

# Unsupervised Classification of Topological Insulators

A Thesis

submitted to

Indian Institute of Science Education and Research Pune

in partial fulfillment of the requirements for the

BS-MS Dual Degree Programme

by

Swarnavo Basu



Indian Institute of Science Education and Research Pune

Dr. Homi Bhabha Road,  
Pashan, Pune 411008, INDIA.

May, 2022

Supervisor: Dr. Mathias S. Scheurer

Institute for Theoretical Physics, University of Innsbruck

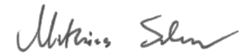
© Swarnavo Basu 2022

All rights reserved



# Certificate

This is to certify that this dissertation entitled “Unsupervised Classification of Topological Insulators” towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Swarnavo Basu, at Indian Institute of Science Education and Research under the supervision of Dr. Mathias S. Scheurer, Assistant Professor, Institute for Theoretical Physics, University of Innsbruck, during the academic year 2021-2022.



Dr. Mathias S. Scheurer

Committee:

Dr. Mathias S. Scheurer

Dr. Sreejith G. J.

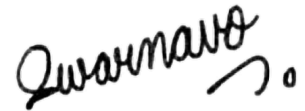


This thesis is dedicated to my parents and my late grandmother.



# Declaration

I hereby declare that the matter embodied in the report entitled “Unsupervised Classification of Topological Insulators”, are the results of the work carried out by me at the Institute for Theoretical Physics, University of Innsbruck, Indian Institute of Science Education and Research, Pune, under the supervision of Dr. Mathias S. Scheurer and the same has not been submitted elsewhere for any other degree.

A handwritten signature in black ink, reading "Swarnavo Basu" with a stylized flourish at the end.

Swarnavo Basu





# Acknowledgements

I would like to express my gratitude towards Professor Mathias S. Scheurer for giving me the opportunity to work on a newly developing field. He patiently guided me throughout the project and was available for help whenever I got stuck. I would also like to thank him for giving me the opportunity to give a virtual talk on my work at the University of Innsbruck and attend their condensed matter seminars. Despite the restrictions imposed by the pandemic, my fifth year was a great experience and I got to learn a lot from him during the whole process.

I would like to thank my friends. I had a memorable five year journey at IISER which would not have been possible without them. I got to learn quite a lot from them and was exposed to different perspectives, not only in academics, but life in general.

Lastly, I would like to express my gratitude towards my parents. They have been a constant pillar of support throughout my life. They have always encouraged me to follow my interests and have always been available for guidance, and I will forever be grateful to them.



# Abstract

Topological insulators are gapped states of quantum matter which cannot be adiabatically connected to conventional insulators [1] and are characterized, among other aspects, by gapless boundary modes. They constitute an active field of research and have the potential to be used for new technologies such as electronic devices with low power consumption, topological quantum computers, etc. While theorists have performed systematic classifications of such topological phases (for example, [2]), the identification of the precise topological characteristics of a given Hamiltonian can still be challenging. Recently, machine learning algorithms have also been employed for “learning” phases and phase transitions in condensed matter systems [3]. However, most of these use supervised algorithms which require a-priori labelling of data sets. Our focus is on unsupervised algorithms, which have recently been used for the classification of topological phases [4, 5]. The advantage of unsupervised algorithms is that they don’t require labelled data. Such algorithms try to extract patterns from the data sets and classify the data into groups. This opens up the possibility of the algorithm to discover new phases or uncover patterns which haven’t been observed before, from raw data. We use the Diffusion Map algorithm to perform topological classification of 1-D 2 band Hamiltonians, and derive the corresponding auxiliary quantum many-body Hamiltonian for this problem.



# Contents

<b>Abstract</b>	<b>xi</b>
<b>1 The Su-Schrieffer–Heeger Model</b>	<b>5</b>
1.1 Description of the Model . . . . .	5
1.2 Bulk Hamiltonian . . . . .	7
1.3 Including the boundary in the SSH model . . . . .	10
<b>2 The Diffusion Map Algorithm</b>	<b>13</b>
2.1 Defining a diffusion process on the data set . . . . .	13
2.2 The Diffusion Distance . . . . .	15
2.3 The Auxiliary Quantum Problem . . . . .	16
<b>3 Unsupervised Classification of 1-D 2-band Hamiltonians</b>	<b>19</b>
3.1 Our Model . . . . .	19
3.2 Choice of Kernel . . . . .	20
3.3 Sampling of the Hamiltonians . . . . .	21
3.4 Discussion . . . . .	23
<b>4 Derivation of the Auxiliary Hamiltonian</b>	<b>25</b>

4.1	The general scheme . . . . .	25
4.2	Simplifying the probability distribution . . . . .	26
4.3	Deriving $E_{\text{eff}}$ . . . . .	28
4.4	Calculating the expression for potential energy . . . . .	29
4.5	Expression for $\mathcal{H}_{\text{eff}}$ . . . . .	33
4.6	Ground state check for $\mathcal{H}_{\text{eff}}$ . . . . .	33
4.7	Hamiltonian in the Continuum Limit . . . . .	35
4.8	Discussion and Outlook . . . . .	38

# List of Figures

1	Winding Numbers . . . . .	2
1.1	Diagrammatic representation of the SSH model . . . . .	5
1.2	Dispersion relation of SSH bulk Hamiltonian . . . . .	8
1.3	Trajectory of the end-point of $\vec{g}(k)$ . . . . .	9
1.4	SSH model energy spectrum and wavefunctions . . . . .	11
2.1	Construction of the diffusion process on data set . . . . .	14
3.1	Results for 1-D 2 band model . . . . .	22





# Introduction

Since the discovery of the major types of forces that are present in the nature and the laws that determine dynamics of particles which are acted upon with those forces, one of the next important steps has been to determine how a system of a large number of particles behaves in the presence of relevant forces. This problem is both important and challenging because in nature we mostly deal with systems that have a rather large number of particles while on the other hand, we know that a problem as simple as the three-body problem is quite challenging to solve analytically. Besides, contrary to what one might expect, a collection of particles has a lot of interesting physics (what is called emergent phenomena) which can be quite different from the behaviour of the individual particles [6]. Thus, this problem has been one of the centres of attraction among physicists and substantial progress has been made in statistical physics and condensed matter physics.

One of the unique characteristics of such many-particle systems is the existence of the so-called various “phases” [7], [8]. A ubiquitous example is water, which can exist in three phases - solid, liquid and steam - based on its temperature, pressure and volume. Phases can be defined as regions in some parameter space of a many-body system within which the system has similar physical properties. Up until a few decades ago, a useful way of identifying the phase of a system using the Ginzburg-Landau theory of phase transitions [9]. The basis of that theory was a local order parameter whose value could determine the phase of the system.

About four decades ago, it was shown that in certain materials (aptly named topological insulators), the phase of a many-body system also depends on the topology [1], [10]. The study of such systems heavily draws concepts from the branch of mathematics called topology. The interest in physics of such materials has been boosted by the discovery of their potential to be used for new technologies such as electronic devices with low power

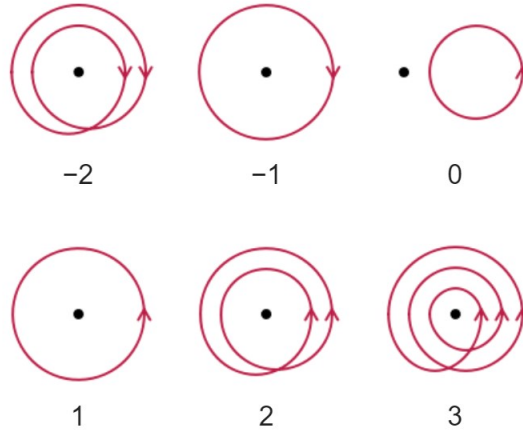


Figure 1: We can topologically classify closed, oriented loops on a two dimensional plane using their winding number. The winding number is defined as the number of loops around the origin that the curve completes. A loop traversed in counter-clockwise direction is considered to be +1 and vice versa. [Image source: wikipedia.org]

consumption, topological quantum computers, etc.

