right] = 0. \tag{17.39}$$

Thus the energy eigenstates, in particular the ground states, can also be simultaneous eigenstates of $\hat{Z}_{\hat{x}}$ and $\hat{Z}_{\hat{y}}$. Therefore the ground state is four-fold degenerate corresponding to the above four states (Eq 17.38).

The ground state in Eq. 17.18, on a 2-torus, corresponds to $|\hat{Z}_{\hat{x}} = +1, \hat{Z}_{\hat{y}} = +1\rangle$. The other three states can be obtained from this by acting $\hat{X}_{\hat{x}}$ and $\hat{X}_{\hat{y}}$ operators on this state. It should be noted that the degeneracy of the above four states is not related to any symmetry. The degeneracy depends on the nature

of the ground state and the definitions of the above operators which in turn depend upon the non-trivial topology of the 2-torus. Hence it is a topological degeneracy. Such topological degeneracies and the resultant topological quantum numbers (eigenvalues of $\hat{Z}_{\hat{x}}$ and $\hat{Z}_{\hat{y}}$ that characterize the ground state) are manifestations of long range quantum entanglement present in the ground state.

Further, symmetry related quantum numbers can be easily destroyed by adding in small local, but, symmetry violating terms to the Hamiltonian. However, this is not true for topological quantum numbers. To see this, let us destroy the exact solvability by putting in a small transverse magnetic field of the form

$$h \sum_i \sigma_i^x \tag{17.40}$$

to the Toric code Hamiltonian of Eq. 17.10. Let us now take the 2-torus that has a dimension $L \times L$, and let us focus on the two states $|1,1\rangle$ and $|-1,1\rangle$. Then the two states are connected at the Lth order of the perturbation theory by the above perturbing term. The effective "tunnelling" Hamiltonian is given by

$$J \begin{bmatrix} 0 & \left(\frac{h}{J}\right)^L \\ \left(\frac{h}{J}\right)^L & 0 \end{bmatrix}. \tag{17.41}$$

This lifts the degeneracy by splitting the energy of the two states. However, since $h/J < 1$ for small perturbations, in the thermodynamic limit ($L \to \infty$), the splitting energy

$$\lim_{L\to\infty} \Delta E \sim J \left(\frac{h}{J}\right)^L \to 0. \tag{17.42}$$

So the topological degeneracy is preserved and hence the topological quantum numbers are robust to such perturbations in the thermodynamic limit.

This completes our discussion of the ideas of using long range entanglement in characterizing quantum order phases. The ideas of topological degeneracy, unusual excitations and topological quantum numbers together describe topological order in the Toric code system. These features are however much more general and form characteristic features of a large class of gapped quantum spin liquids called $\mathbb{Z}_2$ quantum spin liquids.

17.6 Fractionalization of spin and emergent quantum electromagnetism in S=1/2 Pyrochlore XXZ antiferromagnets

Having discussed the idea of topological order in gapped quantum spin liquids, we now turn to the idea of emergent gauge theory and quantum number

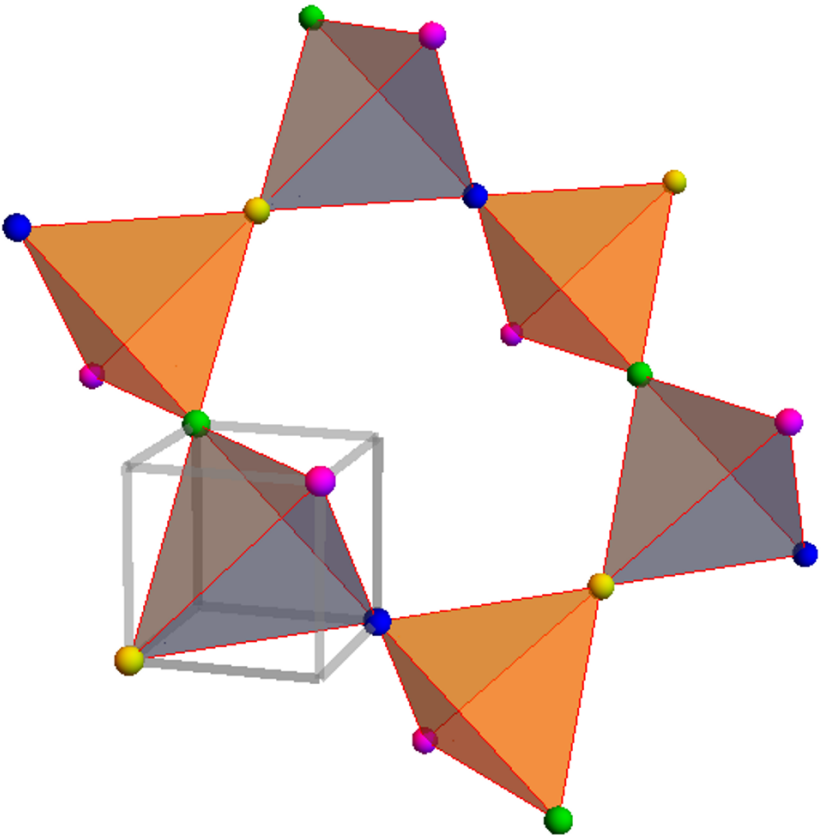


Figure 17.9: The pyrochlore lattice as a network of corner sharing tetrahedra. We choose (without loss of generality) the down tetrahedra as the unit cell. There are four sub-lattices with the underlying Bravais lattice being a face centered cubic lattice.

fractionalization. These features are also manifestation of the underlying long range quantum entanglement present in these systems.

Though these ideas are much more general, like in the previous section where we used the example of Toric code, we shall study these features using the concrete example of in S=1/2 XXZ antiferromagnets on pyrochlore lattice [14].

17.6.1 S=1/2 Pyrochlore XXZ antiferromagnets

Consider a network of corner sharing tetrahedra as shown in Fig. 17.9 with spin-$\frac{1}{2}$s sitting on the vertices. Choosing, without loss of generality, the down

tetrahedra to define the unit cell, let us define XXZ antiferromagnet given by the Hamiltonian [14]

$$H = J_{zz} \sum_{\langle ij \rangle} s_i^z s_j^z - J_{\pm} \sum_{\langle ij \rangle} (s_i^+ s_j^- + s_i^- s_j^+), \tag{17.43}$$

where $J_{zz}, J_{\pm}(> 0)$ denote antiferromagnetic interactions between nearest neighbour spins, $\langle ij \rangle$. Clearly, apart from the lattice symmetries, the above Hamiltonian has a global $U(1)$ symmetry related to the conservation of total, s^z. Therefore $s^z_{Total} = \sum_i s_i^z$ is a good quantum number for this system. We shall be interested in the particular limit where $J_{zz} \gg J_{\pm}$ in the Hamiltonian given by Eq. 17.43. To understand the situation we therefore start with the limit where $J_{\pm} = 0$ and hence we have a purely classical Ising model.

17.6.2 Classical Ising limit: Macroscopic Ground state degeneracy

In this limit, exploiting the corner-sharing geometry, the Hamiltonian can be re-written in the form [22]

$$H = \frac{J_{zz}}{2} \sum_{\boxtimes} \left(\sum_{i \in \boxtimes} s_i^z \right)^2 + \quad \text{constant}, \tag{17.44}$$

where the inner summation refers to the spins belonging to the same tetrahedra and the outer one refers to sum over all such tetrahedra. Clearly the ground state is obtained by minimizing the total s^z separately for each tetrahedra, which is , equivalent to having

$$\sum_{i \in \boxtimes} s_i^z = 0 \tag{17.45}$$

separately for each tetrahedron. Clearly this can be done in many ways. For example in a single tetrahedron, out of the $2^4 = 16$ states, six states satisfy this two-up-two-down condition. The energy cost of violating this constraint is of the order of J_{zz}.

The number of ground states in a pyrochlore lattice in this classical or Ising limit is exponential in the number of spins, a rough estimate of which can be obtained as follows [22]: The pyrochlore lattice contains up and down tetrahedra as shown in the Fig. 17.9 with the up tetrahedra being connected only with the down tetrahedra. Now if there are N spins then there are $N/4$ up and $N/4$ down tetrahedra. Let us first focus on the up tetrahedra and in each such up tetrahedra we have six possible states that satisfy the ground state constraint in Eq. 17.45. However given a configuration of spins on the up tetrahedra that satisfies the constraint (Eq. 17.45), the probability of a down tetrahedra satisfying the constraint is $6/16$. Therefore the a rough estimate of the number of states that satisfy the minimum energy constraint is given by [22]

$$\Omega \approx [6]^{N/4} \left[\frac{6}{16} \right]^{N/4} = \left[\frac{3}{2} \right]^{N/2}. \tag{17.46}$$

Therefore the entropy associated with the states is non-zero in the thermodynamic limit

$$\lim_{n\to\infty} \frac{S}{N} \approx \frac{1}{2} \ln(3/2). \tag{17.47}$$

This entropy is necessarily quenched once quantum terms like $J_\pm$ are turned on. In general, this can happen in two ways, each being interesting in its own right. The first way, which will not be a subject of this article, is called *quantum order by disorder*. This happens when the energy cost of excitations about different classically degenerate groundstates are unequal. The system then selects a classical groundstate around which the cost of excitations are particularly cheap (softer excitations), thereby gaining resonance energy from quantum tunneling which may be favoured by the quantum terms [23]. In this case the system can lower its energy by picking the corresponding ground state and hence undergo ordering.

A second way, which is relevant to XXZ pyrochlore, is that the quantum terms can lead to the quenching of the entropy without favouring any particular classical ground state, but instead choosing a superposition of all/most of them. Such a situation, in contrast to the above, is called "quantum disorder by disorder" [24]. The resultant state too gains resonance energy with respect to the classical state, but there is no ordering. The resultant ground state, due to the extensive superposition of a macroscopic number of (classically degenerate) states, give a long range entangled state as in the present case, as far as we understand [14].

17.6.3 Low energy theory: $U(1)$ quantum spin liquid

At low energies, the effect of the transverse terms in lifting the classical degeneracy, without leading to magnetic ordering (disorder by disorder), can be seen as follows [14]: In the limit $J_\pm/J_{zz} \ll 1$, the effective Hamiltonian can be obtained using degenerate perturbation theory. Since, at this energies, there are no excitations that violate the two-up-two-down condition in Eq. 17.45, the low energy Hamiltonian consists of amplitudes connecting different classically degenerate ground state configurations.

The leading order non-trivial terms in the degenerate perturbation theory, as shown by Hermele *et. al.* (Ref. [14]) (also see Appendix C), come from cooperative flipping of the spins along the smallest closed loop — the hexagons formed by six tetrahedra as shown in Fig. 17.10. This leads to an effective low energy Hamiltonian given by

$$\mathcal{H}_{\text{eff}} = -\frac{J_\pm^3}{J_{zz}^2} \sum_{\hexagon} \left(\mathcal{O}_{\hexagon} + h.c.\right), \tag{17.48}$$

where $\mathcal{O}_{\hexagon} = s_1^+ s_2^- s_3^+ s_4^- s_5^+ s_6^-$ $(1, \ldots 6 \in \hexagon)$ is an operator that flips a loop of spins on hexagons as shown in the Fig. 17.10.

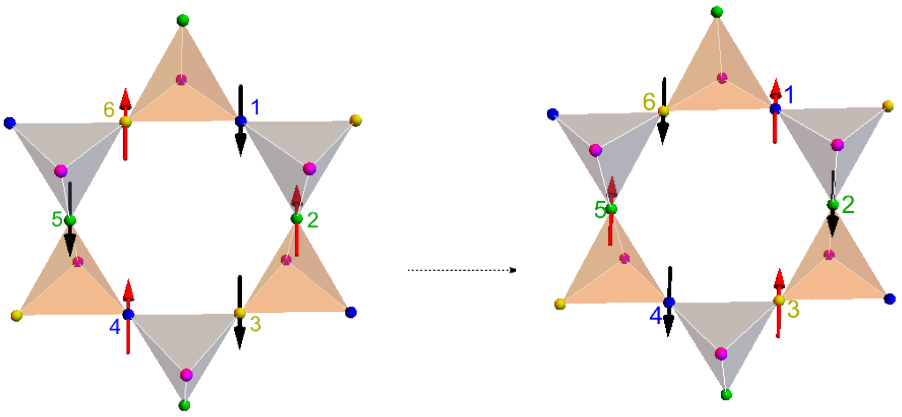


Figure 17.10: The two configurations represent two ground states where the constraint in Eq. 17.45 is obeyed. However the two configurations differ by the orientation of the six spins which form part of the hexagon. Starting with the left configuration, the right configuration can be reached by cooperatively flipping all the six spins as represented by the effective low energy Hamiltonian in Eq. 17.48.

17.6.4 Emergent quantum electromagnetism and gapless gauge boson

The low energy physics of the system encoded within the effective Hamiltonian (Eq. 17.48) becomes transparent following the mapping to an effective problem of quantum electromagnetism [14]. To this end, we note that each site of the pyrochlore lattice can be uniquely identified with a bond of the medial diamond lattice that is obtained by joining the centres of the tetrahedra forming the pyrochlore lattice. The spins sit on the bonds of this diamond lattice. Then, for a spin at site i of the pyrochlore lattice, we write

$$s_i^z = b_{\mathbf{r}\mathbf{r}'}; \qquad s^{\pm} = e^{\pm i a_{\mathbf{r}\mathbf{r}'}}, \tag{17.49}$$

where $\mathbf{r}(\mathbf{r}')$ denotes the centre of down(up) tetrahedra; Here we have introduced two new fields $b_{\mathbf{r}\mathbf{r}'}$ and $a_{\mathbf{r}\mathbf{r}'}$ which resides on the links of the diamond lattice. Starting from the spin commutation, it can be shown that the commutation between $b_{\mathbf{r}\mathbf{r}'}$ and $a_{\mathbf{r}\mathbf{r}'}$ is given by

$$[b_{\mathbf{r}\mathbf{r}'}, a_{\mathbf{r}''\mathbf{r}'''}] = i\left(\delta_{rr''}\delta_{\mathbf{r}'\mathbf{r}'''} - \delta_{\mathbf{r}\mathbf{r}'''}\delta_{\mathbf{r}'\mathbf{r}''}\right), \tag{17.50}$$

This leads us to identify $b_{\mathbf{r}\mathbf{r}'}$ is the emergent magnetic field and $a_{\mathbf{r}\mathbf{r}'}$ is the dual vector potential that is conjugate to $b_{\mathbf{r}\mathbf{r}'}$ [25]. The emergent electric field is

therefore defined by,

$$e_{\mathbf{ss}'} = \sum_{hex} a_{\mathbf{rr}'}, \tag{17.51}$$

where the sum is over all links of the hexagon (in a clockwise manner looking down on the hexagon). The righthand side, therefore is only a discretized version of the closed loop integral $\oint d\mathbf{r}\cdot\mathbf{a}$ where the integral is taken around the hexagon. Now by applying Stokes' theorem, we can convert the line integral to a surface integral of $\mathbf{a}$ over the enclosed area. This identifies the emergent electric field as the curl of the dual vector potential and the above equation is then nothing but a discretized form of this curl (*i.e.*, a lattice curl).