Classification of phases, both conventional and topological, is an important task and various methods have been employed to achieve that. However, in this thesis, we are concerned with the classification of topological phases. The challenge with topological phases is that unlike conventional phases (which depend on a local order parameter), they depend on the geometry (which is a global property of the system). Thus, identification of topological phases involves a global scanning which is a relatively cumbersome task. Usually, such topological phases are identified by a quantity known as the topological invariant. A topological invariant classifies all those geometries which can be achieved via continuous operations (i.e. vaguely speaking, stretching and bending, but not tearing). An example of a topological invariant is the winding number (Fig. 1).

Even though the order parameter or the topological invariant can be calculated analytically, an automated calculation (and subsequent identification of the phase) is both convenient as well as more practical. Towards that end, machine learning algorithms have gained popularity and they have been shown to perform well for this purpose ([3], [11], [12], [13]). However, most of these algorithms use supervised machine learning. One of the problems with supervised learning is that they require a large number of data sets for which the answers are already known. The model is first “trained” using this data set, which is often a

computationally expensive process. One can also argue that albeit such training makes the model very good at predicting answers to similar cases, their performance in a somewhat different context might be questionable. This is where unsupervised algorithms come into the picture.

In principle, unsupervised algorithms do not need any sort of training and they work solely on the principle of identifying patterns in the data and classifying the whole data into groups with similar patterns. Thus, for a data set that contains Hamiltonians which show topological phases, if one can somehow encode the process of recognising topological equivalence in a unsupervised machine learning algorithm then that algorithm should be able to find out the different topological phases of the physical system described by the Hamiltonian. This is precisely what was achieved very recently in [4], which was followed by [5]. Such unsupervised algorithms are usually much lesser computationally expensive (when compared to supervised algorithms and deep neural networks) and they also have the potential of discovering new phases without prior training.

In this thesis, we first briefly study how topology can affect physical properties of a many-body system by using the example of the Su-Schrieffer-Heeger model. We look at the concept of winding number as a topological invariant. We then study the Diffusion Map algorithm, which is the machine learning algorithm that was used in [4], and [5] for unsupervised classification of topological phases by calculating the winding number. We also see how the Diffusion Map (for any general setting) can be mapped to a quantum mechanical problem. Finally, using this background knowledge, we apply the Diffusion Map algorithm to classify one-dimensional two-band Hamiltonians that are drawn from a given probability distribution. We then derive the corresponding quantum many-body Hamiltonian for the Diffusion Map used in our case.



# Chapter 1

## The Su-Schrieffer–Heeger Model

In this chapter, we will look at a very simple one-dimensional model that exhibits topological phases, characterised by the existence/absence of zero-energy edge states. We will also see how the topological invariant acts as the identifier of the different topological phases. This is important as this is the quantity that we calculate using the machine learning algorithm. The presentation in this chapter has been inspired from [14] and [15].

### 1.1 Description of the Model

We consider a one-dimensional chain with  $N$  unit cells, with each unit cell having two sites (sub-lattice points  $X$  and  $Y$ ). We neglect interactions between the electrons and consider them to be spin-less. The Su-Schrieffer–Heeger Model (SSH model) [16] describes hopping

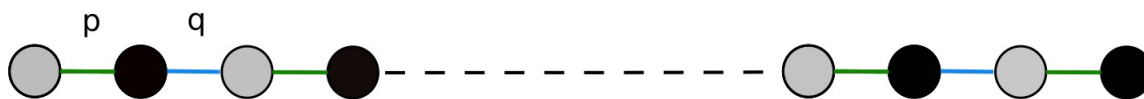


Figure 1.1: Diagrammatic representation of the SSH model. The grey circles represent sub-lattice  $X$  and black circles represent sub-lattice  $Y$ . Both of them together constitute one unit cell, and the chain is constructed by repeating the unit cell one after the other.  $p$  and  $q$  are the intra-cell and inter-cell hopping amplitudes, respectively.

of electrons in such a one-dimensional chain with varying hopping amplitudes (as shown in the Fig. 1.1). The only possible dynamics in such a system will be hopping of electrons between neighbouring sites. The Hamiltonian of the SSH model is thus given by:

$$\hat{H} = p \sum_{n=1}^N (|n, X\rangle \langle n, Y| + h.c.) + q \sum_{n=1}^{N-1} (|n+1, X\rangle \langle n, Y| + h.c.) \quad (1.1)$$

Here,  $m$  is the index of the unit cell and  $X$  and  $Y$  are the sub-lattice points. The abbreviation  $h.c.$  stands for Hermitian conjugate of the terms that are followed by  $h.c.$

We can separate the states using the tensor product basis and write:

$$|n, X(Y)\rangle \longrightarrow |n\rangle \otimes |X(Y)\rangle$$

Thus, we can write the SSH Hamiltonian using the tensor product basis by separating the external (unit cells) and internal (sub-lattice sites) degrees of freedom. To make the expression simpler, we use the Pauli matrices, which are defined as:

$$\sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad (1.2)$$

Now we can write the Hamiltonian in terms of the Pauli matrices defined above as follows:

$$\hat{H} \longrightarrow \mathcal{H}_{\text{ext}} \otimes \mathcal{H}_{\text{int}} = p \sum_{n=1}^N |n\rangle \langle n| \otimes \sigma_1 + q \sum_{n=1}^{N-1} \left( |n+1\rangle \langle n| \otimes \frac{\sigma_1 + i\sigma_2}{2} + h.c. \right) \quad (1.3)$$

Since this is a finite open chain, we will treat the edges and the bulk part of the chain separately because it is easier to solve the bulk part, for which we can assume periodic boundary conditions.

## 1.2 Bulk Hamiltonian

In any sort of physical realisation of a one dimensional chain, we will have a large value of  $N$ . This lets us impose the thermodynamic limit wherein the bulk will constitute the major part and will determine the physical properties of the chain. Thus, it makes sense to look at the bulk Hamiltonian. For ease of calculation, we impose periodic boundary conditions. Thus the bulk Hamiltonian  $H_b$  is defined as:

$$H_b = \sum_{n=1}^N (p |n, Y\rangle \langle n, X| + q |(n \bmod N) + 1, X\rangle \langle n, Y| + h.c.) \quad (1.4)$$

The bulk part has the property of translational invariance and as a result, we can apply Bloch's theorem and write plane wave basis states for the external Hilbert space as follows:

$$|k\rangle = \frac{1}{\sqrt{N}} \sum_{n=1}^N e^{ink} |n\rangle \quad (1.5)$$

Using the above basis states, we can write  $H_b$  in the following way:

$$H_b = \sum_k (p |k, Y\rangle \langle k, X| + p |k, X\rangle \langle k, Y| + q e^{ik} |(k, X)\rangle \langle k, Y| + q e^{-ik} |(k, Y)\rangle \langle k, X|)$$

We can separate the external and internal Hilbert space and write  $H_b$  as:

$$H_b = \sum_k |k\rangle \langle k| \otimes \{(p + q e^{-ik}) |Y\rangle \langle X| + (p + q e^{ik}) |(X)\rangle \langle Y|\}$$

We can further simplify the Hamiltonian by writing it succinctly using the band Hamiltonian.

$$H_b = \sum_k |k\rangle H(k) \langle k|$$

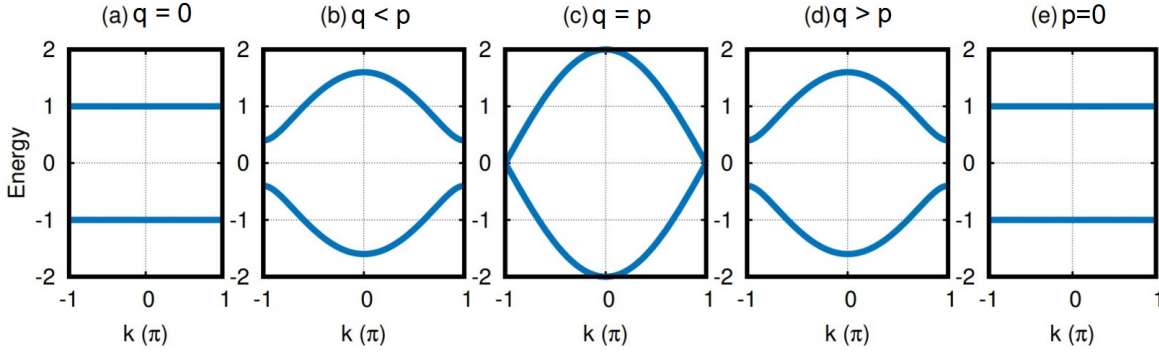


Figure 1.2: Dispersion relation of the SSH bulk Hamiltonian for various combinations of the hopping amplitudes  $p$  and  $q$ . Image source: [14] (after slight modification).

where, the internal Hilbert space band Hamiltonian  $H(k)$  is defined as:

$$H(k) = \begin{pmatrix} 0 & p + qe^{ik} \\ p + qe^{-ik} & 0 \end{pmatrix} \quad (1.6)$$

In order to find the band structure of the Hamiltonian, we have to diagonalise it and find the Energy dispersion relation. On diagonalising, we find that the eigenvalues and the eigenstates ( $E(k)$  and  $\psi(k)$ ) of  $H(k)$  are given by:

$$\begin{aligned} E(k) &= \pm \sqrt{p^2 + q^2 + 2pq \cos(k)} \\ \pm \psi(k) &= \begin{pmatrix} \pm e^{-i\theta(k)} \\ 1 \end{pmatrix} \\ \theta(k) &= \tan^{-1} \left( \frac{q \sin k}{p + q \cos k} \right) \end{aligned} \quad (1.7)$$

We can plot the energy dispersion for different values of  $p$  and  $q$ , as shown in Fig 1.2. We see that other than the  $p = q$  case, the model is that of an insulator because there is a band gap. Looking at the dispersion plots, it seems that the cases  $p < q$  and  $p > q$  are similar to each other, but in order to get complete information about the system, we have to look at the eigenvectors as well.



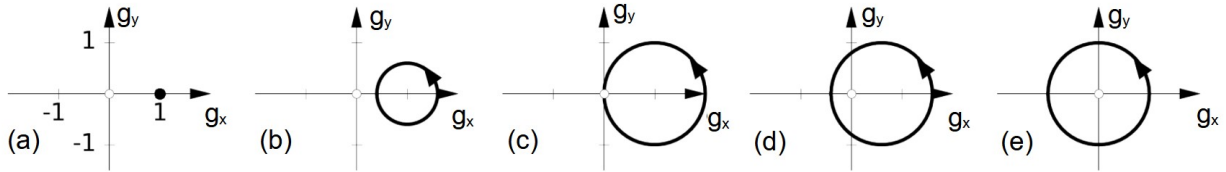


Figure 1.3: Trajectory of the end-point of  $\vec{g}(k)$  for the cases (a)  $p = 1, q = 0$ ; (b)  $p = 1, q = 0.6$ ; (c)  $p = 1, q = 1$ ; (d)  $p = 0.6, q = 1$ ; (e)  $p = 0, q = 1$ . Image source: [15] (after modification).

Using the Pauli matrices defined in (1.2), we can write the band Hamiltonian in a more compact way as shown below:

$$H(k) = \vec{g}(k) \cdot \vec{\sigma}, \quad (1.8)$$

where  $\vec{g}(k)$  and  $\vec{\sigma}$  are defined as follows:

$$\begin{aligned} \vec{\sigma} &= \{\sigma_i\}, \quad i \in \{0, 1, 2, 3\} \\ g_x(k) &= p + q \cos k, \\ g_y(k) &= q \sin k, \\ g_z(k) &= 0. \end{aligned} \quad (1.9)$$

As a result of the above definition, we see that the parameter that governs the eigenstates,  $\theta(k)$ , is now equal to  $\tan^{-1} \left( \frac{g_y(k)}{g_x(k)} \right)$ . This way of writing the band Hamiltonian in terms of  $\vec{g}(k)$  is particularly useful because when we look in the  $g_x - g_y$  plane (because  $g_z(k) = 0$ ), we see that the direction of the vector (which depends on  $\theta(k)$ ) represents the eigenvectors and the magnitude of the vector represents the eigenvalues of our band Hamiltonian. This representation captures full information about the Hamiltonian. On plotting this vector in the  $g_x - g_y$  plane, we can see a difference between the  $p < q$  and  $p > q$  cases, as illustrated in Fig 1.3.

The winding number of a path can be defined as the total number of complete loops around the origin that the path has traversed. In this process, we take the clockwise and anti-clockwise direction to be opposite to each other and thus one clockwise loop followed

by one anti-clockwise loop will result in a winding number of zero. Fig. 1.3 shows the path traversed by the vector  $\vec{g}(k)$ . These plots illuminate a difference between the hitherto similar cases of  $p < q$  and  $p > q$ . We see that in the case where  $p < q$ , the winding number of the curve about the origin is +1 whereas in the case where  $p > q$ , the winding number is 0. We can also see that in the process of transitioning from winding number 0 to winding number 1, the curve has to go through the case where  $p = q$  where the winding number is undefined because the curve passes through the origin itself. This point is the point of topological phase transition between the two topologically different phases ( $p < q$  and  $p > q$ ) characterised by their winding number. To see how this topological difference manifests itself physically, we will have to look at the original SSH model which was a finite open chain (i.e. without the periodic boundary conditions). Thus, we will have to include the boundary of the chain as well, rather than just focusing on the bulk.

### 1.3 Including the boundary in the SSH model

Once we include the boundary, we lose the periodic boundary conditions and as a result, we can no more solve this analytically. To see how the difference in topology manifests itself physically, we will have to look at the eigenstates of the SSH chain.

In Fig.1.4, we plot the energy eigenvalues and the eigenspectrum for the case of  $N = 10$ . In Fig.1.4(a), we see that there exist approximately zero energy states as long as  $p < q = 1$ . Fig.1.4 (b) and Fig.1.4 (c) show the nature of the eigenstates corresponding to such zero energies. These eigenstates are localised at the ends. These kind of states are called edge states. We cannot have such (approximately) zero energy edge states once  $p > q$ . This is how the topological difference within regions in the parameter space of  $p - q$  is expressed in the physical properties of the chain.

This is an example of the bulk-boundary correspondence. We looked at the bulk Hamiltonian and by calculating the topological invariant, we found that there are two topologically different regions in the parameter state which gives rise to two physically different phases. Thus, analysing the topological properties of the bulk Hamiltonian allows us to predict the existence of the edge states (and vice versa).