The emergent electric field, $e_{\mathbf{ss}'}$, therefore, is defined on the links of the dual diamond lattice whose links are denoted by $\mathbf{ss}'$ in accordance with the right-hand rule.

On the other hand the ground state constraint, Eq. 17.45, can now be written as

$$\sum_{\mathbf{r}'} b_{\mathbf{rr}'} = 0, \tag{17.52}$$

which is nothing but the Gauss law for the magnetic field in absence of magnetic monopoles [25]. In terms of the new variables, the effective low energy Hamiltonian (Eq. 17.48) can be recast as

$$\mathcal{H}_{\text{eff}} = \frac{U}{2} \sum_{\langle \mathbf{rr}' \rangle} b_{\mathbf{rr}'}^2 - \frac{J_{\pm}^3}{J_{zz}^2} \sum_{\langle \mathbf{ss}' \rangle} \cos[e_{\mathbf{ss}'}] \tag{17.53}$$

where the second term comes straight-forwardly from Eq. 17.48 following the above mapping. The first term, however needs explanation. Due to the identification in Eq. 17.49 between the emergent magnetic field, $b_{\mathbf{rr}'}$ and s_i^z, we note that the magnetic field cannot be trivially zero as $s_i^z = \pm 1/2$. For the above Hamiltonian (Eq. 17.53), this is softly implemented by putting an energy cost of U and the above model is a faithful representation of the spin model in Eq. 17.48, in the $U/(J_{\pm}^2/J_{zz}^2) \to \infty$ limit. However, universal low energy features of Eqs. 17.48 and 17.53 remain identical over a parameter range away from this strict microscopic limit, as can be explicitly checked through numerical calculations [14, 26]. Hence we shall focus on Eq. 17.53 which is easier to work and gives the universal physics of Eq. 17.48 in the right parameter regime $U/(J_{\pm}^2/J_z z^2) \gg 1$.

Clearly this Hamiltonian is invariant under the transformation

$$a_{\mathbf{rr}'} \to a_{\mathbf{rr}'} + \theta_{\mathbf{r}} - \theta_{\mathbf{r}'}; \qquad b_{\mathbf{rr}'} \to b_{\mathbf{rr}'} \tag{17.54}$$

where $\theta_{\mathbf{r}}$ is a function of the sites. This is nothing but a U(1) gauge transformation. Indeed this low energy effective theory is nothing but a pure U(1) lattice gauge theory in (3+1) dimensional space-time and hence similar to regular quantum electrodynamics. However, a major difference from the regular

quantum electrodynamic is the fact that unlike in regular quantum electrodynamics, in the present case the electric flux can change by 2π due to the cosine term in Eq. 17.53. Thus what we have in the present case is called a compact U(1) gauge theory as opposed to a non-compact U(1) gauge theory for regular electromagnetism [14, 27].

While a detailed discussion of pure compact U(1) gauge theories would lead us too far afield and interested readers can consult Ref. [27, 28], we note that such theories can have two phases — (1) the confined phase, and (2) the deconfined phase. It is easy to understand these phases qualitatively in the light of the discussion in the above paragraph. If the quantum fluctuations associated with the change of electric flux by 2π are important, then it means that the electric field in the medium is rapidly changing. Suppose, in such an environment, we put two oppositely charged test magnetic charges. Because of the random change in the surrounding electric field, these test magnetic charges see a random Aharonov-Bohm phase and their amplitude of propagation from one point in space to another is severely affected by such random phases causing destructive interferences. Spoken in another way, a magnetic charge cannot propagate in a background of rapidly changing electric field. However, if the two oppositely charged magnetic test charges combine to form a neutral particle, this neutral particle can then propagate almost freely as it is almost insensitive to the fluctuations of the electric field. Thus, in this phase, oppositely charged magnetic charges are "confined" to each other to form neutral particles. This is the confined phase noted above, and it is not of direct interest to us here since the U(1) quantum spin liquid is described by the deconfined phase of the above gauge theory which we briefly describe below [14, 27, 28].

In contrast to the confined phase, in the deconfined phase, the quantum fluctuations leading to the the change in the electric flux by 2π are not important. Hence the fluctuations in the background electric field are smooth, and as an extension of the ideas presented in the previous paragraph, it can be argued that test magnetic charges are now free to propagate in the deconfined phase as the effect of destructive interference is much less due to the absence random Aharonov-Bohm phases. Therefore, free magnetic charges can almost freely propagate in the deconfined phase and hence they are "deconfined" as opposed to the previous phase. This deconfined phase corresponds to a U(1) quantum spin liquid phase which is of interest to us.

In this deconfined phase, as the fluctuations of the emergent electric field are smooth, we can expand the cosine terms in Eq. 17.53. While we can keep on working with the lattice theory in Eq. 17.53, it is often more insightful to derive a continuum limit of this theory that is valid at long wavelengths and low energies. The continuum limit is obtained by taking

$$\sum \to \frac{1}{l^3} \int d^3\mathbf{r}; \qquad e_{\mathbf{ss}'} = l\mathbf{e} \cdot \hat{\mathbf{l}}_{\mathbf{ss}'}; \qquad b_{\mathbf{rr}'} = l\mathbf{b} \cdot \hat{\mathbf{t}}_{\mathbf{rr}'},$$

where l is a lattice length scale. Here, $\hat{\mathbf{t}}_{\mathbf{rr}'}$ and $\hat{\mathbf{l}}_{\mathbf{ss}'}$ are the unit vectors in the direction from $\mathbf{s}$ to $\mathbf{s}'$ that forms the links of the direct and dual diamond

lattice respectively. Thus $\sum_{\mathbf{r}'} \hat{t}^{\alpha}_{\mathbf{r}\mathbf{r}'} \hat{t}^{\beta}_{\mathbf{r}\mathbf{r}'} = \sum_{\mathbf{s}'} \hat{l}^{\alpha}_{\mathbf{s}\mathbf{s}'} \hat{l}^{\beta}_{\mathbf{s}\mathbf{s}'} = \frac{4}{3}\delta^{\alpha\beta}$ as can be checked explicitly. The continuum Hamiltonian is given by:

$$\mathcal{H}_0^{\text{continuum}} = \frac{1}{2} \int d^3\mathbf{r} \left[\mathcal{U}\ \mathbf{b}^2 + \mathcal{K}_0 \mathbf{e}^2 \right], \tag{17.55}$$

with $\mathcal{U} = 4U/3l$ and $\mathcal{K}_0 = \frac{4J_{\pm}^3}{3lJ_{zz}^2}$. This is clearly the Hamiltonian for non-compact U(1) gauge theory [14, 25, 27], similar to the theory of quantum electromagnetism which supports gapless excitations akin to photons [29]. Here we call such excitations as emergent photons.

The gaplessness of the photon is protected by the gauge invariance at low energy (Eq. 17.54) which prevents the photons from acquiring a mass. This is very different from the gapless spin-waves discussed earlier as examples of Nambu-Goldstone bosons which are also massless but due to a spontaneously broken continuous symmetry (spin rotation symmetry in our case). In the case of spin-waves, a gap can be easily obtained by explicitly breaking the spin-rotation down to a discrete group. However to produce a photon mass, one has to break the gauge structure.

17.6.5 Fractionalization of the spin

The spin of the spin-waves or the magnons is clearly 1 as it is obtained by flipping a single spin. However, in case of the above model, if we flip one spin in any of the classical ground states of our present system, we violate the ground state constraint of Eq. 17.45 in two tetrahedra which shares the spin. Thus two defects are created which cost an energy of the order J_{zz} and they together have $s^z = 1$ as only one spin has been flipped. These point defects are excitations and the quantum terms (proportional to $J_{\pm}$) can move these excitations to form a gapped band. It is easy to see that these particles obey a bosonic statistics and since they violate zero divergence constraint for the magnetic field in Eq. 17.52, they can be identified with magnetic monopoles [14, 30].

Clearly, the combined s^z of two such magnetic monopoles is 1. However, such magnetic monopoles are deconfined in a sense that, once created, they can be separated from each other without further energy cost by flipping a string of spins [14,30]. Also, the spin quantum numbers of the magnetic monopoles must be identical as there is no distinction between one from the other. Thus each magnetic monopole must have $s^z = 1/2$. This clearly is different from magnons, in comparison to which the spin is fractionalized. However, the present magnetic monopoles do not carry charge of the underlying electrons from which these spins are made. Hence, they represent excitations with fractionalized quantum numbers. Such fractionalization of quantum numbers, again, has been possible due to the presence of underlying long range quantum entanglement.

17.7 Conclusion

In this article (and indeed in the set of lectures), I have tried to explain the major features of quantum ordered states in frustrated magnets through the example of two well-known systems — (1) Toric code model [12,13], and (2) Pyrochlore XXZ antiferromagnets [14]. At the same time, I have tried to contrast their features from the conventional magnetically ordered state. Together, the two examples provide glimpses of the new framework required to understand quantum phases that lie beyond the conventional paradigm of spontaneous symmetry breaking.

Needless to say, several things have been left out from the above discussion. From the theoretical side, the general mean field theory to describe the above type of fractionalized particles in quantum spin liquid theory, variously called slave boson/ slave fermion/parton mean field theories [11] have been left out. These mean field theories often form the first step of understanding the nature of the fractionalization present in the quantum spin liquid where the fractionalized excitations transform under projective representation of the symmetry group of the system [6, 10]. Such projective classifications currently appears to provide a strong framework to describe the interplay of symmetry fractionalization, topological order and long range quantum entanglement in quantum spin liquids [6, 10]. However, keeping in mind that the set of lectures was supposed to be an introduction to the physics of quantum spin liquids, we thought that providing explicit examples where some of the basic features of quantum spin liquids emerge would help the readers to get to the core of the problem at once (within the time-frame of a set of two lectures). The framework required for the generalization of some of the ideas introduced here then form the important second step for readers aiming to work in this area.

Another very serious lacuna of the present article is the lack of descriptions of many exciting experiments that, in a way, have driven the field of strongly correlated system and provided the constant motivation to develop a more general framework of condensed matter. However, for that I must refer to other reviews [7, 8]. Some of these experimental results are very much a topic of present research and the theoretical ideas described here are intended to provide a platform for readers to try to seek answers to some of these questions.

Acknowledgements

I would like to acknowledge people with whom I have interacted with over the years for research in this area. I have immensely benefitted from various discussions during these collaborations. In no particular order, they are: R. Moessner, Y. C. He, F. Pollmann, Y. Fuji, K. Raichaudhuri, Y. B. Kim, S.-S. Lee, SungBin Lee, D. H. Lee, E. K. H. Lee, G. J. Sreejith, R. Schaffer, K. Hwang, J. Rau, H. Y. Kee, A. Paramekanti, K. Damle, A. Sen, K. Sengupta, G. Baskaran, R. Shankar, S. Mandal, V. B. Shenoy, J. Knolle, T. Senthil, T. Dodds and S. Zherlitsyn. I would like to thank the organizers of the SERC school for

giving me this opportunity to deliver the set of lectures and particularly S. M. Bhattacharjee for his patience with me regarding the typing up of these lecture notes. I would also like to thank the hospitality at IACS, IIT-Kharagpur, IISER-BHOPAL, IIT-Bombay, and IMSc where various versions of these lectures were delivered. I would also like to acknowledge DAE for funding a part of this research though ICTS-TIFR, and G. Dasgupta for her help with proof reading.

Appendix A: From electron to spin Hamiltonians

Consider electrons hopping on the lattice (like in Fig. 17.1) and repelling each other through Coulomb repulsion. The long range part of the Coulomb repulsions are often screened leading to short range repulsive interactions. These class of systems, then can be described by Hubbard type of minimal Hamiltonians where only on-site repulsion of electrons are retained. Such Hamiltonians are given as

$$H = \sum_{ij} t_{ij}^{\sigma\sigma'} c_{i\sigma}^{\dagger} c_{j\sigma'} + U \sum_{i} n_{i\uparrow} n_{i\downarrow}, \tag{17.56}$$

where $c_{i\sigma}^{\dagger}$ creates an electron of spin $\sigma = \uparrow, \downarrow$ at site i. U represents the on-site Coulomb repulsion. On the other hand $t_{ij}^{\sigma\sigma'}$ represents the amplitude of an electron, at site j and spin σ', to hop to site i with spin σ. In Eq. 17.56, we have written an effectively single band model since there are no-orbital index. In absence of spin-orbit coupling, spin-rotation symmetry is present in terms of the physical spin σ, σ'. In this case we must have

$$t_{ij}^{\sigma\sigma'} = t_{ij}^{0} \delta_{\sigma\sigma'}. \tag{17.57}$$

However, in presence of spin-orbit coupling, σ, σ' represents a pseudospin-$\frac{1}{2}$ that is a combination of the orbital and spin angular momenta. Such a situation has recently attracted much interest in context of $5d$ transition metal Iridium where a pseudo-spin-$\frac{1}{2}$ ($J = L + s = \frac{1}{2}$) emerges as the right low energy atomic orbital in terms of which a single band Hubbard model (like in Eq. 17.56) can be written down. In such cases, the hopping amplitude depends non-trivially on the "spin" indices and is actually a 2×2 matrix in this space which has the following form

$$t_{ij}^{\sigma\sigma'} = t_{ij}^{0} \delta_{\sigma\sigma'} + i \mathbf{t}_{ij} \cdot \boldsymbol{\sigma}_{\sigma\sigma'} \tag{17.58}$$

where $i\boldsymbol{\sigma}$ represents the Pauli matrices. $\mathbf{t}_{ij}$ transforms as a pseudovector under lattice transformations and can be non-zero in presence of spin-orbit coupling representing the spin-flip hopping processes.