Thus, through the simple example of the SSH model, we saw the existence of topological

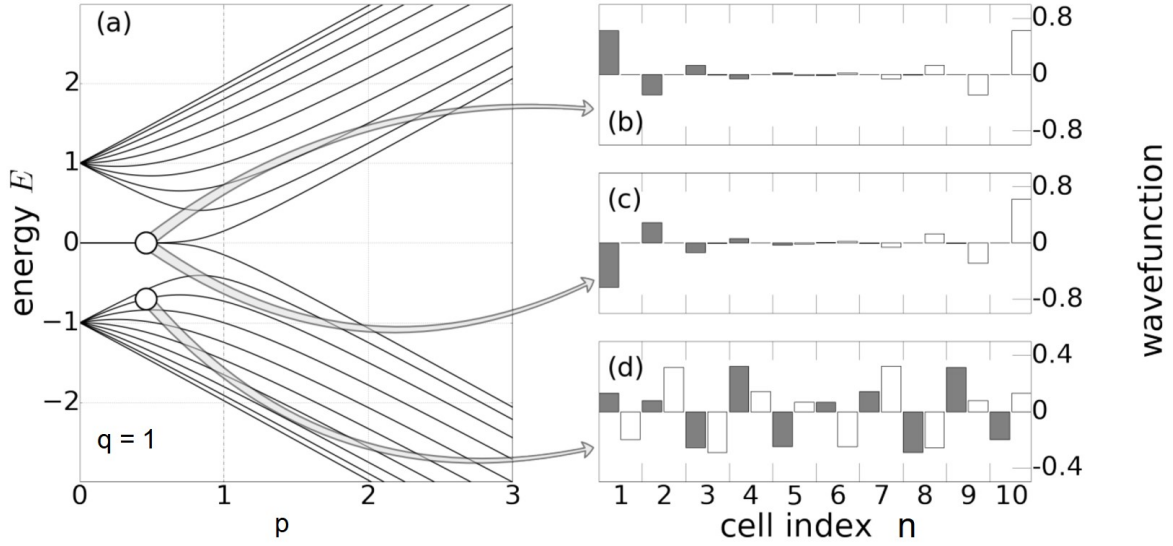


Figure 1.4: Energy spectrum and eigenstates for a finite sized SSH model with  $N = 10$  (a) Plot of the energy eigenvalues as a function of  $p$  with  $q = 1$ ; (b), (c) Plots of two approximately zero energy eigenfunctions when  $p < q$  ; (d) a generic delocalised eigenstate, when  $p > q$ . Image source: [15]

phases, their physical consequences and their characterisation with the help of a topological invariant - which was the winding number in our case. It is this winding number that we will calculate using unsupervised machine learning algorithm.



# Chapter 2

## The Diffusion Map Algorithm

This chapter deals with the machine learning approach called the Diffusion Map algorithm [17], [18]. It was originally proposed as a dimensionality reduction method. In essence, such methods search for some principle variables out of the many variables present in the data set and as a result, they classify the data. We will first go over the basics of the algorithm and then see how it can be used to perform classification of topological insulators. For simplicity, we will proceed by assuming that the data set describes some physical system that we eventually want to classify.

### 2.1 Defining a diffusion process on the data set

The first step in the algorithm is to define a diffusion process on the data set. Let us denote our data set with  $X = \{x_m \mid m = 1, 2, \dots, N\}$  where each data point  $x_m$  describes a physical system and lies in the configuration space  $\Omega$ . Now we treat each data point as a node on a graph and we define a function that determines the transition probability from one node (data point) to the other. Definition of the function, known as the kernel, is the key point here and it should be such that the value of the kernel is low between data points that are not “similar” to each other. Fig. 2.1 shows a pictorial representation of the diffusion process and the type of kernel that is ideally desirable.

To define the transition probabilities and subsequently the transition matrix for the

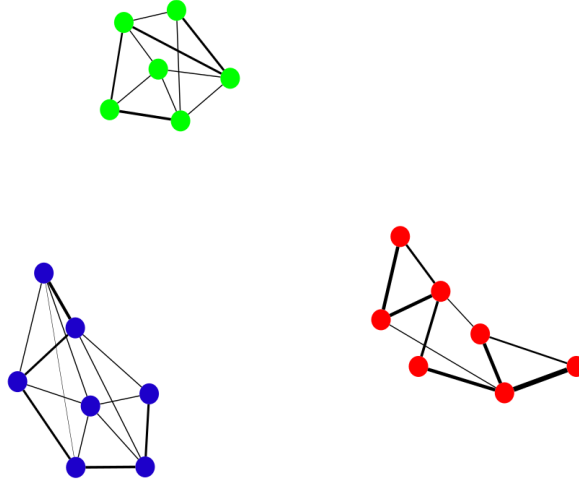


Figure 2.1: Each node represents a data point. Nodes with the same colour represent similar data points. The width of the edges denote the magnitude of transition probability. Ideally, the transition probabilities should be close to zero for non-similar data points.

process, we first define a kernel  $k(x, y)$  that maps  $X \times X$  to a non-negative real number and is symmetric in its arguments. Let  $t_{m,m'}$  be the transition probability from  $x_m$  to  $x_{m'}$ . We can define  $t_{m,m'}$  as:

$$t_{m,m'} = \frac{k(m, m')}{z_m}, \quad \text{where } z_m = \sum_{m'} k(m, m'). \quad (2.1)$$

The above definition defines a transition matrix  $T$  whose elements are given by  $t_{m,m'}$ . The transition matrix follows the following conditions:

$$\begin{aligned} t_{m,m'} &= t_{m',m}, \\ t_{m,m'} &\geq 0, \\ \sum_{m'} t_{m,m'} &= 1. \end{aligned} \quad (2.2)$$

The Markov process defined by the transition matrix has a stationary distribution which is given by:

$$\pi_m = \frac{z_m}{\sum_{m'} z_{m'}} \quad (2.3)$$

We also see that the process follows the detailed balanced equation, and thus, it is reversible. These properties allow us to perform its spectral analysis. We can say that the matrix  $T$  has a set of eigenvalues,  $\lambda_m$ , and eigenvectors,  $\psi_m$ , that satisfy:

$$T\psi_m = \lambda_m\psi_m \quad (2.4)$$

## 2.2 The Diffusion Distance

In order to perform classification of the data set, we look at the ‘‘Diffusion Distance’’ [17]. The Diffusion Distance between two data points  $m$  and  $m'$  is obtained after performing the diffusion process defined by the matrix  $T$  for a certain number of times, say  $2p$  times. For large values of  $p$ , this metric captures the similarity between the two samples  $m$  and  $m'$ . The Diffusion Distance after  $2p$  steps is defined as:

$$D_{2p}(m, m') := \sum_{m''} \frac{1}{z_{m''}} ((t^p)_{m, m''} - (t^p)_{m', m''})^2 \quad (2.5)$$

Using (2.1), we can write the above expression as:

$$D_{2p}(m, m') = \frac{1}{z_m} (t^{2p})_{m, m} + \frac{1}{z_{m'}} (t^{2p})_{m', m'} - \frac{2}{z_{m'}} (t^{2p})_{m, m'} \quad (2.6)$$

The above makes it obvious that the  $2p$ -step Diffusion distance between data point  $m$  and  $m'$  is smaller when these two points are maximally connected (i.e.  $\frac{2}{z_{m'}} (t^{2p})_{m, m'}$  is large). Taking the example of the extreme case where  $m = m'$ , we see that the Diffusion distance is zero for that case. Thus, more similar the data points, lower is the diffusion distance.

As shown in [17], we can write the Diffusion distance in terms of the eigenvalues and eigenvectors defined in (2.4). Such an expression of (2.5) is given by:

$$D_{2p}(m, m') = \sum_{n=0}^N \lambda_n^{2p} [(\psi_n)_m - (\psi_n)_{m'}]^2 \quad (2.7)$$

If we order the eigenvalues and eigenvectors such that  $\lambda_0 \geq \lambda_1 \geq \dots \geq \lambda_{N-1}$ , then since  $T$  is the transition matrix of a diffusion process,  $\lambda_0 = 1$  and the eigenvector corresponding to that doesn't contribute to (2.7). Thus we have the Diffusion Map:

$$x_m \longrightarrow \begin{pmatrix} (\psi_1)_m \\ (\psi_2)_m \\ \vdots \\ (\psi_k)_m \end{pmatrix} \quad (2.8)$$

where  $k \leq N - 1$ . For large enough  $k$ , it can be shown that the approximate Diffusion distance defined as,

$$D_{2p}(m, m') = \sum_{n=1}^k \lambda_n^{2p} [(\psi_n)_m - (\psi_n)_{m'}]^2 \quad (2.9)$$

is approximately equal to the original Diffusion distance defined in (2.7). Thus, the mapping (2.8), gives a lower dimensional representation of the data wherein the dimensions have been reduced from  $N$  to  $k$ . We will see that in our case, after appropriate definition of the kernel, this  $k$  will be equal to the total number of topological phases.

## 2.3 The Auxiliary Quantum Problem

As shown in [18], we can derive a quantum mechanical problem corresponding to the Diffusion Map that we get. This Hamiltonian has the property that its low energy spectrum determines the  $\lambda_m$  and  $\psi_m$  of the Diffusion Map (up to a certain value of  $m$ ). In order to derive the Hamiltonian, we first need to take the continuum limit of the diffusion process.



### 2.3.1 Diffusion Map in the continuum limit

We first embed our configuration space  $\Omega$  in Euclidean space  $\mathbb{R}^n$ . We also assume the following form of the kernel (which we will use later for our particular case):

$$k(x_m, x_{m'}) = h\left(\frac{\|x_m - x_{m'}\|^2}{\epsilon}\right) \quad (2.10)$$

where  $h(x)$  is a function that decays exponentially when the argument is large,  $\|\cdot\|$  denotes the Euclidean norm in  $\mathbb{R}^n$  and  $\epsilon$  is the parameter that determines how fast the kernel decays. We also assume that  $N$  is large enough so that we can assume  $N \rightarrow \infty$ . Under these conditions, we can view the summations to be Monte-Carlo sampled integrals. Thus, quantities get redefined in the following way:

$$z_\epsilon(x_m) := \int_{\Omega} dx_{m'} \rho(x_{m'}) k_\epsilon(x_m, x_{m'}) \quad (2.11)$$

where  $\rho(x_{m'})$  is the probability distribution from which the samples are drawn. The transition probabilities in this case get redefined as:

$$t_\epsilon(x_m, x_{m'}) = \frac{k_\epsilon(x_m, x_{m'})}{z_\epsilon(x_m)}, \quad (2.12)$$

and the eigenvalue equation (2.4) becomes:

$$\int_{\Omega} dx_{m'} \rho(x_{m'}) t_\epsilon(x_m, x_{m'}) \psi_n(x_{m'}) = \lambda_n \psi_n(x_m) \quad (2.13)$$

For the case when  $N \rightarrow \infty$ , the eigenvectors  $\psi_n(x_m)$  are approximately equal to  $(\psi_n)_m$  defined in (2.4). We can define an operator  $\hat{T}_\epsilon$  that acts on functions in  $\Omega$ :

$$\left(\hat{T}_\epsilon f\right)(x_m) := \int_{\Omega} dx_{m'} \rho(x_{m'}) t_\epsilon(x_m, x_{m'}) f(x_{m'}) \quad (2.14)$$

Thus in the limit  $N \rightarrow \infty$ ,  $(\psi_n)_m$  and  $\lambda_n$  (defined in (2.4)) approach the solution of:

$$\hat{T}_\epsilon \psi_n = \lambda_n \psi_n \quad (2.15)$$

### 2.3.2 The auxiliary Hamiltonian

We assume that our data set  $X$  is of a physical system that has an energy functional  $E(x_m)$ . Thus, the probability distribution is given by:

$$\rho(x_m) = \frac{e^{-E(x_m)}}{Z} \quad (2.16)$$

where  $Z$  is the partition function.

Here, we use the result of [18] which states that the generator

$$\hat{G}_\epsilon = \frac{1 - \hat{T}_\epsilon}{\epsilon} \quad (2.17)$$

satisfies

$$\hat{G}_\epsilon f \sim -\frac{\hat{\Delta} f \rho}{\rho} + \frac{(\hat{\Delta} \rho)}{\rho} f, \quad \epsilon \rightarrow 0 \quad (2.18)$$

for  $f$  defined on  $\Omega$ . Here,  $\hat{\Delta} = \hat{\nabla} \cdot \hat{\nabla}$ .

For  $\rho(x_m) = \frac{e^{-E(x_m)}}{Z}$ , in the limit  $\epsilon \rightarrow 0$ , the solutions  $\{\lambda_n, \psi_n\}$  in (2.15) satisfy the following Schrödinger equation:

$$\mathcal{H} \Psi_n = E_n \Psi_n \quad (2.19)$$

where  $\mathcal{H}$  is

$$\mathcal{H} = -\hat{\Delta} + V(x), \quad V(x) = \|\hat{\nabla} E\|^2 - \hat{\Delta} E \quad (2.20)$$

As  $\epsilon \rightarrow 0$ , we have:

$$\lambda_n \sim 1 - \epsilon E_n, \quad \psi_n \sim \frac{\Psi_n}{\rho} \quad (2.21)$$

We use this result later to derive the auxiliary Hamiltonian for our case.



# Chapter 3

## Unsupervised Classification of 1-D 2-band Hamiltonians

We will now use the Diffusion Map algorithm to perform classification of topological phases in physical systems, which was first done in [4] where it was shown that it can detect the phases in the  $XY$  model. Subsequently, it was also used for classification of band Hamiltonians in [5]. In this work, we extend this approach and classify an ensemble of 1-dimensional 2-band Bloch Hamiltonians with chiral symmetry, that are drawn from a specific probability distribution.

### 3.1 Our Model

We sample 2 band Hamiltonians in 1D,  $h_k = h_k^\dagger \in \mathbb{C}^{2 \times 2}$ , with chiral symmetry  $C = i\sigma_2$ . Therefore, the Hamiltonians obey  $h_k, C = 0$ . Similar to the approach in chapter 1, we write the Hamiltonian using the Pauli matrices (1.2) in the following form:

$$h_k = g_k^x \sigma_1 + g_k^z \sigma_3, \quad g_k^i = g_{k+2\pi}^i \quad (3.1)$$

Rather than performing the path-finding approach used in [5], we perform importance sampling of  $h_k$ . Similar to random-matrix ensembles, we consider the ensemble of  $2 \times 2$ -

matrix-valued functions  $h_k$  which are  $2\pi$  periodic and sufficiently smooth in  $k$  and obey  $\{h_k, C\} = 0$ . We sample  $h_k$  distributed according to:

$$P(\{h_k\}) = \frac{1}{Z} e^{-\frac{\alpha}{2} \int_k \text{tr}[h_k^\dagger h_k] - \frac{\beta}{2} \int_k \text{tr}[\partial_k h_k^\dagger \partial_k h_k]} \prod_k \sigma(\{h_k, C\}). \quad (3.2)$$

where  $\int_k \dots = \int_0^{2\pi} dk \dots$ . Due to the parameterisation in (3.1), the above probability distribution can be written as:

$$P(\{\vec{g}_k\}) = \frac{1}{Z} e^{-E[\vec{g}_k]}, \quad E[\{\vec{g}_k\}] = \alpha \int_k \vec{g}_k^2 + \beta \int_k (\partial_k \vec{g}_k)^2 \quad Z = \int \prod_k d^2 \vec{g}_k e^{-E[\vec{g}_k]}, \quad (3.3)$$

where we defined  $\vec{g}_k = (g_k^x, g_k^y)^T$ ; this form of the probability density is much easier to sample since the chirality constraint is already taken into account.