If the system is at half filling *i.e.*, one electron per site on average, then in the limit $t_{ij}^{0}, \mathbf{t}_{ij} \gg U$, the system is in a metallic state where all the electrons can move freely. This state is adiabatically connected to the $U = 0$ free electron limit. In the opposite limit, however, the repulsion term tries to force the

electrons to stay put individually at the lattice sites as double occupancy costs an energy of the order $\sim U$. The electron spins are however free to fluctuate and the ground state in the $U = \infty$ limit (at finite $t^0_{ij}, \mathbf{t}_{ij}$) is macroscopically degenerate leading to finite ground state entropy. This entropy is quenched by quantum fluctuation of the electrons which try to gain delocalisation energy through virtual hopping (to the leading order) from site i to j (say) and back with the appropriate hopping amplitude. The intermediate virtual state where there are two electrons at the site j costs an energy $\sim U$. From Eq. 17.58, it is easy to see that there are three kinds of such processes which are:

1. The electron hops between site $i \leftrightarrow j$ through the spin conserving amplitude t^0_{ij}. The resultant scalar coupling is $\sim (t^0_{ij})^2/U$ which describes the scalar interaction $\mathbf{s}_i \cdot \mathbf{s}_j$ where the spin operators are related to the electron operators as:

$$\mathbf{s}_i = \frac{1}{2} c^\dagger_{i\sigma} \boldsymbol{\sigma}_{\sigma\sigma'} c_{j\sigma'} \tag{17.59}$$

2. The electron hops from $i \rightarrow j$ through the spin conserving hopping t^0_{ij}, but while returning it uses spin-flipping amplitude, $\mathbf{t}_{ij}$. Thus the coupling constant is proportional to $\mathbf{D}_{ij} \sim t^0_{ij}\mathbf{t}_{ij}/U$ which is a pseudovector. This couples to a pseudovector operator $\mathbf{s_i} \times \mathbf{s_j}$. Note that the opposite order of t^0_{ij} and $\mathbf{t}_{ij}$ also gives the same term.

3. Finally, the electron can hop between $i \leftrightarrow j$ using spin-slip hopping $\mathbf{t}_{ij}$, which gives a symmetric tensor coupling constant $t^\alpha_{ij} t^\beta_{ij}/U$ where $\alpha, \beta = x, y, z$. This couples to a symmetric tensorial operator $s^\alpha_i s^\beta_j + s^\beta_i s^\alpha_j$.

The standard way to perform this strong coupling perturbation theory is discussed in several textbooks, such as Ref. [15]. All the above terms are of the form of the first (quadratic) term of Hamiltonian in Eq. 17.3. Extending this perturbation theory to higher orders in t/U gives further terms like the second (quartic) term of Hamiltonian in Eq. 17.3.

Thus we find that in the $4t/U \ll 1$ limit, the electrons themselves cease to be the right degree of freedom in terms of which the higher energy Hubbard model (Eq. 17.56) is written down. At low energies (smaller compared to U), spins emerge as the right degree of freedom in terms of which the low energy spin Hamiltonian is written down. In this sense, the spins emerge as effective degrees of freedom at low energies from the high energy degrees of freedom, the electron.

Appendix B: Introduction to Entanglement entropy and its connection to characterizing quantum spin liquids

In this appendix, we give a short introduction to entanglement entropy as a measure of long range quantum entanglement. For a quantum many-body system in state $|\psi\rangle$, it is worthwhile to consider a sub-part of the system denoted by

A (say) and the rest of the system by $\bar{A}$ (say). A measure of such entanglement is obtained by considering the density matrix for the entire system $\hat{\rho} = |\psi\rangle\langle\psi|$ and integrating out the degrees of freedom in $\bar{A}$ to obtain the reduced density matrix

$$\hat{\rho}_A = Tr_{\bar{A}} [\hat{\rho}] . \tag{17.60}$$

The entropy of this reduced density matrix

$$S_{A\bar{A}} = -Tr_A [\hat{\rho}_A \ln \hat{\rho}_A] , \tag{17.61}$$

then gives a measure for the entanglement between A and $\bar{A}$. Note that there may be other measures of entanglement which we do not discuss here.

Clearly for a product state like the approximate ground state of ferromagnet in Eq. 17.7 or the random product state in Eq. 17.19, $S_{A\bar{A}} = 0$. Adding small superposition to such states induces small amount of entanglement and hence gives a non-zero value of entanglement entropy of the form

$$S_{A\bar{A}} = \#L, \tag{17.62}$$

where L is the length of the perimeter of the boundary between A and $\bar{A}$ and $\#$ is a number that is zero for an exact product state.

A naive look at the ground state of Toric code (Eq. 17.18) may suggest that since the spins are randomly up or down, it may give an entanglement entropy of the form given above. However, a more careful thought suggest that there is a pattern in this massively superposed state — the down spins always lie on a closed loop. Therefore, when we divide the entire system into subsystems A and $\bar{A}$, then some of these loops cross the boundary. The number of intersections from each loop with the boundary must be an even number as the loops are closed. Therefore, the state in Eq. 17.18 has this one qubit of information encoded in it that the number of such intersection is $0(mod\ 2)$. This translates into an extra $-\ln 2$ factor in the entanglement entropy, which cannot be changed continuously (unlike the prefactor of L) and corresponds to the presence of long-range quantum entanglement in the ground state of the Toric code model.

Appendix C: Low energy effective Hamiltonian for XXZ pyrochlore antiferromagnets

In the limit J_{zz} is larger than $J_{\pm}$ in the Hamiltonian in Eq. 17.43, we can derive the effective low energy (below the energy scale of $O(J_{zz})$) [14]. This effective Hamiltonian is given by

$$H_{eff} = \mathcal{P} [H_p + H_p G_0' H_p + \cdots] \mathcal{P}, \tag{17.63}$$

where $\mathcal{P}$ is the projector to the classical ground state manifold satisfying Eq 17.45 and

$$\begin{aligned} H_p &= -J_{\pm} \sum_{\langle ij \rangle} (s_i^+ s_j^- + s_i^- s_j^+); \\ G_0' &= (1-\mathcal{P}) \frac{1}{E - H_{zz}} (1-\mathcal{P}); \\ H_{zz} &= J_{zz} \sum_{\langle ij \rangle} s_i^z s_j^z, \end{aligned} \tag{17.64}$$

First order: At the first order all terms in H_p take the state out of the classical ground-state manifold. So there are no first order contributions since the projection operator $\mathcal{P}$ in Eq. 17.63 kills all such terms.

Second order: The second order term has the form (see Eq. 17.63)

$$H_{eff}^{(2)} = \mathcal{P} \left[H_p G_0' H_p \right] \mathcal{P}. \tag{17.65}$$

This can only lead to non-zero contribution when the spins flipped by each of the two H_p refer to the same bond. In this case, however, the ground state remains the same, and hence this leads to a constant term of the order $O(J_{\pm}^2)$ [14]. We neglect such constants.

Third order The third order term has the form

$$H_{eff}^{(3)} = \mathcal{P} \left[H_p G_0' H_p G_0' H_p \right] \mathcal{P}. \tag{17.66}$$

The non trivial contributions come from the hexagons shown in Fig. 17.10, and the two classical ground-states they connect, differ in configuration with respect to the six spins forming the flippable hexagon. This contribution is of the type [14]

$$\sim -\frac{J\pm^3}{J_{zz}^2} \sum_{\hexagon, (1,\ldots 6 \in \hexagon)} \left(s_1^+ s_2^- s_3^+ s_4^- s_5^+ s_6^- + h.c. \right), \tag{17.67}$$

which is same as the effective low energy Hamiltonian is Eq. 17.48.

References

[1] P. W. Anderson *Basic notions of condensed matter physics* (Benjamin-Cummings, 1984).

[2] S. Ma, *Modern theory of critical phenomena*, 46 (Da Capo Press, 2000).

[3] T. Senthil, International Journal of Modern Physics B **20**, 2603 (2006).

[4] K. v. Klitzing, G. Dorda, and M. Pepper, Phys. Rev. Lett. **45**, 494 (1980).

[5] D. C. Stormer, and A. C. Gossard, Phys. Rev. Lett. **48**, 1559 (1982).

[6] X.-G. Wen, *Quantum Field Theory of Many-body Systems* (Oxford University Press Inc., New York. ISBN 019853094., 2004).

[7] P. A. Lee, Science (New York, NY) **321**, 1306 (2008).

[8] L. Balents, Nature **464**, 199 (2010).

[9] P. W. Anderson, Materials Research Bulletin **8**, 153 (1973).

[10] X.-G. Wen, Physical Review B **65**, 165113 (2002)..

[11] G. Baskaran, Z. Zou, and P. Anderson, Solid state communications **63**, 973 (1987).

[12] A. Y. Kitaev Annals of Physics **303**, 2 (2003).

[13] A. Kitaev Annals of Physics **321**, 2 (2006).

[14] M. Hermele, M. P. A. Fisher, and L. Balents, Phys. Rev. B **69**, 064404 (2004).

[15] A. Auerbach, *Interacting electrons and quantum magnetism* (Springer Science & Business Media, 2012).

[16] A. Zee, *Quantum field theory in a nutshell* (Princeton university press, 2010).

[17] V. Berezinskii, Soviet Journal of Experimental and Theoretical Physics **32**, 493 (1971).

[18] V. Berezinskii, Soviet Journal of Experimental and Theoretical Physics **34**, 610 (1972).

[19] J. M. Kosterlitz and D. J. Thouless, Journal of Physics C: Solid State Physics **6**, 1181 (1973).

[20] N. D. Mermin Reviews of Modern Physics **51**, 591 (1979).

[21] D. Thouless, *Topological quantum numbers in nonrelativistic physics* (World Scientific Publishing Co Pte Ltd, 1997).

[22] J. T. Chalker, *Introduction to frustrated magnetism* (Springer, 2011) pp. 3–22.

[23] E. F. Shender, Sov. Phys. JETP **56**, 178 (1982).

[24] P. Fazekas and P. W. Anderson, Phil. Mag. **30**, 23 (1974).

[25] J. D. Jackson, *Classical electrodynamics* (Wiley, 1999).

[26] O. Benton, O. Sikora, and N. Shannon, Phys. Rev. B **86**, 075154 (2012).

[27] R. Savit, Rev. Mod. Phys. **52**, 453 (1980).

[28] J. B. Kogut, Rev. Mod. Phys. **51**, 659 (1979).

[29] J. J. Sakurai and J. Napolitano, *Modern quantum mechanics* (Addison-Wesley, 2011).

[30] C. Castelnovo, R. Moessner, and S. L. Sondhi, Nature **451**, 42 (2008).

18

Topological aspects of quantum information processing

Ville Lahtinen and Jiannis K. Pachos

In this review we present an introduction to topological quantum computation – quantum computing with anyons. This approach is inherently resilient against errors, thus promising to overcome one of the main obstacles for the realisation of quantum computers. We first provide an introduction to anyon models and discuss the general steps how to use them to encode and process quantum information. Then we present a toy microscopic model, Kitaev's p-wave wire, that supports localised Majorana modes. These are the simplest non-Abelian anyons, which are the subject of intense theoretical and experimental research. We outline how quantum computation could be carried out in this microscopic system in a topologically protected manner, discuss the nature of the topological fault tolerance and review the recent experimental developments in realizing Majorana modes.

18.1 Introduction

Physics should remain unchanged when two identical particles are exchanged. This is a fundamental symmetry with far reaching consequences. In three spatial dimensions it dictates that only bosons and fermions can exist as point-like particles. A wave function describing them acquires a $+1$ or a -1 phase, respectively, whenever two particles are exchanged. However, when one goes down to two spatial dimensions, perhaps somewhat counter-intuitively as the number of degrees of freedom is reduced, a much richer variety of statistical behaviours is allowed. In addition to bosonic and fermionic behaviour, arbitrary phase fac-

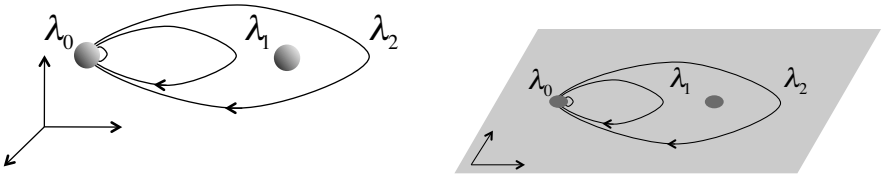
S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6_18

tors, or even non-trivial unitary evolutions, can be obtained when two particles are exchanged [1]. Particles with such an exotic statistics are called *anyons*.

The fundamental difference between two and three spatial dimensions relies on the topology of space-time evolutions of point-like particles. To understand this principle consider the exchange processes of two particles illustrated in Figure 18.1. In three dimensions the loop λ_2 drawn by the encircling particle is always continuously deformable to loop λ_1 that does not encircle the other particle. This loop, in turn, is fully contractible to a point, which means that the wave function of the system must satisfy

$$\text{3D}: \qquad |\,\Psi(\lambda_2)\rangle = |\,\Psi(\lambda_1)\rangle = |\,\Psi(0)\rangle\,. \tag{18.1}$$

As one particle encircles the other twice, the evolution of the system can be represented by the exchange operator R such that $|\,\Psi(\gamma_2)\rangle = R^2\,|\,\Psi(0)\rangle$. The contractibility of the loop requires that $R^2 = 1$, which has only the solutions $R = \pm 1$ that correspond to the exchange statistics of either bosons or fermions. Since the order and the orientation of the exchanges are not relevant, the statistics of point-like particles in three spatial dimensions are described by the simple permutation group.



(a) (b)

Figure 18.1: Exchange statistics in two vs. three spatial dimensions. (a) In three spatial dimensions the loop λ_2 describing two particle exchanges is continuously deformable to λ_1 that encloses no particles, which in turn is contractible to a point λ_0. (b) In two spatial dimensions, however, the loops λ_2 and λ_1 are topologically inequivalent, while λ_1 is still contractible to a point.

This contrasts with two spatial dimensions, where the loop λ_2 is no longer continuously deformable (the path is not allowed to cross the encircled particle) to the fully contractible loop λ_1. This means that the final state $|\,\Psi(\lambda_2)\rangle$ no longer needs to be equal to the initial state $|\,\Psi(0)\rangle$

$$\text{2D}: \qquad |\,\Psi(\lambda_2)\rangle \neq |\,\Psi(\lambda_1)\rangle = |\,\Psi(0)\rangle\,. \tag{18.2}$$

Hence the exchange operator R is no longer constrained to square to identity either. Instead, it can be represented by a complex phase, or even a unitary matrix. The only constraints on it are a set of consistency conditions on the order and orientation of the exchanges. These derive from a mathematical structure known as *the braid group*, which describes all topologically distinct evolutions

of point-like particles in two spatial dimensions. It is this description of the statistics by the braid group that allows anyons to exist.

Quantum computation with anyons is referred to as *topological quantum computation*. We illustrate the basic principles behind any topological scheme for quantum computation and motivate why pursuing such an exotic route to fault-tolerance is to be taken seriously. We start with reviewing the general structure of consistent *anyon models*. These models emerge in the low-energy behaviour of topologically ordered many-body systems. Having thus set the table, we outline how the anyons could in principle be used to encode and process quantum information. Importantly we present the main appeal of such exotic sounding idea: the high fault-tolerance provided by a topology based encoding. While the discussion up to this point takes place at an abstract general level, we then illustrate all these concepts in the context of simple microscopic model – Kitaev's p-wave wire [20] – whose experimental realization is a subject of intense contemporary research. This model supports Majorana modes, simplest types of non-Abelian anyons. We outline how the encoding and processing of quantum information by topological means could be carried out in an array of such wires, and conclude by reviewing recent advances to go beyond Majorana modes.