## 3.2 Choice of Kernel

We use the local similarity measure which was defined in [5]. For a data set comprising of  $N_b$ -band Hamiltonians  $h_k^m$  (where  $m$  denotes the data point) that solve the eigenvalue equation:

$$h_k^m |\psi_{nk}^m\rangle = E_{nk}^m |\psi_{nk}^m\rangle, \quad (3.4)$$

we use the local similarity measure between Hamiltonians  $m$  and  $m'$  given by:

$$S_{m,m'} = \frac{1}{N_k} \sum_k \frac{1}{N_b} \sum_{n=1}^{N_b} |\langle \psi_{nk}^m | \psi_{nk}^{m'} \rangle|^2 \quad (3.5)$$

where  $N_k$  is the total number of  $k$ -points.

In our case, because of the parameterisation, the above similarity measure is given by the following expression:

$$S_{m,m'} = \frac{1}{2} \left[ 1 + \frac{1}{2\pi} \int_k \hat{g}_k^m \cdot \hat{g}_k^{m'} \right], \quad \hat{g}_k^m := \frac{\vec{g}_k^m}{|\vec{g}_k^m|}. \quad (3.6)$$

For our purpose, we will need a discretised version, which is given by:

$$S_{m,m'} = \frac{1}{2} \left[ 1 + \frac{1}{N_k} \sum_k \hat{g}_k^m \cdot \hat{g}_k^{m'} \right], \quad \hat{g}_k^m := \frac{\vec{g}_k^m}{|\vec{g}_k^m|}. \quad (3.7)$$

The above similarity measure reduces the quantum problem to a classical problem. Using this measure, we define the kernel for our Diffusion Map

$$k(x_m, x_{m'}) = e^{-(1-S_{m,m'})/\epsilon} \quad (3.8)$$

which in turn defines the transition matrix elements  $T_{m,m'}$

$$T_{m,m'} = \frac{k(x_m, x_{m'})}{z_m}, \quad z_m = \sum_{m'} k(x_m, x_{m'}). \quad (3.9)$$

We use this transition matrix to proceed with the Diffusion Map and classify the Hamiltonians based on their topology.

### 3.3 Sampling of the Hamiltonians

To sample the Hamiltonians, we first convert (3.3) to a discrete version

$$E[\vec{g}_k] = \alpha \sum_k g_k^2 + \alpha' \sum_k (g_{k+1} - g_k)^2 - \beta \sum_k g_k^2 \cos(\theta_{k+1} - \theta_k) \quad (3.10)$$

where we have reparameterised  $\vec{g}_k = g_k(\cos \theta_k, \sin \theta_k)^T$ . Using this energy functional, we perform Monte-Carlo sampling of the Hamiltonians. The parameter values for sampling, corresponding to which the plots were obtained are:  $N_k = 30$ ,  $\alpha = 1.0$ ,  $\alpha' = 0.5$ ,  $\beta = 0.58$ . We used the kernel as defined in (3.8), with  $\epsilon = 0.055$ , to obtain the transition matrix  $T$ . The results have been plotted in 3.1. We discuss the results in the next section.

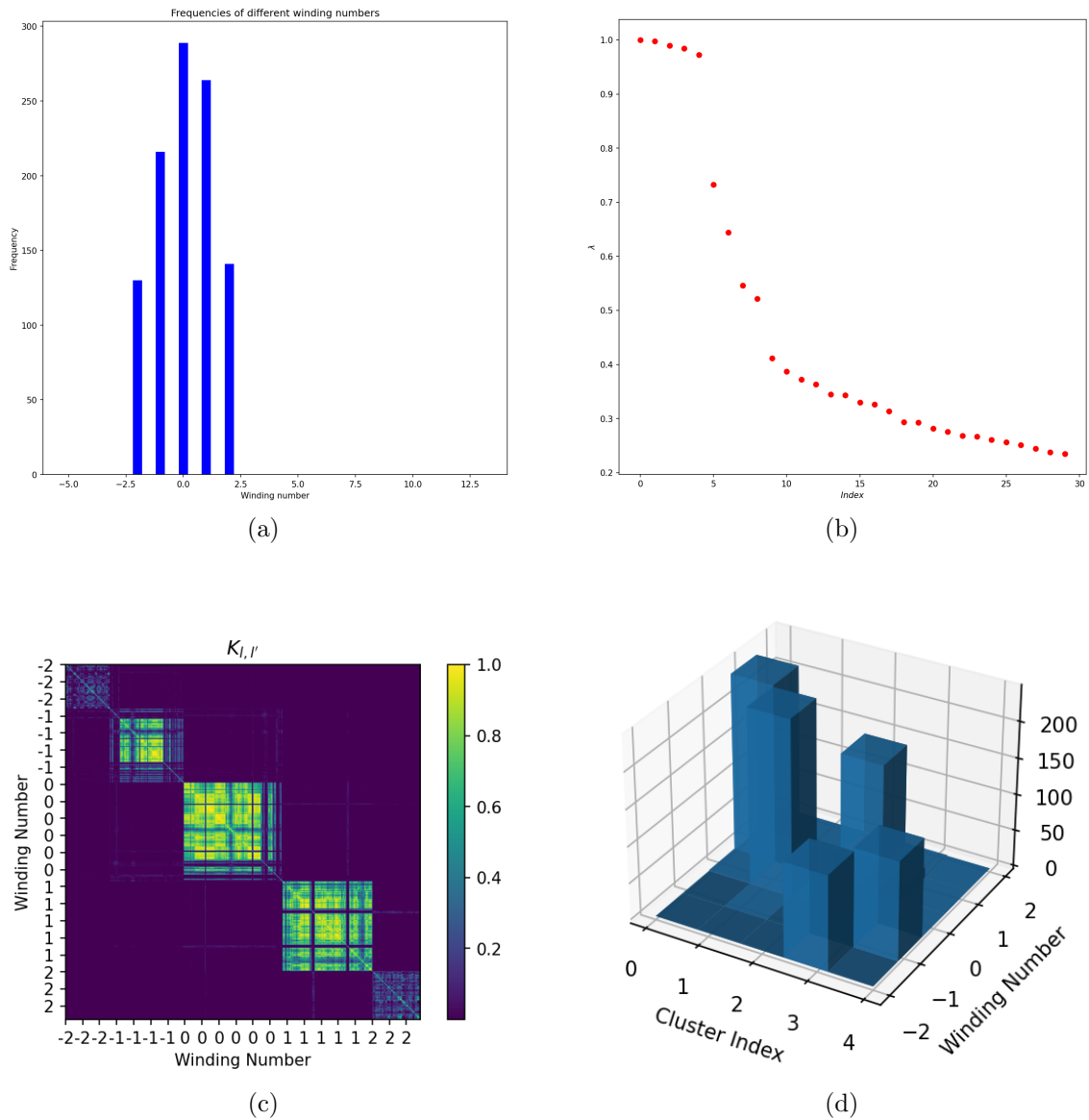


Figure 3.1: **Results for 1-D 2 band model** (a) Number of samples  $vs$  their winding number (b) Eigenvalues ( $\lambda_i$ ) of the transition matrix  $T$ . We see that there are 5 leading eigenvalues, corresponding to the five topologically different types of Hamiltonians that are present in the data set (c) The transition probability matrix  $T$  (here denoted as  $K$ ). The samples have been rearranged in the order of increasing winding number. Value of matrix element  $K_{l,l'}$  is higher when samples  $l$  and  $l'$  are topologically similar (i.e. have the same winding number) (d) We perform K-Means clustering after plotting the eigenvectors  $\Psi_i$  ( $i = 1, \dots, 4$ ) in 4 dimensions. K-Means associates each sample to a cluster. Here, we plot a histogram between the cluster index and the winding number for all the samples.

### 3.4 Discussion

We started with chains that had winding numbers starting from 0, going all the way up to  $\pm 10$ . Then we allowed random fluctuations on random sites and performed Monte-Carlo sampling. The resulting distribution with respect to the winding numbers has been plotted. As expected, we get a Gaussian like distribution which favours lower winding numbers, so much so that all the chains with winding number  $> 2$  eventually ended up lying in range of  $-2$  to  $2$ .

3.1(b) shows the eigenvalues of the transition matrix  $T$ . We see that there are five leading eigenvalues. Going back to the definition of Diffusion Distance in (2.7), which was defined as:

$$D_{2p}(m, m') = \sum_{n=0}^N \lambda_n^{2p} [(\psi_n)_m - (\psi_n)_{m'}]^2,$$

we can say that the Diffusion Distance can now be approximated by:

$$D_{2p}(m, m') = \sum_{n=1}^4 \lambda_n^{2p} [(\psi_n)_m - (\psi_n)_{m'}]^2,$$

since for large values of  $p$ , the contribution due to other eigenvectors will be close to zero as  $\lambda_i^{2p}$  will be very small. Thus, the Diffusion Map reduced the data set into four eigenvectors which can give us the five clusters that are associated to the five topologically different Hamiltonians (the idea which was first proposed in [4]).

Thus, for any data set involving Hamiltonians that have topological phases, the total number of leading eigenvalues gives us the number of topologically different phases present in the data set.

We then rearrange the transition matrix rows based on the winding number of samples (which is calculated from the angle rotated by the vector  $\vec{g}$  as we move from  $k = 1$  to  $k = N_k$  in the chain). After rearrangement, we plot the matrix in 3.1 (c). We see that the transition probability (which is proportional to the similarity between samples, as defined in (3.8)), is higher for samples with same winding number. Thus it shows that the algorithm was able to classify the Hamiltonians correctly.



To further see this, we plotted the eigenvectors in four-dimensional space and performed K-Means clustering (refer Appendix). The K-Means algorithm associates every data point with a cluster centre. Thus, our samples now have two labels - the K-means cluster centre and the winding number. If the classification has been done correctly, then these labels will match for majority of the samples and indeed, we see that is the case in 3.1 (d).

To summarise overall, we extended the study of diffusion map algorithm for topological classification ([4], [5]) and applied it to classify data set *sampled* from a probability distribution. This further goes on to show the robustness and applicability of this unsupervised technique.



# Chapter 4

## Derivation of the Auxiliary Hamiltonian

Now we move on to the final part, which is the derivation of the auxiliary quantum many-body Hamiltonian for the Diffusion process defined on our data set. We use the general approach outlined in section 2.3.

### 4.1 The general scheme

We know that when the samples,  $x_t$ , for the Diffusion Map algorithm are drawn from a probability distribution of the form

$$\rho = e^{-E(x)}/Z, \tag{4.1}$$

then the corresponding dual quantum mechanical Hamiltonian is given by:

$$\mathcal{H}_{\text{eff}} = -\hat{\Delta} + V(x), \tag{4.2}$$

where  $V(x)$  is given by:

$$V(x) = \|\hat{\nabla}E\|^2 - \hat{\Delta}E. \tag{4.3}$$

## 4.2 Simplifying the probability distribution

In our case, the sample consists of Hamiltonians of the following form:

$$h_k = g_k^x \sigma_x + g_k^z \sigma_z, \quad g_k^j = g_{k+2\pi}^j \in \mathbb{R}. \quad (4.4)$$

We have shown that the ML algorithm successfully clusters  $\{h_k\}$  which are  $2\pi$  periodic, sufficiently smooth in  $k$ , obey  $\{h_k, C\} = 0$  and are distributed according to:

$$P(\{h_k\}) = \frac{1}{Z} e^{-\frac{\alpha}{2} \int_k \text{tr}[h_k^\dagger h_k] - \frac{\beta}{2} \int_k \text{tr}[\partial_k h_k^\dagger \partial_k h_k]} \prod_k \sigma(\{h_k, C\}).$$

Due to the parameterisation in (4.4), we have:

$$P(\{\vec{g}_k\}) = \frac{1}{Z} e^{-E[\vec{g}_k]}, \quad E[\{\vec{g}_k\}] = \alpha \int_k \vec{g}_k^2 + \beta \int_k (\partial_k \vec{g}_k)^2 \quad Z = \int \prod_k d^2 \vec{g}_k e^{-E[\vec{g}_k]}, \quad (4.5)$$

where we defined  $\vec{g}_k = g_k \hat{g}_k = (g_k^x, g_k^y)^T$ . On discretising the energy function, we get:

$$E[\{\vec{g}_k\}] = \alpha \sum_k \vec{g}_k^2 + \beta \sum_k (\vec{g}_{k+1} - \vec{g}_k)^2$$

Consequently, the discretised expression for the corresponding probability distribution becomes:

$$P(\{\vec{g}_k\}) = \frac{1}{Z} e^{-[\alpha \sum_k \vec{g}_k^2 + \beta \sum_k (\vec{g}_{k+1} - \vec{g}_k)^2]}$$

Since the topological phases were only dependent on the angle rotated by  $\vec{g}_k$  as we move along  $k$ , and not on the variable magnitude at every  $k$ -point, we take the conditional probability distribution that depends only on the directions, i.e.  $\{\hat{g}_k\}$ . Thus, we integrate out  $g_k$  and get the following expression for  $P(\{\hat{g}_k\})$ :

$$\begin{aligned} P(\{\hat{g}_k\}) &= \frac{1}{Z} \int \prod_k dg_k g_k e^{-E[\vec{g}_k]} \\ &= \frac{1}{Z} \int \prod_k \left( dg_k g_k e^{-[\alpha \sum_k \vec{g}_k^2 + \beta \sum_k (\vec{g}_{k+1} - \vec{g}_k)^2]} \right) \\ &= \langle e^{-\beta \sum_k (\vec{g}_{k+1} - \vec{g}_k)^2} \rangle \end{aligned}$$

where,

$$\langle \dots \rangle := \frac{1}{Z} \int \prod_k dg_k g_k e^{-\alpha \sum_k g_k^2}.$$

In order to obtain the Hamiltonian, it is convenient to have a probability distribution of the form shown in (4.1). The following calculations are done to achieve such a form for the probability distribution. We start by expanding the exponential up to second order, which gives:

$$\begin{aligned} P(\{\hat{g}_k\}) &\sim \langle 1 - \beta \sum_k (\vec{g}_{k+1} - \vec{g}_k)^2 + \frac{\beta^2}{2} \sum_{k,k'} (\vec{g}_{k+1} - \vec{g}_k)^2 (\vec{g}_{k'+1} - \vec{g}_{k'})^2 \rangle \\ &= \langle 1 - \beta \sum_k (g_{k+1}^2 + g_k^2 - 2g_{k+1}g_k \hat{g}_{k+1} \cdot \hat{g}_k) \\ &\quad + \frac{\beta^2}{2} \sum_{k,k'} \left\{ g_{k+1}^2 g_{k'+1}^2 + g_{k+1}^2 g_{k'}^2 - 2g_{k+1}^2 g_{k'+1} g_{k'} \hat{g}_{k'+1} \cdot \hat{g}_{k'} \right. \\ &\quad \quad + g_k^2 g_{k'+1}^2 + g_k^2 g_{k'}^2 - 2g_k^2 g_{k'+1} g_{k'} \hat{g}_{k'+1} \cdot \hat{g}_{k'} \\ &\quad \quad - 2g_{k+1} g_k \hat{g}_{k+1} \cdot \hat{g}_k g_{k'+1}^2 - 2g_{k+1} g_k \hat{g}_{k+1} \cdot \hat{g}_k g_{k'}^2 \\ &\quad \quad \left. + 4g_{k+1} g_k \hat{g}_{k+1} \cdot \hat{g}_k g_{k'+1} g_{k'} \hat{g}_{k'+1} \cdot \hat{g}_{k'} \right\} \rangle \\ &= \langle 1 - \beta \sum_k [2g_k^2 - 2g_{k+1}g_k (\hat{g}_{k+1} \cdot \hat{g}_k)] \\ &\quad + \frac{\beta^2}{2} \sum_{k,k'} \left\{ 4g_k^2 g_{k'}^2 + 4g_{k+1}g_k g_{k'+1}g_{k'} (\hat{g}_{k+1} \cdot \hat{g}_k) (\hat{g}_{k'+1} \cdot \hat{g}_{k'}) - 8g_k^2 g_{k'+1}g_{k'} (\hat{g}_{k'+1} \cdot \hat{g}_{k'}) \right\} \rangle \\ &= \langle 1 - 2\beta \sum_k g_k^2 + 2\beta^2 \sum_{k,k'} g_k^2 g_{k'}^2 + \left( 2\beta - 4\beta^2 \sum_{k'} g_{k'}^2 \right) \sum_k g_{k+1}g_k (\hat{g}_{k+1} \cdot \hat{g}_k) \\ &\quad + 2\beta^2 \sum_{k,k'} g_{k+1}g_k g_{k'+1}g_{k'} (\hat{g}_{k+1} \cdot \hat{g}_k) (\hat{g}_{k'+1} \cdot \hat{g}_{k'}) \rangle \end{aligned}$$