In this review we focus on the recent developments on the realization of scalable topological quantum computing with anyons. The closely related topic of (topological) quantum memories and error correction is covered in other recent reviews [76, 77]. Before proceeding we provide a brief history of the key developments in topologically ordered systems and in their application to quantum computation.

18.1.1 A brief history of anyonic quantum computation

The study of anyons started as a theoretical toy model for quasiparticles with fractional spin in two spatial dimensions [2]. They were promoted from a theoretical curiosity to a serious research direction after the discovery of the fractional quantum Hall effect [3, 4]. This effect emerges when a gas of electrons confined to two spatial dimensions is subjected to a strong magnetic field and very low temperatures. Then several distinct phases of matter can be experimentally distinguished by precisely quantised conductivity, which is proportional to a topological invariant characterising the ground state [5]. Thus, such states became known as *topologically ordered*. The theoretical breakthrough for understanding this behaviour was provided by Laughlin, who explained the physics of these strongly correlated phases in terms of simple effective trial wave functions [6]. The remarkable property of these fractional quantum Hall wave functions was the prediction that they could support quasiparticles carrying fractional charge and behaving as anyons. While the experimental verification of this prediction was challenging at the time, the fractionalised charge of the quasiparticles has since been confirmed [69, 73]. The conclusive experimental verification of the anyonic statistics, however, has remained elusive.

Due to the rapid progress, the early studies on topological order focused on fractional quantum Hall states. As the experiments got more precise, a zoo of different potential topological ordered states at different filling fractions were discovered and numerous new trial wave functions were proposed to describe them. Out of these particularly interesting was the Moore-Read state at the filling fraction 5/2 [24], since it was the simplest state supporting the so called non-Abelian anyons – anyons associated with non-local degeneracies and statistics represented by unitary matrices (as opposed to phase factors of Abelian anyons). While much effort was dedicated to verifying the existence of these anyons, the experiments proved challenging [73]. Fortunately, as the understanding of the physics of topologically ordered states grew, it was realised that fractional quantum Hall systems are not the only place to look for them. A seminal paper by Read and Green showed that they could occur also in superconductors with special pairing symmetries [21]. Of particular interest of these were the p-wave paired superconductors that were predicted to support the same type of anyons as the Moore-Read states. In these systems they would manifest themselves as fractionalized fermion modes, or *Majorana modes*, bound to vortex excitations [23].

Parellel to the study of topological order in condensed matter systems, a lot of research was directed to a promising new field of quantum computation and quantum information [7]. There the challenge is to store and manipulate quantum information in a robust manner, which had lead to the study of quantum error correcting codes. While working on a class of them known as surface codes, Kitaev realised that some of these error correcting codes could be written as spin lattice models with special Hamiltonians [39]. The ground states of these models exhibit decoherence-free subspaces that resembled much the properties of topologically ordered systems. The simplest of such models came to be known as the Toric Code – nowadays the archetypal topological quantum memory– where the errors in the encoded information could be understood as anyonic excitations. Kitaev then showed that instead of being harmful in this particular model, anyons in some more complicated system could actually be used to both encode and process quantum information in a fault tolerant manner. Thus topological quantum computation was born [38, 39].

The main advantage for using anyons for quantum computation is the existence of a decoherence-free subspace, which can only be evolved through non-local operations by moving the anyons around each other. Local noise operators such as the ones employed in realistic noise models, can not cause errors in this space. In other words, errors would be suppressed at the level of hardware. The operator of the computer, on the other hand, can in principle implement such operations through adiabatic transport of the anyons, and thereby implement accurately topologically protected quantum gates. This may sound straightforward, but the challenge is to first find a system that supports anyons and enables them to be manipulated in a robust manner. The models emerging from Kitaev's construction required many-body interactions and were thus beyond immediate experimental relevance. A challenge to topological quantum

computation is also that not all anyons are universal for quantum computation. Sufficiently complex systems are required in order to obtain a universal gate set only by braiding the anyons. The zoo of potential fractional quantum Hall states held the initial promise. The putative filling fraction 12/5 was proposed to be described by the Read-Rezayi state [8] that was conjectured to support so called Fibonacci anyons. These, as opposed to the Ising anyons appearing in the Moore-Read state (equivalent to Majorana modes for most practical purposes), were universal for quantum computation. Unfortunately, but to no one's surprise, the robust experimental realisation of the Read-Rezayi state has proved even more challenging as the already elusive Moore-Read state.

While progress with the fractional quantum Hall systems has been made, the current push for topological quantum computation comes from topological superconductors. Even though also p-wave superconductors are yet to be experimentally confirmed in materials, it was realised that qualitatively same physics could occur when a topological insulator [46], a spin-orbit coupled semiconductor [47,50] or a a chain of magnetic atoms [54,55] is placed in the proximity of a regular s-wave superconductor. In particular, wires made of these materials and deposited on top of a superconductor were predicted to host Majorana modes at their ends, which could be probed through simple conductance measurements [20,48,49]. While the explicit verification of their braiding properties is yet to be carried, several experiments on microscopically distinct setups strongly support the existence of Majorana modes [40–42, 78]. Experiments have also been proposed for cold atoms in optical lattices that due to their inherent cleanness and high degree of control will hopefully yield further evidence [56,57,90].

From the point of view of quantum computation this breakthrough is slightly bitter sweet, since Majorana modes are not universal for quantum computing by purely topologically protected operations. However, as they are the first realisable non-Abelian anyons, they hold great promise to experimentally test the two key elements of topological quantum computation – topological protection of quantum encoding and the implementation of quantum gates by braiding anyons. In the following we provide an introduction to the concept of anyonic models [22, 32] and describe the steps required to realise anyonic quantum computation.

18.2 Anyon models

The hallmark of topologically ordered phases in two spatial dimensions is the existence of fractionalised quasiparticles that obey anyonic statistics. Different types of anyons can be divided into two general classes. If the exchanges of two quasiparticles makes the wave function of the system to acquire a complex phase factor the quasiparticles are referred to as *Abelian anyons*. If the presence of the quasiparticles is associated with ground state degeneracy, then an exchange can lead to a unitary evolution in the ground state manifold. In this case the quasiparticles are known as *non-Abelian anyons*. While both types of anyons

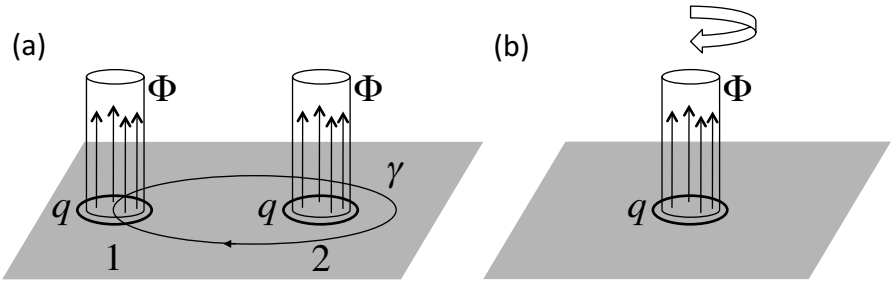


Figure 18.2: Toy model for anyons as charge-flux composites. (a) Anyons can be described effectively as magnetic flux Φ confined to a tube that is encircled by a ring of electric charge q. When anyon 1 moves around anyon 2 along loop λ, its charge (flux) circulates the flux (charge) of the other anyon. The Aharonov-Bohm effect gives rise to the complex phase $e^{2iq\Phi}$, which describes the mutual statistics of the composite objects. For $2q\Phi \neq 2\pi$ they are Abelian anyons. (b) When the composite particle rotates around itself by 2π, the system acquires the phase factor $e^{iq\Phi}$ as its charge circulates its own flux. Thus the composite particle has the spin $h = q\Phi/(2\pi)$.

have applications to quantum information, it is predominantly the latter that one refers to when talking about topological quantum computation.

The full microscopic description of strongly correlated systems is often very complicated. Nevertheless, if such systems are topologically ordered then the low energy description can be given only in terms of the anyonic quasiparticles. Due to their topological nature the possible evolutions are limited to three simple scenarios: (i) Anyons are created or annihilated in pairwise fashion, (ii) they fuse to form other types of anyons and (iii) they can be exchanged (adiabatically moved around each other). The physical framework capturing these properties in a unified fashion goes under the name of a *topological quantum field theory* [17]. However, often in physics the details of this rigorous description can be omitted and only a minimal set of data is required to specify the properties of the anyons corresponding to a particular topological quantum field theory. Here, as also often is the case in literature, we refer to this minimal information as an *anyon model*. Before describing the general structure of anyon models (for a thorough account we refer to [22]), we warm-up by presenting a simple toy model for anyons in terms of Aharonov-Bohm effect.

18.2.1 Toy anyons from the Aharonov-Bohm Effect

Let us think of anyons as being composite particles consisting of a magnetic flux Φ confined inside a small solenoid and a ring of electric charge q around it [2], as illustrated in Figure 18.2. If one particle encircles the other, then its charge q goes around the flux Φ, and vice versa. Due to the celebrated Aharonov-

Bohm effect [10], the wave function of the system will acquire a phase factor $U = e^{2iq\Phi}$, even if there is no direct interaction between the particles. Since all the magnetic flux is confined to the solenoid, this phase factor does not depend in the local details of the path. It depends only on the number of times one particle circulates the other, which makes it topological in nature. These observations imply that if

$$2q\Phi \neq 2\pi n, \qquad n = 1, 2, \dots \ , \tag{18.3}$$

then the wave function will evolve exactly in the same way as a system of two Abelian anyons described by the exchange operator $R = U^{1/2}$. This condition is satisfied if either the charge or the flux is *fractionalised*, i.e., $q \neq n$ (in units of the electron charge e) or $\Phi \neq 2\pi n$, respectively.

The picture of anyons as fractional charge-flux composites, while being a toy model, demonstrates the intimate connection between anyons and fractionalisation [2]. Indeed, as anyons emerge always as localised quasiparticle collective states of some more fundamental particles (fermions, bosons or spins), one can also associate them a charge, which the theory predicts to be fractionalised. This has been experimentally verified [69], for instance, for the charge $q = 1/3$ Abelian anyons in the Laughlin fractional quantum Hall state [6]. One should keep in mind though that while this picture provides a simple mechanism for the origin of anyonic statistics, in real many-body systems anyons are not in general composites of physical flux and charge. Instead, the charge and flux of the anyons emerge due to an effective gauge theory that describes the low energy behaviour of the model [11, 17]. With these simple insights in mind, we now proceed to describe the structure of actual anyon models.

18.2.2 Ising anyons

The principles underlying topological quantum fields theories allow for many different self-consistent anyon models [32]. We illustrate their structure using one of the simplest ones – the Ising anyon model. This anyonic model describes the low energy physics of topologically ordered phases that support localised Majorana modes. It is currently the most topical of non-Abelian anyon models due to its likely relevance to numerous research directions, including p-wave superconductors [21], the fractional quantum Hall states at filling fraction 5/2 [24], spin lattice models [22] and heterostructures of topological insulators / spin-orbit coupled semiconductors in the proximity of a normal s-wave superconductor [46, 47, 50, 54, 55]. Note that in general there are several different anyon models that all can be described in terms of Majorana modes, but which have slightly different statistical properties [22]. We will not be concerned with this subtlety though and will use Ising anyons and Majorana modes interchangeably throughout this review.

To specify an anyon model, one needs to list what kinds of particles exist in the theory (anyon types), how they behave when combined (fusion rules) and how the system evolves when the particles are exchanged (topological spin of

the particles, or equivalently, the braid matrices). These are most conveniently presented in terms diagrams of the worldlines of anyons as they are exchanged or fused, as illustrated Figure 18.3. Such diagrams also vividly capture the topological nature of such processes – two diagrams that can be continuously deformed into each other will correspond to the same evolution of the system. The worldline diagrams can also be extended to account for the topological spins h of the anyons by replacing the lines with ribbons. As illustrated in Figure 18.4, a rotation of particle around itself by 2π can be conveniently represented by a twist in the ribbon. This twist results in the overall phase factor of $e^{2i\pi h}$. By spin-statistics theorem the evolution of the particle twist relates to the exchange evolution R [45], as we shall see later on.

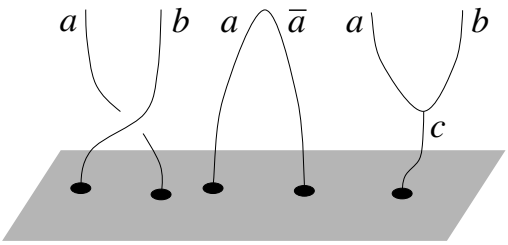


Figure 18.3: Worldline representation of anyon dynamics in $2+1$ dimensions with time flowing downwards. *Left:* A counter-clockwise exchange of two particles, a and b, is depicted in terms of braided worldlines. *Middle:* A pair-creation or annihilation of particle a with its antiparticle $\bar{a}$ corresponds to two worldlines starting or ending at the same position, respectively. *Right:* Fusion of particles a and b that gives particle c (a and b are in the fusion channel c) corresponds to a Y-junction of the worldlines.

The Ising anyon model consists of three particles: 1 (vacuum), ψ (fermion) and σ (anyon). The statistics follow from their topological spins that are given by $h_1 = 0$, $h_\psi = 1/2$ and $h_\sigma = 1/16$, respectively. The fractional spin of the σ particle makes it an anyon, but it does not specify whether it is an Abelian or a non-Abelian one. This information is encoded in the fusion rules

$$\begin{gathered} 1 \times 1 = 1, \quad 1 \times \psi = \psi, \quad 1 \times \sigma = \sigma, \\ \psi \times \psi = 1, \quad \psi \times \sigma = \sigma, \\ \sigma \times \sigma = 1 + \psi. \end{gathered}$$

As a vacuum label 1 fuses trivial with the two non-trivial particles. The fusion rule $\psi \times \psi = 1$ implies that when brought together two fermions behave like there is no particle, while $\psi \times \sigma = \sigma$ implies that ψ with a σ is indistinguishable from a single σ. The non-Abelian nature of the σ particles is encoded in the last fusion rule which says that two of them can behave either as the vacuum or as a fermion. In other words, for two well separated σ particles there is a non-local degree of freedom associated with their *fusion channel*. It is this property that makes them attractive for quantum computing as we discuss below.

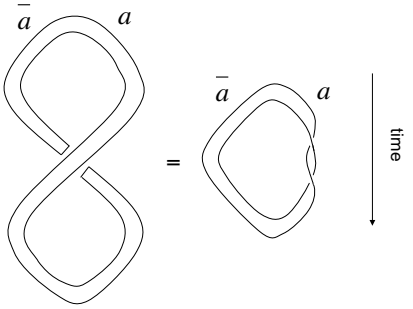


Figure 18.4: The spin-statistics theorem for anyons. Consider a process where an anyon a and its anti-particle $\bar{a}$ are first created from vacuum, then exchanged and finally fused back to vacuum. The process is topologically equivalent to pair creation -annihilation process with one of the anyons rotated by 2π at the intermediate step (the world-ribbon is twisted, which can be nicely verified with a belt). Due to the spin h of the anyon a, such twist is equivalent to overall phase factor of $e^{2\pi i h}$. Thus we arrive at the spin statistics theorem that related exchange statistics of two particles to their spins [45].

Physically, the fusion rules can be understood, for instance, in the context of a topological p-wave superconductor [21]. There, the vacuum 1 is a condensate of Cooper pairs. The fermions ψ are Bogoliubov quasiparticles that can pair into a Cooper pair and thus vanish into the vacuum. The σ particles, on the other hand, correspond to vortices that bind localised Majorana modes. As we will explain below, a Majorana mode corresponds to a "half" of a complex fermion. A pair of such vortices carries thus a single non-local fermion mode, the ψ particle, that can be either unoccupied (fusion channel $\sigma \times \sigma \to 1$) or occupied (fusion channel $\sigma \times \sigma \to \psi$).

18.2.3 The Fusion space

The non-Abelian fusion rule for the σ particles implies that there is a two dimensional Hilbert space associated with a pair of them. The basis in this Hilbert space can be associated with the two fusion channels and denoted by

$$\{|\, \sigma \times \sigma \to 1\rangle\, , |\, \sigma \times \sigma \to \psi\rangle\}. \tag{18.4}$$

However, the fusion rule for two ψ particles implies that the Ising anyon model only conserves ψ parity, which means that the two states associated with the pair of σ particles also belong to different parity sectors. In order to have a non-trivial fusion space in the same parity sector, one needs to consider at least four σ particles. Then in the even parity sector, consisting of fusion channels that

give either 1 or an even number of ψ's, the basis can be given by the states

$$\begin{aligned} |\,0\rangle &\equiv |\,(\sigma\times\sigma)\times(\sigma\times\sigma)\to 1\times 1=1\rangle\,, \\ |\,1\rangle &\equiv |\,(\sigma\times\sigma)\times(\sigma\times\sigma)\to \psi\times\psi=1\rangle\,. \end{aligned} \tag{18.5}$$

These correspond to both σ pairs fusing independently either to 1 or to ψ. We have named these states as $|\,0\rangle$ and $|\,1\rangle$, because in the next section they are employed as the logical states of a qubit.

Of course, these are not the only possible sequences to make all the particles fuse to 1. We could have chosen either of the fusion paths (parenthesis denotes the order of fusing the particles)

$$[\sigma\times(\sigma\times\sigma)]\times\sigma \;\;=\;\; [\sigma\times(1 \text{ or } \psi)]\times\sigma = [\sigma]\times\sigma\to 1, \tag{18.6}$$

that differ by the intermediate step yielding either 1 or ψ. Let us denote the two states associated with these fusion paths as $|\,\bar{0}\rangle$ and $|\,\bar{1}\rangle$, anticipating the connection to a different basis of logical states. There must then exist a unitary relating these two choices of basis, as diagrammatically illustrated in Figure 18.5. This can be obtained by solving the so called *pentagon equations* that are consistency conditions for the fusion rules [22]. For the Ising anyon model one finds that $|\,0\rangle = F\,|\,\bar{0}\rangle$ and $|\,1\rangle = F\,|\,\bar{1}\rangle$, where the F-matrix is given by

$$F=\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}. \tag{18.7}$$

This corresponds to $|\,\bar{0}\rangle = (|\,0\rangle + |\,1\rangle)/\sqrt{2}$ and $|\,\bar{1}\rangle = (|\,0\rangle - |\,1\rangle)/\sqrt{2}$. In other words, if the fusion outcome is completely fixed in one fusion order, then it is completely random in the other and vice versa. Thus the different fusion orders correspond to different bases exactly as the basis for a qubit could be chosen along the z-axis as $\{|\,0\rangle\,,|\,1\rangle\}$ or along x-axis as $\{|\,\bar{0}\rangle\,,|\,\bar{1}\rangle\}$. As the different basis states correspond to different pairs of σ particles fusing into 1 or ψ, detecting the fusion outcomes provides a natural way to perform measurements on different bases of the fusion space.

The structure of the fusion space for more σ's generalises in a straightforward manner. For $2N$ particles the dimension of the fusion space in each parity sector is given by 2^{N-1} with the bases associated with the different fusion paths. As the dimension of the Hilbert space doubles for every added σ pair, one may define the dimensional contribution per particle as $d_\sigma = \sqrt{2}$. This quantity is referred to as the *quantum dimension* and it is larger than unity only for non-Abelian anyons. For Abelian anyons, as well as for bosons and fermions, the quantum dimension is always $d = 1$. One may also define the total quantum dimension of an anyon model as

$$D=\sqrt{\sum_a d_a^2}, \tag{18.8}$$

where a runs over all distinct particle types of the theory. For Ising anyon model one obtains $D_{\text{Ising}} = 2$. It is an important quantity since it appears as a constant

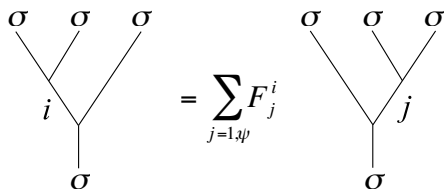


Figure 18.5: Every fusion diagram corresponds to a state in the fusion space. When the particles are arranged on a line, the basis can be chosen to coincide with different fusion orders. The corresponding states must be related by unitary, which called the F-matrix. In the two dimensional fusion space associated with three σ particles in the σ sector (equivalent to four σ particles in the global vacuum sector), the two states correspond to fusing first the two particles on the left or on the right.

term in the entanglement entropy of topologically ordered states [18, 19]. As different anyons models can have the same total quantum dimension, it is not a unique characteristic of a particular anyon model. It is still an important concept though, because it can tell whether a given state is topologically ordered and provide partial information about the nature of the state.

18.2.4 Braiding evolutions

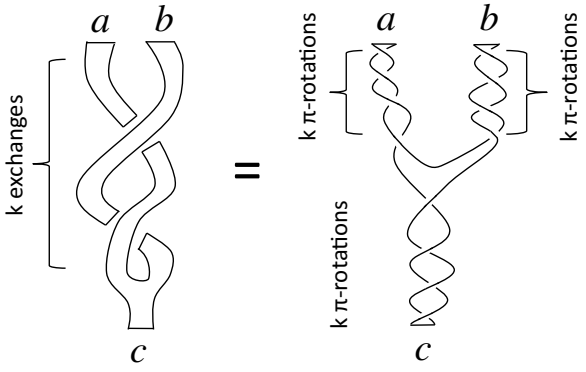


Figure 18.6: Consider a process where two particles a and b are exchanged k times and then fused to particle c. The corresponding diagram is topologically equivalent to twisting k times each the worldribbon by π, such that a and b are twisted counter-clockwise, while c is twisted clockwise. Thus the effect of braiding can be fully given in terms of the topological spins of the particles.

To evolve a state in the fusion space one needs to move the σ particles around each other. This is in general known as *braiding* the anyons, with the corresponding evolutions depending both on the order and orientation of the way the particles are exchanged. The nature of the evolution dependents on the type of the anyonic model. The corresponding exchange or braid matrices R can be obtained by solving another set of consistency equations known as the *hexagon equations*. However, there is also a simpler way to obtain the braid matrices from the topological spins of the particles.

By the well known spin-statistics theorem, a rotation (twisting) of a particle a with spin h_a by 2π will result in the wave function of the system acquiring an overall phase factor of $\theta_a = \exp(2\pi h_a i)$. Consider then the process depicted in Figure 18.6 where two σ particles are first exchanged twice clockwise and then fused. By a purely topological argument this must be equivalent to twisting first both σ particles counter-clockwise, fusing them and then twisting the resulting 1 or ψ particle clockwise. In other words, when the two σ's fuse to $a = 1, \psi$, there must hold

$$(R^a_{\sigma\sigma})^2 = \frac{\theta_a}{\theta_\sigma \theta_\sigma} = \exp[2\pi i(h_a - 1/8)]. \tag{18.9}$$

If such a braiding was performed on either the left or the right σ pair, then a generic state expressed in the basis (18.5) would evolve according to the matrix

$$R^2 = \begin{pmatrix} (R^1_{\sigma\sigma})^2 & 0 \\ 0 & (R^\psi_{\sigma\sigma})^2 \end{pmatrix} = e^{-i\pi/4} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{18.10}$$

Were the σ's braided counter-clockwise, then the evolution would be described by $(R^\dagger)^2$. We immediately see that if we encoded a qubit in the fusion space associated with four σ particles, such a braiding would implement a phase gate.

Since the basis (18.5) is chosen such that both the left and the right pair both fuse simultaneously either to 1 or to ψ, one may ask what is the nature of the evolution if instead of braiding particles from same pairs, one would braid them from different pairs. The corresponding evolution can be obtained by first using an F-matrix to go to the basis (18.6) where the two middle particles (let's label them 2 and 3) are fused first, braiding them using the R-matrix and then returning back to the original basis using the inverse F-matrix. More precisely, we obtain

$$R^2_{23} = F^{-1} R^2 F = e^{-i\pi/4} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{18.11}$$

In other words, if the four σ's were employed to encode a qubit, this braid would have implemented a logical NOT-gate. Physically, this corresponds to changing the intermediate fusion channels of the two σ pairs, as illustrated in Figure 18.7.

Arbitrary evolutions in the fusion space of $2N$ σ's can be obtained using these simple examples. To work out the effect of braiding any two particles, one must first use some sequence of F-matrices defined in this larger Hilbert space

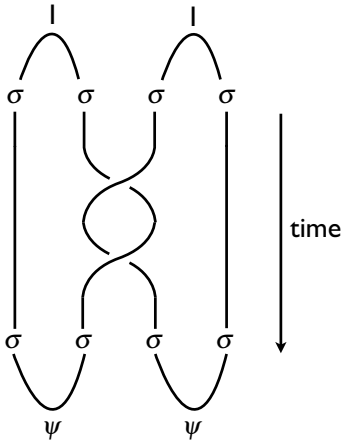


Figure 18.7: When σ particles from two different pairs are braided, the fusion channel of both changes from the 1 channel to the ψ channel. This is allowed, because of the fusion rule $\psi \times \psi = 1$ both states belong to the global vacuum sector.

to move to a basis where the desired particles are directly fused and then apply the braid matrix R. The evolution in the original basis is obtained by applying afterwards the inverse sequence of the F-matrices. For Ising anyons one can implement through braiding any operation that is in the Clifford group [33, 34]. This means that Ising anyons, while being non-Abelian, are not universal for quantum computation by braiding. To overcome this shortcoming, non-topological schemes have been devised to promote their computational power to universality [36, 37]. While the need for such non-topological operations makes the system more susceptible to errors, Ising anyons are still the best candidates we currently have to experimentally test the principles of topological quantum computation.

In the next section we describe in more detail how such computation would proceed. Before doing so, we wrap up this section by discussing how the defining properties of non-Abelian anyons – the fusion space and the braid statistics – would in general manifest themselves in a quantum many-body system.

18.2.5 Anyons in many-body systems: Ground state degeneracy and Berry phases

So far we have discussed non-Abelian anyons at an abstract level. In reality, the anyons appear as collective quasiparticles in strongly correlated many-body systems. There the fusion space appears as degeneracy of the ground state in a fixed quasiparticle number sector. Braiding manifests itself as the evolution in this degenerate manifold when the quasiparticles are transported adiabatically

around each other. Physically this is captured by the non-Abelian Berry phase acquired by the system's wave function under such a process.

Let us consider a system of N Ising anyons σ on the plane, where N is even. The d_σ^N-dimensional fusion space manifests itself as the d_σ^N degenerate many-body ground states given by

$$|\,\Psi_n(z_1, z_2, \ldots, z_N)\rangle\,, \qquad n = 1, 2, \ldots, d_\sigma^N, \tag{18.12}$$

where z_j are the coordinates of the N anyons on the plane. These states are separated from all other states in the spectrum by some energy gap ΔE. This contrasts with the ground state in the presence of Abelian anyons, which is always unique on a plane regardless of the number of anyons present.

To implement braiding of the anyons, one transports them adiabatically around each other. The corresponding evolution of the many-body states (18.12) is given in terms of the geometric Berry phase [12,13]. Let λ be a cyclic path in the anyon coordinates that winds one anyon around another. If one changes the parameters z_j slowly in time compared to the energy gap Δ, then the transport is adiabatic and the system evolves only within the degenerate ground state manifold spanned by the states (18.12). This evolution is in general given by

$$|\,\Psi_n(z_1, z_2, \ldots, z_N)\rangle \to \sum_{m=1}^{d_\sigma^N} \Gamma_{nm}(\lambda)\,|\,\Psi_m(z_1, z_2, \ldots, z_N)\rangle\,, \tag{18.13}$$

where the non-Abelian Berry phase is defined by [14]

$$\Gamma(\lambda) = \mathbf{P} \exp \oint_\lambda \mathbf{A} \cdot d\mathbf{z}. \tag{18.14}$$

Here $\mathbf{P}$ denotes path ordering and the components of the non-Abelian Berry connection are given by

$$(A^j)_{mn} = \langle \Psi_m(z_1, z_2, \ldots, z_N)\,|\,\frac{\partial}{\partial z_j}\,|\,\Psi_n(z_1, z_2, \ldots, z_N)\rangle\,. \tag{18.15}$$

The geometric phases due to the cyclic evolution in anyonic coordinates z_j do not depend on the time it takes to traverse the path λ as long as it is long enough for the evolution to be adiabatic. Nor do they depend on the exact shape of the path. Thus they are topological in nature and capture the non-Abelian statistics of the anyons.

When λ corresponds to an arbitrary sequence of exchanges in a system of four σ particles, then the resulting non-Abelian Berry phase $\Gamma(\lambda)$ evaluates the corresponding product of the R and F matrices. Any more complicated braid in the presence of more σ's could also be evaluated in the same way. Indeed, it has been explicitly demonstrated by Arovas, Schrieffer and Wilczek [15], that the statistics of Abelian anyons for the Laughlin fractional quantum Hall state can be expressed as such a Berry phase. Similar explicit calculation was recently

generalised also to the non-Abelian Ising anyons emerging in the Moore-Read state [68]. The non-Abelian statistics as Berry phases have also been evaluated in the context of p-wave superconductors [16], Moore-Read and Read-Rezayi states [91, 92] Kitaev's honeycomb lattice model [28, 79] as well as bosonic fractional quantum Hall systems [80].

18.3 Quantum computation with anyons

As we discussed above, the characteristic of non-Abelian anyons is the existence of a non-local fusion space $\mathcal{F}$, that is spanned by a basis corresponding to the different possible outcomes and orders of fusing all the anyons. In an ideal situation (infinite spatial anyon separation and infinite energy gap), the states in this space have three very attractive properties from the point of view of quantum computation:

(i) All the states are perfectly degenerate.

(ii) They are indistinguishable by local operations.

(iii) They can be evolved by braiding the anyons, with the evolutions only depending on the topology of the braids.

If this space of states is used as the computational space of a quantum computer, property (i) implies that the encoded information is free of dynamical dephasing, while property (ii) means that it is also protected against any uncorrelated and local noise operator. Property (iii) means that errors could only occur under unlikely non-local noise that would create virtual anyons and propagate them around the encoding ones. However, braiding of the encoding anyons could be carried out in a robust manner by the controlled operations of the computer. Property (iii) implies that the corresponding quantum gates should be error-free and constructible from the F- and R-matrices.

All together these properties mean that quantum computation by such means would heavily suppress errors already at the level of the hardware, with little need for often resource intensive quantum error correction. While these conditions are highly idealised, their desirability led around 2003 to the exotic idea of employing non-Abelian anyons for quantum computation [38, 39]. We now discuss the general steps required to carry out such a topological quantum computation.

18.3.1 The topological qubit and initialisation

The first step of a topological quantum computation is to identify the computational space $\mathcal{H} = \mathbb{C}^{2n}$ of n qubits and initialise it in a given state. A slight complication with the fusion spaces $\mathcal{F}$ is that they do not necessarily admit a tensor product structure. However, usually a subspace $\mathcal{H} \subset \mathcal{F}$ can be identified, where quantum information can be encoded in the usual way.

Because of their particular fusion rules, this subtlety is not a concern when employing Ising anyons. For $2N$ sigma particles the fusion space has the dimensionality $\dim(\mathcal{F}) = d_a^{2N}$, which increases $\mathcal{F}$ by $d_\sigma^2 = 2$ for every additional σ pair. This means that the fusion space of Ising anyons naturally has a tensor product structure $\mathcal{F} = \mathbb{C}^{2N}$. We restrict to the global vacuum sector, which is a natural assumption as the particles are always created pairwise from the vacuum. Then six σ particles can encode two qubits, eight σ's three qubits, and so on. The fusion diagrams corresponding to the basis states are given in Figure 18.8.

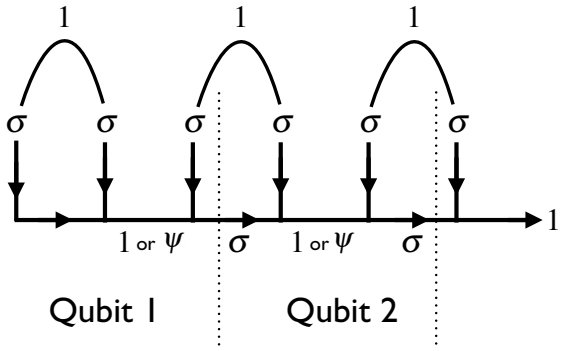


Figure 18.8: The fusion diagram for six Ising anyons, restricted to the global vacuum sector, with the order of fusing proceeding from left to right. Due to the fusion rule $\sigma \times \sigma = 1 + \psi$, the fusion diagram contains two identical, but independent sections with two possible fusion outcomes (1 or ψ). Thus the fusion space of six σ's has tensor product structure $\mathcal{F} = \mathbb{C}^2 \otimes \mathbb{C}^2$. Such Hilbert space can encode two qubits, each associated with three σ particles. When the σ pairs are created from the vacuum as shown, the two qubit system is initialised in the product state $|\,0\rangle\,|\,0\rangle$ that corresponds to both fusion degrees of freedom being fixed to the vacuum channel.

Consider the case where every neighbouring pair of anyons is created from the vacuum and the pairs are kept far from each other. This forces the non-local degree of freedom of each pair to be initialised, independently of each other, in the fusion channel $\sigma \times \sigma \to 1$. Defining the logical states as (18.5), this corresponds to initialising the computational space in the product state $|\,0\rangle\,|\,0\rangle$.

18.3.2 Topological gates and measurements

To perform a computation in the fusion space is equivalent to specifying a braid – a sequence of exchanges of the anyons that corresponds to a sequence of logical gates. Each exchange evolves the computational space. The evolution corresponding to the full braid can be constructed as a product of the R- and

F-matrices. In Figure 18.9 we illustrate a simple braid and we evaluate the corresponding evolution in the fusion space of six Ising anyons. Of course, for any useful algorithm a much larger number of anyons is required and the braids required to approximate desired evolutions are much more complicated [43,44]. Still, any algorithm on a topological quantum computer can be written as such a braid.

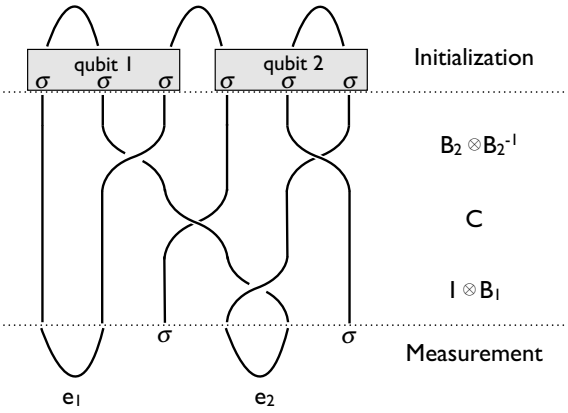


Figure 18.9: An example of a topological quantum computation. Creating three pairs of σ particles from the vacuum initialises the fusion space in the state $|0\rangle|0\rangle$, where the computational basis is defined diagrammatically in Figure 18.8. The shown braid evolves this state according to the unitary $U = (1 \otimes B_1)C(B_1 \otimes B_2^{-1})$, where $B_1 = R$ and $B_2 = F^{-1}RF$. The gate $C = e^{i\pi/4\sigma^x \otimes \sigma^z}$ due to braiding anyons from different qubits can be obtained by solving the Yang-Baxter equations $(B_2 \otimes 1)C(B_2 \otimes 1) = C(B_2 \otimes 1)C$ and $(1 \otimes B_1)C(1 \otimes B_1) = C(1 \otimes B_1)C$. Fusing the two shown particles and observing either 1 or ψ will give outcomes e_1 and e_2, which correspond to measuring both qubits in the z-basis.

The final step of a computation is the read-out. Since the basis states in the fusion space correspond to the possible intermediate fusion outcomes of the non-Abelian anyons, measurements in the computational basis are then equivalent to detecting them. To do this, one needs to bring the particles together. For an encoding based on Ising anyons, a measurement corresponds to bringing two σ particles together and detecting whether they fuse to 1 or ψ. In the basis (18.5) fusing σ's from the same pairs would amount to a measurement in the z-basis. Fusing them from different pairs would amount to measurement in x-basis, as we illustrate with our toy computation in Figure 18.9.

In a nutshell, these are the basic steps of operating a topological quantum computer. How they are carried out precisely depends on the microscopics of the system that supports the anyons. In the next section we describe how a

topological quantum computation could in principle be carried in one microscopic system that supports Ising anyons, the Kitaev's p-wave wire.

18.4 Ising anyons as Majorana modes in a microscopic model

Since the work of Ivanov [23], it has been known that quasiparticles in a two-dimensional system described by Majorana operators would exhibit the braiding statistics of Ising anyons. Formally, Majorana modes are "half" of a complex fermion. By this we mean that if f_i is a fermion operator satisfying $\{f_i^\dagger, f_j\} = \delta_{ij}$, one can always write

$$f_i = \frac{1}{2}\left(\gamma_1 + i\gamma_2\right), \tag{18.16}$$

where $\gamma_i = \gamma_i^\dagger$ are Hermitian Majorana operators satisfying $\{\gamma_i, \gamma_j\} = 2\delta_{ij}$ and $\gamma_i^2 = 1$. If two Majorana modes, γ_1 and γ_2, would exist as localised quasiparticles, then the occupation $f_i^\dagger f_i = 0, 1$ of the complex fermion shared by them would constitute a non-local degree of freedom. This would precisely correspond to the non-local degree of freedom of two σ particles of the Ising anyon model. If the fermionic mode is unoccupied ($f_i^\dagger f_i = 0$), then the two Majorana particles would behave like the vacuum, 1, when brought together. If it is occupied ($f_i^\dagger f_i = 1$), then the fusion of two such quasiparticles would leave behind a fermion that would correspond to the ψ particle. The question then is: could localised Majorana modes actually exist in solid state systems?

18.4.1 Kitaev's toy model for a topological nanowire

Theory suggests that Majorana modes could appear in vortices in exotic superconductors, such as the p-wave superconductor [21], or as quasiparticles in fractional quantum Hall states, like the Moore-Read state proposed for the filling fraction 5/2 [24]. However, as strongly correlated systems, the experimental verification of such materials is still an open question. Building on the physics of p-wave superconductors, Kitaev proposed in 2001 a simplified one-dimensional model where Majorana modes could appear at the ends of a superconducting wire [20]. While not having a clear experimental realisation at the time, this simple toy model provided valuable insights into the mechanisms that give rise to localised Majorana modes.

The remarkable thing about Kitaev's toy model is that it is no longer a toy model. In a seminal work Fu and Kane [46] discovered that topological insulators in proximity of a normal s-wave superconductor could reproduce physics similar to a p-wave superconductor. This soon lead to the prediction that the same could be achieved with an even simpler settings by replacing the topological insulator with a spin-orbit coupled semiconductor [47, 50]. When these ideas were applied to one-dimensional nanowires, it was found that they could host Majorana end states [48, 49]. This prediction was supported by experiments a few years later [40–42, 78]. Such nanowires are no longer the only

potential realisations of Kitaev's toy model. Proposals have been put forward to realise them also in optical lattices [56, 57, 90], cavity arrays [53], magnetic molecules [55], nanoparticles [54] and half-metals [51, 52]. Regardless of the diversity of microscopic realisations, the low energy physics in all of them can always be cast in the form of the toy model, that we now discuss.

Let us assume the model is defined on a chain of L sites and that the fermions in the system are spinless (or equivalently spin polarised). The superconducting Hamiltonian for the system can then be written as

$$H = \sum_{j=1}^{L} \left[-w \left(f_j^\dagger f_{j+1} + f_{j+1}^\dagger f_j \right) - \mu \left(f_j^\dagger f_j - \frac{1}{2} \right) + \left(\Delta f_j f_{j+1} + \Delta^* f_{j+1}^\dagger f_j^\dagger \right) \right], \tag{18.17}$$

where w is the tunnelling amplitude, μ is the chemical potential and $\Delta = |\Delta| e^{i\theta}$ is the superconducting pairing potential. Following [20] we express this Hamiltonian for L complex fermions in terms of $2L$ Majorana operators by employing the decomposition (18.16). Including the superconducting phase θ in their definition, we define $f_j = e^{-i\theta/2}(\gamma_{2j-1} + i\gamma_{2j})/2$. In terms of the Majorana operators the Hamiltonian takes the form

$$H = \frac{i}{2} \sum_{j=1}^{L} \left[-\mu \gamma_{2j-1}\gamma_{2j} + (w + |\Delta|)\gamma_{2j}\gamma_{2j+1} + (-w + |\Delta|)\gamma_{2j-1}\gamma_{2j+2} \right], \tag{18.18}$$

which allows us to explore the phase diagram and the edge properties of the model. There are two special limits in which the ground state can be obtained immediately.

Trivial phase: When the chemical potential term dominates, we can set $|\Delta| = w = 0$. The Hamiltonian is then given by

$$H = \frac{i}{2} \sum_{j=1}^{L} -\mu \gamma_{2j-1}\gamma_{2j} = -\mu \sum_{j=1}^{L} (f_j^\dagger f_j - \frac{1}{2}), \qquad \mu \gg w, |\Delta|. \tag{18.19}$$

The ground state of such a system is trivial. It is given by having a fermion ($f_j^\dagger f_j = 1$) on every site, as illustrated in Figure 18.10.

Topological phase: The other limit is to have the kinetic term be comparable to the pairing potential and dominating over the chemical potential. Setting $w = |\Delta|$ and $\mu = 0$, we obtain the Hamiltonian

$$H = iw \sum_{j=1}^{L} \gamma_{2j}\gamma_{2j+1} = 2w \sum_{j=1}^{L-1} (\tilde{f}_j^\dagger \tilde{f}_j - \frac{1}{2}), \qquad w = |\Delta| \gg \mu, \tag{18.20}$$

where we have defined a new set of fermionic operators by combining the Majoranas as $\tilde{f}_j = e^{-i\theta/2}(\gamma_{2j} + i\gamma_{2j+1})/2$. As illustrated in Figure 18.10, the Majorana operators γ_1 and γ_{2L} completely decouple from the Hamiltonian that now describes interactions only between $L - 1$ complex fermions. The missing

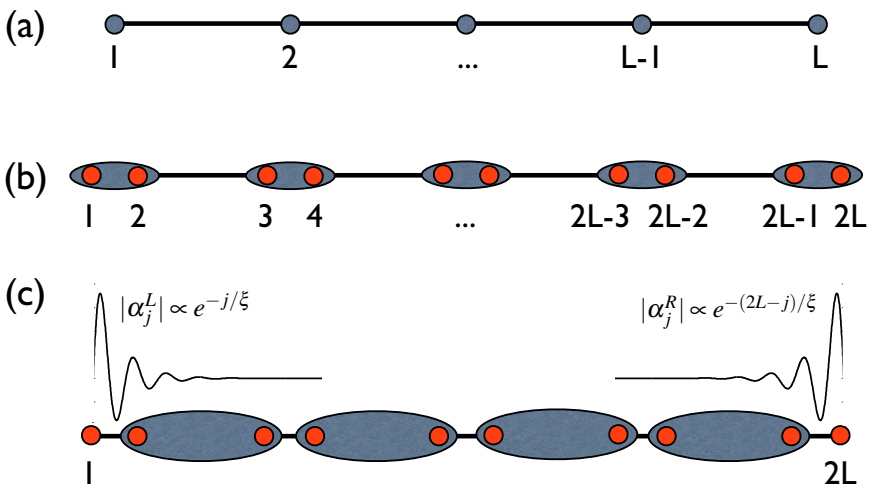


Figure 18.10: Kitaev's toy model for a p-wave paired superconducting wire [20]. (a) The original Hamiltonian (18.17) in terms of complex fermions $f_j^\dagger$ is defined on a one-dimensional lattice of L sites. (b) When each complex fermion operator is decomposed into two Majoranas by writing $f_j = e^{-i\theta/2}(\gamma_{2j-1} + i\gamma_{2j})/2$, the Hamiltonian (18.18) describes free Majorana fermions on a chain of length $2L$ (red sites). When $\mu \gg w, |\Delta|$, the system is in the trivial phase described by the Hamiltonian (18.19), whose unique ground state has all the Majorana modes paired. (c) In the opposing limit, the system is in the topological phase where the Majorana operators γ_1 and γ_{2L} completely decouple from the Hamiltonian (18.20). The ground state is two-fold degenerate. Both states are localised at the wire ends (the figure shows a schematic of the Majorana wave functions (18.21) when $\mu \neq 0$) and correspond to the occupation number of the delocalised fermion $d = (\gamma_1 + i\gamma_{2L})/2$.

fermion can be described by the operator $d = e^{-i\theta/2}(\gamma_1 + i\gamma_{2L})/2$ that is delocalised between the two ends of the wire. Since $[d^\dagger d, H] = 0$, the ground state of this idealised system is two-fold degenerate parameterised by the population of the d mode. Explicit evaluation of the ground state wave functions reveals that they are indeed localised at the ends of the wires. This confirms the qualitative argument for this phase of the system supporting localised Majorana modes. As opposed to the trivial phase in the $w = |\Delta| = 0$ limit, we call this phase a topological phase. In addition to supporting fractionalised excitations, the ground state of the system in this phase can be characterised by a topological invariant [20].

Of course the coupling regimes $w = |\Delta| = 0$ or $\mu = 0$ are limiting cases of idealised models. Were such a model realised in the topological phase, one would hope to achieve a parameter regime where $w \approx |\Delta| \gg \mu$. This is fine as

the system remains in the topological phase that supports localised Majorana modes at wire ends for $2w > |\mu|$. The consequence of being away from the extreme $\mu = 0$ point is that the Majorana modes will be exponentially localised at the wire ends rather than being positioned on a single site. The operators describing them at the left (L) and right (R) ends of the wire take the form

$$\Gamma_{L/R} = \sum_{j=1}^{2L} \alpha_j^{L/R} \gamma_j, \tag{18.21}$$

where the normalised amplitudes will decays as $|\alpha_j^L| \propto e^{-j/\xi}$ and $|\alpha_j^R| \propto e^{-(2L-j)/\xi}$, where $\xi \propto |\Delta|^{-1}$ is the coherence length. This means that for finite wires of length L, the wave functions of the two Majoranas will in general overlap, which in turn results in a finite energy splitting $\Delta E \propto e^{-L/\xi}$ between the two ground states.

The energy splitting described above is the first example how the idealised assumptions of anyon models (in this case perfect degeneracy of the fusion states) are lifted due to the specific microscopics of the system. It serves as an explicit reminder that topologically ordered phases emerge in strongly correlated many-body systems, where the anyon model provides an exact effective low energy description only in an idealised limit of infinite system size and energy gap. The anyons themselves are always collective quasiparticle states of underlying more fundamental particles (electrons in the wire example), that have also microscopic dynamics of their own. In general they are not negligible and endow the anyons with non-topological properties, such as degeneracy-lifting interactions [63] that can even induce topological phase transitions when the anyons form arrays [81–83]. The existince of such interactions has been verified in numerous microscopically distinct systems that support Majorana modes [64–66]. The conditions for the formations of the collective states of Majorana modes, that constitutes the ultimate failure of a topological quantum computer, have also been studied in several recent works [67, 84–89].

18.4.2 The Majorana qubit

Let us view the two localised Majorana modes γ_1 and γ_{2L} as a pair of σ particles corresponding to the Ising anyon model. The occupation of the non-local fermion shared by them is described by the operator $d^\dagger d = (1 + i\gamma_1\gamma_{2L})/2$. For $d^\dagger d = 0$ the two σ's will fuse to vacuum 1, while for $d^\dagger d = 1$ they will leave behind a fermion ψ. These two degenerate ground states in the topological phase can be identified with the two fusion channel states (18.4)

$$\begin{aligned} i\gamma_1\gamma_{2L} \,|\, \sigma \times \sigma \to 1\rangle &= -\,|\, \sigma \times \sigma \to 1\rangle\,, \\ i\gamma_1\gamma_{2L} \,|\, \sigma \times \sigma \to \psi\rangle &= +\,|\, \sigma \times \sigma \to \psi\rangle\,, \end{aligned} \tag{18.22}$$

where

$$|\, \sigma \times \sigma \to \psi\rangle = d^\dagger \,|\, \sigma \times \sigma \to 1\rangle\,.$$

However, these two states cannot form a basis for a qubit, because they belong to different parity sectors of the wire. The fermion parity is described by the operator $\mathcal{P} = \exp(i\pi \sum_j f_j^\dagger f_j)$, which is an exact symmetry of the Hamiltonian (18.17). In the two-fold degenerate ground state manifold this operator acts as $\mathcal{P} = i\gamma_1\gamma_{2L}$, which means that the two ground states cannot mix in a closed system. Like in our discussion in Section 18.2.3, to form a qubit one needs two wires, whose ground state manifold contains two states belonging to the same parity sector. Choosing the even parity sector ($\mathcal{P} = 1$), the computational basis states can be identified with the fusion channel states (18.5) as

$$\begin{aligned} |\,0\rangle &\equiv |\,(\sigma\times\sigma)_1 \times (\sigma\times\sigma)_2 \to 1\times 1 = 1\rangle\,, \\ |\,1\rangle &\equiv |\,(\sigma\times\sigma)_1 \times (\sigma\times\sigma)_2 \to \psi\times\psi = 1\rangle\,, \end{aligned} \tag{18.23}$$

where

$$|\,(\sigma\times\sigma)_1 \times (\sigma\times\sigma)_2 \to \psi\times\psi = 1\rangle = d_1^\dagger d_2^\dagger\,|\,(\sigma\times\sigma)_1 \times (\sigma\times\sigma)_2 \to 1\times 1 = 1\rangle\,.$$

The subscripts in the non-local fermion operators d_i refer to the two wires hosting the anyon pair denoted by $(\sigma\times\sigma)_i$.

18.4.3 Manipulating the Majorana qubit

Let us next consider how braiding, and hence topological quantum gates, could be implemented in the topological nanowire system. As braiding involves moving the anyons around each other, the obvious problem seems to be that the Majorana modes are stuck at the wire ends. This obstacle can be overcome by realising that Majorana modes exist not only at the wire ends, but also at the interfaces between topological ($2w > |\mu|$) and non-topological ($2w < |\mu|$) phases.

One can then envisage a scenario where by external control of the chemical potential part of the wire is made to be in a topological phase and the other is not. By adiabatically tuning the local chemical potential one can move the interface, and thus the trapped Majorana mode, back and forth on the wire. One can then consider a T-junction of these wires, as illustrated in Figure 18.11. This junction enables braiding operations to be executed by manipulating the local couplings on the wires. It has been explicitly shown [25], that when such exchanges are performed by adiabatically tuning the chemical potential, the resulting non-Abelian Berry phase (18.13) coincide with the R-matrix of the Ising anyon model. Correspondingly, the computational states (18.23) transform as $|\,0\rangle \to |\,0\rangle$ and $|\,1\rangle \to i\,|\,1\rangle$. Thus the simple microscopic toy model reproduces all the properties expected by the abstract anyon model.

We should point out that while braiding can be implemented by actually transporting the anyons around each other, it is not the only possibility for implementing the desired evolutions. On way around it is to use measurement-only topological quantum computation, i.e., using fusion measurements to evolve a

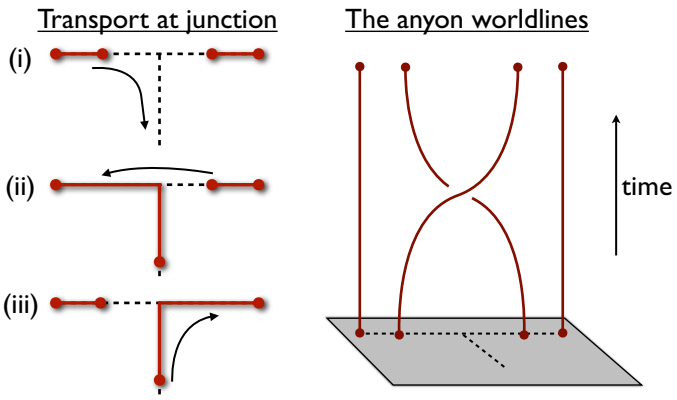


Figure 18.11: Braiding Majoranas at a T-junction [25]. By locally tuning the chemical potential μ along the wires, one can change between the topological phase ($|\mu| < 2w$, solid red lines) and the trivial phase ($|\mu| > 2w$, dashed black lines). The shown sequence (i)-(iii) results in the Majorana modes bound at the domain walls (red dots) to be exchanged such that their worldlines are braided.

subspace of the fusion space [70, 71]. Another option is to control the microscopic interactions to the same end [26, 27, 31].

18.4.4 How protected is the Majorana qubit?

In Section 18.3 we outlined the general steps for operating a topological quantum computer and in this section we have described how those steps could in principle be carried out in a microscopic model. While doing so, we discovered that the ideal properties of anyons, like the exact degeneracy of the non-local fusion states that makes topological quantum computing appealing, are not perfectly manifested in microscopic systems. We discuss now how these more realistic conditions affect the protection provided by topological encoding.

The explicit example we encountered was the lifting by $\Delta E \sim e^{-L/\xi}$ of the degeneracy of the states shared by the localised Majorana modes. This is due to finite wire length L and finite energy gap Δ, which gives rise to a finite coherence length $\xi \sim \Delta^{-1}$. In fact, such exponential degeneracy lifting is expected not to apply only to p-wave wires, but to all microscopic topological phases. This has been explicitly verified for p-wave superconductors [64], fractional quantum Hall states [65] and spin lattice models [66]. It can rarely be ignored as a macroscopic number of anyons is required for quantum computation [44], and thus in any realistic system they must be in proximity of each other. So, what does such degeneracy lifting mean in general for the promised protection provided by the anyons?

In essence, the degeneracy lifting means two things. First, logical states with different energies will dephase with time. While they remain insensitive to

local operations, the Hamiltonian of the system will distinguish between them. Thus topological qubits, like non-topological ones, will decohere with time unless error correction is applied [93]. Second, quantum gates by braiding are no longer exact. Finite energy splitting between the ground states implies that the evolution must be fast enough at the scale of ΔE for the states to appear degenerate, while still being slow enough at the scale of the energy gap Δ not to excite the system [28–30]. In general this balancing between the two energy scales means that there are small errors in the implemented logical operations. If they accumulate and are not error corrected, they will become a source of decoherence. Nevertheless, it is worth noting that the degeneracy lifting can be exponentially suppressed by the distance between interacting anyons, which in principle gives a very powerful way to keep it under control. The catch is that the number of anyons needed to be kept under control for a robust implementation of quantum algorithms ranges from 10^3 Fibonacci anyons to a whopping 10^9 Ising anyons [44]. These numbers, while being rough estimates, vividly illustrate the degree of accuracy required in order not to accumulate errors due to any non-topological microscopics. Alternatively, braiding can be simulated by microscopic control of interaction [27,31] or through measurements [71], but the challenge to make these schemes robust at the energy scales of experiments remains.

These two general properties imply that while topological quantum computation provides much enhanced protection compared to more conventional schemes, it is no panacea by itself. Topological qubits can decohere due to the microscopics, and thus it is expected that any scalable topological quantum computer architecture will be a hybrid employing both the hardware level protection provided by the anyons and the software level protection of quantum error correction [31, 93]. For architectures based on Ising anyons, error correction will be necessary, also because non-topological operations are needed for universal quantum computation. Since only quantum gates that are in the Clifford group can be implemented by braiding, for universality one needs in addition to implement the $\pi/8$-gate. This could be performed, for instance, via allowing the anyons to interact in a controlled manner [36, 37]. The noise in such unprotected operations could then be compensated by using distillation schemes [72].

Like with conventional qubits, another source of decoherence is the uncontrolled coupling to a reservoir. In topologically ordered systems the reservoir could appear in two ways. Either there are additional anyons in the system, i.e., the full fusion space does not coincide with the computational space, or there is a reservoir nearby from where quasiparticles can tunnel into and out of the system. In the p-wave wire the first scenario could occur due to local random disorder in the wire. As has been studied in [61,62], sufficiently strong spatial fluctuations in the chemical potential μ can split the wire into several domains, some of which are in the topological phase ($2w > |\mu|$ locally) and some of which are in the trivial phase ($2w < |\mu|$ locally). The domain walls will also host additional Majorana modes that couple to the wire end Majoranas

encoding the topological qubit. Unless one keeps track of the fusion space associated with them, encoded information will leak into this subspace. It is a general challenge of topological quantum computing schemes to keep precise track of all the quasiparticles in the system and thereby prevent such leakage.

Coupling to an external reservoir of anyons should in general be preventable by simply isolating the topologically ordered system, that only occurs under very precise conditions. However, for Majorana qubits in nanowires or other hybrid architectures this can still pose a problem. Since the Majorana qubit is encoded essentially in the ψ fermion parity of a wire, the qubits are protected only as long as the parity described by the operator $\mathcal{P} = 1\gamma_1\gamma_{2L}$ is a good quantum number. This is in general a strong demand for hybrid realizations of p-wave wires, because in all of them superconductivity is induced to the otherwise semiconducting wire by the means of a proximity to an s-wave superconductor [48,49,55]. If Cooper pairs can tunnel in and out of the wire to this reservoir, why could not also fermionic ψ quasiparticles? Indeed, Majorana qubits have been shown to be susceptible to decoherence of this type [58–60,97].

Finally, a big question is the robustness of topological schemes at finite temperature. In principle, the energy gap of topologically ordered systems does protect the encoded information bt exponentially suppressing thermal fluctuations that would excite virtual anyons and propagate them around the system. However, studies on Abelian anyon-based topological quantum memories suggest that any finite temperature can be a problem [74, 75], if the system does not encode some, so far artificial mechanism that suppresses the spontaneous creation of stray anyons [94, 95]. Experimentally the situation is even more challenging as the energy gaps tend to be small and thus only formidably low temperatures can be tolerated even for short times. Fortunately, recent studies show that Majorana qubits can tolerate small temperatures [96, 98, 99], which should thus not pose a fundamental obstacle.

18.5 Outlook

In this review we have outlined the three general steps that constitute topological quantum computation: (i) To have access to a system supporting non-Abelian anyons, (ii) to be able to adiabatically move them around each other and (iii) to be able measure their fusion channels. We discussed the implementation of these steps in the context of p-wave paired superconducting wires, that have several microscopically distinct experimental realisation, and which support Majorana modes. Considering the recent progress in experimentally identifying robust signatures of Majorana modes in them [40–42,78], these systems are also the current prime candidates to experimentally verify anyonic statistics [25,100]. Anyon based schemes to encode and process quantum information promise hardware level protection, but with their own weaknesses they are no panacea. However, considering the challenges faced by non-topological schemes, overcoming these challenges is a fair price to pay for the robustness that comes with topological quantum computation.

While Majorana modes are likely to be the first anyons whose braid statistics are to be probed experimentally [25, 100], they are insufficient to perform universal quantum computation. Recently, progress along the lines of hybrid schemes for p-wires was made to go beyond Majorana modes. Replacing the spin-orbit coupled semiconductors with edge states of Abelian fractional quantum Hall states, it shown that one can realise parafermion modes that allow for a larger, although still non-universal gate set [101, 102]. Collective states of such generalisations of Majorana modes can, however, in principle give rise to a state that supports the Holy Grail of topological quantum computation – the universal Fibonacci anyons [103]. While beyond current technology, strong hope remains that a scalable fully topological quantum computer can one day see the daylight. A fascinating future awaits.

References

[1] J. M. Leinaas and J. Myrheim, Nuovo Cimento B 37, 1 (1977).

[2] F. Wilczek, Phys. Rev. Lett. 49, 957 (1982).

[3] K. von Klitzing, G. Dorda, and M. Pepper, Phys. Rev. Lett. **45**, 494 (1980).

[4] D. C. Tsui, H. L. Stormer and A. C. Gossard, Phys. Rev. Lett. 48, 1559 (1982).

[5] D. J. Thouless, M. Kohmoto, M. P. Nightingale and M. den Nijs, Phys. Rev. Lett. **49**, 405 (1982).

[6] R. B. Laughlin, Phys. Rev. Lett. 50, 1395 (1983).

[7] M. A. Nielsen and I. L. Chuang, Quantum Computation and Quantum Information, (Cambridge, 2000).

[8] N. Read and E. Rezayi, Phys. Rev. B **59**, 8084 (1999).

[9] C. Nayak, S. H. Simon, A. Stern, M. Freedman and S. Das Sarma, Rev. Mod. Phys. 80, 1083 (2008).

[10] Y. Aharonov and D. Bohm, Phys. Rev. 115, 485 (1959).

[11] Z. Nussinov and G. Ortiz, Proc. Nat. Ac. Sc. USA 106, 16944 (2009).

[12] M. V. Berry, Roc. R. Soc. A 392, 45 (1984).

[13] J. K. Pachos and P. Zanardi, Int. J. Mod. Phys. B 15, 1257 (2001).

[14] F. Wilczek and A. Zee, Phys. Rev. Lett. 52, 2111 (1984).

[15] D. Arovas, J.R. Schrieffer and F. Wilczek, Phys. Rev. Lett 53, 722 (1984).

[16] N. Read, Phys. Rev. B 79, 045308 (2009).

[17] E. Witten, Commun. Math. Phys. 121, 351 (1989).

[18] A. Kitaev and J. Preskill, Phys. Rev. Lett. 96, 110404 (2006).

[19] M. Levin and X.-G. Wen, Phys. Rev. Lett. 96, 110405 (2006).

[20] A. Kitaev, Phys.-Usp. **44**, 131 (2001).

[21] N. Read and D. Green, Phys. Rev. B **61**, 10267 (2000).

[22] A. Kitaev, Ann. Phys. **321**, 2 (2006).

[23] D. A. Ivanov, Phys. Rev. Lett. **86**, 268 (2001).

[24] G. Moore and N. Read, Nucl. Phys. B **360**, 362 (1991).

[25] J. Alicea, Y. Oreg, G. Refael, F. von Oppen, and M.P.A. Fisher, Nat. Phys. **7**, 412 (2011).

[26] J. D. Sau, D. J. Clarke, and S. Tewari, Phys. Rev. B **84**, 094505 (2011).

[27] B. van Heck, A.R. Akhmerov, F. Hassler, M.Burrello, and C.W.J. Beenakker, New J. Phys. **14**, 035019 (2012).

[28] V. Lahtinen and J. K. Pachos, New J. Phys. **11**, 093027 (2009).

[29] M. Cheng, V. Galitski and S. Das Sarma, Phys. Rev. B **84**, 104529 (2011).

[30] I. C. Fulga, B. van Heck, M. Burrello and T. Hyart, arXiv:1308.0244.

[31] T. Hyart, B. van Heck, I. C. Fulga, M. Burrello, A. R. Akhmerov and C. W. J. Beenakker, Phys. Rev. B **88**, 035121 (2013).

[32] E. Rowell, R. Stong and Z. Wang, Comm. Math. Phys. **292**, 343 (2009).

[33] S. Bravyi and A. Kitaev, Ann. Phys. **298**, 210 (2002).

[34] A. Ahlbrecht, L. S. Georgiev and R. F. Werner, Phys. Rev. A **79**, 032311 (2009).

[35] S. Bravyi and A. Kitaev, Phys. Rev. A **71**, 022316 (2005).

[36] S. Bravyi, Phys. Rev. A **73**, 042313 (2006).

[37] P. Bonderson, D. J. Clarke, C. Nayak and K. Shtengel, Phys. Rev. Lett. **104**, 180505 (2010).

[38] M. H. Freedman, A. Kitaev, M. J. Larsen, and Z. Wang Bull. Amer. Math. Soc. **40**, 31 (2003).

[39] A. Kitaev, Annals Phys. **303**, 2 (2003).

[40] V. Mourik, K. Zuo, S. M. Frolov, S. R. Plissard, E. P. A. M. Bakkers, and L. P. Kouwenhoven, Science **336**, 1003 (2012).

[41] A. Das, Y. Ronen, Y. Most, Y. Oreg, M. Heiblum and H. Shtrikman, Nature Physics **8**, 887-895 (2012).

[42] H. O. H. Churchill, V. Fatemi, K. Grove-Rasmussen, M. T. Deng, P. Caroff, H. Q. Xu, C. M. Marcus, Phys. Rev. B **87**, 241401(R) (2013).

[43] N. E. Bonesteel, L. Hormozi, G. Zikos, S. H. Simon, Phys. Rev. Lett. **95**, 140503 (2005).

[44] M. Baraban, N. E. Bonesteel and S. H. Simon, Phys. Rev. A **81**, 062317 (2010).

[45] D. Finkelstein and J. Rubinstein, J. Math. Phys. **9**, 1762 (1968).

[46] L. Fu and C. L. Kane, Phys. Rev. Lett. **100**, 096407 (2008).

[47] J. D. Sau, R. M. Lutchyn, S. Tewari and S. Das Sarma, Phys. Rev. Lett. **104**, 040502 (2010).

[48] Y. Oreg, G. Refael, and F. von Oppen, Phys. Rev. Lett. **105**, 177002 (2010).

[49] R. M. Lutchyn, J. D. Sau, and S. Das Sarma, Phys. Rev. Lett. **105**, 077001 (2010).

[50] J. Alicea, Phys. Rev. B **81**, 125318 (2010).

[51] M. Duckheim and P. W. Brouwer Phys. Rev. B **83**, 054513 (2011).

[52] S.- B. Chung, H.-J. Zhang, X.-L. Qi, and S.-C. Zhang, Phys. Rev. B **84**, 060510(R) (2011).

[53] C.-E. Bardyn and A. Imamoglu, Phys. Rev. Lett. **109**, 253606 (2012).

[54] T.-P. Choy, J. M. Edge, A. R. Akhmerov, C. W. J. Beenakker Phys. Rev. B **84**, 195442 (2011).

[55] S. Nadj-Perge, I. K. Drozdov, B. A. Bernevig, A. Yazdani, Phys. Rev. B **88**, 020407(R) (2013).

[56] L. Jiang, T. Kitagawa, J. Alicea, A. R. Akhmerov, D. Pekker, G. Refael, J. I. Cirac, E. Demler, M. D. Lukin and P. Zoller, Phys. Rev. Lett. **106**, 220402 (2011).

[57] S. Diehl, E. Rico, M.A. Baranov, P. Zoller, Nat. Phys. **7**, 971 (2011).

[58] J. C. Budich, S. Walter and B. Trauzettel, Phys. Rev. B **85**, 121405 (2012).

[59] D. Rainis and D. Loss, Phys. Rev. B **85**, 174533 (2012).

[60] M. J. Schmidt, D. Rainis, D. Loss, Phys. Rev. B **86**, 085414 (2012).

[61] P. W. Brouwer, M. Duckheim, A. Romito, and F. von Oppen, Phys. Rev. Lett. **107**, 196804 (2011).

[62] A. M. Lobos, R. M. Lutchyn, S. Das Sarma, Phys. Rev. Lett. **109**, 146403 (2012).

[63] P. Bonderson, Phys. Rev. Lett. **103**, 110403 (2009).

[64] M. Cheng, R.M. Lutchyn, V. Galitski and S. Das Sarma, Phys. Rev. Lett. **103**, 107001 (2009).

[65] M. Baraban, G. Zikos, N. Bonesteel and S.H. Simon, Phys. Rev. Lett. **103**, 076801 (2009).

[66] V. Lahtinen, New. J. Phys. **13**, 075009 (2011).

[67] V. Lahtinen, A.W.W. Ludwig, J.K. Pachos and S. Trebst, Phys. Rev. B **86**, 075115 (2012).

[68] P. Bonderson, V. Gurarie and C. Nayak, Phys. Rev. B **83**, 075303 (2011).

[69] F. E. Camino, W. Zhou and V. J. Goldman, Phys. Rev. Lett. **98**, 076805 (2007).

[70] P. Bonderson, M. Freedman and C. Nayak, Phys. Rev. Lett. **101**, 010501 (2008).

[71] P. Bonderson, Phys. Rev. B **87**, 035113 (2013).

[72] H. Bombin and M. A. Martin-Delgado, Phys. Rev. Lett. **97**, 180501 (2006).

[73] R. L. Willett, L. N. Pfeiffer and K. W. West, Proc. Nat. Acad. Sci, **106**, 8853 (2009).

[74] C. Castelnovo and C. Chamon, Phys. Rev. B 76, 184442 (2007).

[75] S. Iblisdir, D. Perez-Garcia, M. Aguado and J. K. Pachos, Thermal States of Anyonic Systems, Nucl. Phys. B **829**, 401 (2010).

[76] J. R. Wootton, J. Mod. Opt. **59**, 1717 (2012).

[77] Benjamin J. Brown, Daniel Loss, Jiannis K. Pachos, Chris N. Self, James R. Wootton, arXiv:1411.6643.

[78] Stevan Nadj-Perge, Ilya K. Drozdov, Jian Li, Hua Chen, Sangjun Jeon, Jungpil Seo, Allan H. MacDonald, B. Andrei Bernevig, Ali Yazdani, Science **346**, 602 (2014).

[79] A. T. Bolukbasi and J. Vala, New J. Phys. **14**, 045007 (2012).

[80] E. Kapit, P. Ginsparg and E. Mueller, Phys. Rev. Lett. **108**, 066802 (2012).

[81] Adrian Feiguin, Simon Trebst, Andreas W. W. Ludwig, Matthias Troyer, Alexei Kitaev, Zhenghan Wang, Michael H. Freedman, Phys. Rev. Lett. **98**, 160409 (2007).

[82] C. Gils, E. Ardonne, S. Trebst, A. W. W. Ludwig, M. Troyer, Z. Wang, Phys. Rev. Lett. **103**, 070401 (2009).

[83] A. W.W. Ludwig, D. Poilblanc, S. Trebst, M. Troyer, New J.Phys **13**, 045014 (2011).

[84] G. Kells, D. Meidan, P. W. Brouwer, Phys. Rev. B **85**, 060507(R) (2012).

[85] V. Lahtinen, A. W. W. Ludwig, S. Trebst, Phys. Rev. B **89**, 085121 (2014).

[86] G. Kells, V. Lahtinen, J. Vala, Phys. Rev. B **89**, 075122 (2014).

[87] D. Wang, Z. Huang, C. Wu, Phys. Rev. B **89**, 174510 (2014).

[88] J. M. Murray, O. Vafek, Phys. Rev. B **92**, 134520 (2015).

[89] T. Liu, M. Franz, Phys. Rev. B **92**, 134519 (2015).

[90] A. Bhler, N. Lang, C. V. Kraus, G. Mller, S. D. Huber, H. P. Bchler, Nat. Comm. **5**, 4504 (2014).

[91] Y. Tserkovnyak and S. H. Simon, Phys. Rev. Lett. **90** 016802 (2003).

[92] Y.-L. Wu, B. Estienne, N. Regnault, B. Andrei Bernevig, Phys. Rev. Lett. **113**, 116801 (2014).

[93] J. R. Wootton, J. Burri, S. Iblisdir, D. Loss, Phys. Rev. X **4**, 011051 (2014).

[94] A. Hamma, C. Castelnovo and C. Chamon, Phys. Rev. B **79**, 245122 (2009).

[95] F. L. Pedrocchi, A. Hutter, J. R. Wootton, Daniel Loss, Phys. Rev. A **88**, 062313 (2013).

[96] M. Cheng, R. M. Lutchyn, S. Das Sarma, Phys. Rev. B **85**, 165124 (2012).

[97] S.-H. Ho, S.-P. Chao, C.-H. Chou, F.-L. Lin, New J. Phys. **16**, 113062 (2014).

[98] L. Mazza, M. Rizzi, M. D. Lukin, and J. I. Cirac, Phys. Rev. B **88**, 205142 (2013).

[99] F. Lin and V. W. Scarola, Phys. Rev. Lett. **111**, 220401 (2013).

[100] J. Li, T. Neupert, B. A. Bernevig and A. Yazdani, arxiv:1404.4058.

[101] D. J. Clarke, J. Alicea, K. Shtengel, Nature Commun. **4**, 1348 (2013).

[102] N. H. Lindner, E. Berg, G. Refael, A. Stern, Phys. Rev. X **2**, 041002 (2012).

[103] R. S. K. Mong, D. J. Clarke, J. Alicea, Netanel H. Lindner, P. Fendley, C. Nayak, Y. Oreg, A. Stern, E. Berg, K. Shtengel, M. P. A. Fisher, Phys. Rev. X **4**, 011036 (2014).

Index

S. M. Bhattacharjee et al. (eds.), *Topology and Condensed Matter Physics*, Texts and Readings in Physical Sciences 19,
https://doi.org/10.1007/978-981-10-6841-6