To simplify the above expression, we define the following quantities:

$$\begin{aligned} \tilde{\alpha} &:= \langle 1 - 2\beta \sum_k g_k^2 + 2\beta^2 \sum_{k,k'} g_k^2 g_{k'}^2 \rangle \\ \tilde{\beta} &:= \langle \left( 2\beta - 4\beta^2 \sum_{k'} g_{k'}^2 \right) g_{k+1}g_k \rangle \quad \forall k, \\ \hat{\beta}_{k,k'} &:= \langle 2\beta^2 g_{k+1}g_k g_{k'+1}g_{k'} \rangle. \end{aligned} \tag{4.6}$$

We note the following property of  $\hat{\beta}_{k,k'}$  which follows from its definition and translation symmetry in the problem:

$$\hat{\beta}_{k,k'} = \hat{\beta}_{l,l'}, \quad \forall \quad |k - k'| = |l - l'|. \quad (4.7)$$

Thus, we can write  $\hat{\beta}_{k,k'}$  as  $\hat{\beta}_\eta$ , where  $\eta = |k - k'|$ . Using (4.6), we can write  $P(\{\hat{g}_k\})$  in a much more simpler form:

$$P(\{\hat{g}_k\}) = \tilde{\alpha} + \tilde{\beta} \sum_k (\hat{g}_{k+1} \cdot \hat{g}_k) + \sum_{k,k'} \hat{\beta}_{k,k'} (\hat{g}_{k+1} \cdot \hat{g}_k) (\hat{g}_{k'+1} \cdot \hat{g}_{k'}) \quad (4.8)$$

Due to the integral involved in  $\langle \dots \rangle$  and the translational symmetry of the problem, the seemingly  $k$  dependence of the parameters vanish as the integral is over the full chain.

### 4.3 Deriving $E_{\text{eff}}$

Let us assume for some  $E_{\text{eff}}$ , we can get a probability distribution of the form shown in (4.1) that equals the probability that we obtained in (4.8). Let  $E_{\text{eff}}$  be:

$$E_{\text{eff}} = -\sigma - \epsilon \sum_k (1 - \hat{g}_{k+1} \cdot \hat{g}_k) - \sum_{k,k'} f_{k,k'} (1 - \hat{g}_{k+1} \cdot \hat{g}_k) (1 - \hat{g}_{k'+1} \cdot \hat{g}_{k'})$$

The corresponding probability (up to second order in  $(1 - \hat{g}_{k+1} \cdot \hat{g}_k)$ ) is given by:

$$\begin{aligned} e^{-E_{\text{eff}}} &\sim 1 + \sigma + \frac{\sigma^2}{2} + (\epsilon + \sigma\epsilon) \sum_k (1 - \hat{g}_{k+1} \cdot \hat{g}_k) \\ &\quad + \sum_{k,k'} (f_{kk'} + \frac{\epsilon^2}{2} + \sigma f_{kk'}) (1 - \hat{g}_{k+1} \cdot \hat{g}_k) (1 - \hat{g}_{k'+1} \cdot \hat{g}_{k'}) \end{aligned} \quad (4.9)$$

Now we group the terms together so that the expression becomes comparable to (4.8). On comparing, we find the relations between the old parameters  $\tilde{\alpha}$ ,  $\tilde{\beta}$ ,  $\tilde{\beta}_{k,k'}$  and  $\sigma$ ,  $\epsilon$ ,  $f_{k,k'}$ .

Note that the properties mentioned in (4.7) apply to  $f_{k,k'}$ .

$$\begin{aligned}
e^{-E_{\text{eff}}} &= 1 + \sigma + \frac{\sigma^2}{2} + \underbrace{\sum_k (\epsilon + \sigma\epsilon) + \sum_{k,k'} (f_{kk'} + \frac{\epsilon^2}{2} + \sigma f_{kk'})}_{\tilde{\alpha}} \\
&\quad + \underbrace{\left( -\epsilon - \sigma\epsilon - 2 \sum_{k'} (f_{kk'} + \frac{\epsilon^2}{2} + \sigma f_{kk'}) \right)}_{\tilde{\beta}} \sum_k (\hat{g}_{k+1} \cdot \hat{g}_k) \\
&\quad + \sum_{k,k'} \underbrace{(f_{kk'} + \frac{\epsilon^2}{2} + \sigma f_{kk'})}_{\hat{\beta}_{k,k'}} (\hat{g}_{k+1} \cdot \hat{g}_k) (\hat{g}_{k'+1} \cdot \hat{g}_{k'})
\end{aligned} \tag{4.10}$$

This shows that we can effectively write (4.8) using  $E_{\text{eff}}$  given by:

$$\boxed{E_{\text{eff}} = -\sigma - \epsilon \sum_k (1 - \hat{g}_{k+1} \cdot \hat{g}_k) - \sum_{k,k'} f_{k,k'} (1 - \hat{g}_{k+1} \cdot \hat{g}_k) (1 - \hat{g}_{k'+1} \cdot \hat{g}_{k'})} \tag{4.11}$$

This in turn, enables us to write the probability distribution in the form shown in (4.1), given by  $P(\{\hat{g}_k\}) = e^{-E_{\text{eff}}}$ .

## 4.4 Calculating the expression for potential energy

Since we now have a convenient expression for the probability distribution, we can now straightaway use the formulae given in (4.3) and (4.2) to write the auxiliary Hamiltonian by replacing  $E$  with  $E_{\text{eff}}$ . To make the analysis easy, we define the following two quantities:

$$\begin{aligned}
\phi &:= -\epsilon \sum_k (1 - \hat{g}_{k+1} \cdot \hat{g}_k) && \text{first order} \\
\omega &:= \sum_{k,k'} f_{k,k'} (1 - \hat{g}_{k+1} \cdot \hat{g}_k) (1 - \hat{g}_{k'+1} \cdot \hat{g}_{k'}) && \text{second order}
\end{aligned} \tag{4.12}$$

Now, we will calculate the expression for potential energy up to second order in  $(1 - \hat{g}_{k+1} \cdot \hat{g}_k)$  using (4.3). To proceed further, we also note that in our case, the following holds

true:

$$\begin{aligned}
\hat{g}_k &= (\cos \theta_k, \sin \theta_k)^T \\
\implies \hat{g}_{k+1} \cdot \hat{g}_k &= (\cos \theta_{k+1}, \sin \theta_{k+1}) (\cos \theta_k, \sin \theta_k)^T \\
&= \cos \theta_{k+1} \cos \theta_k + \sin \theta_{k+1} \sin \theta_k \\
&= \cos (\theta_{k+1} - \theta_k)
\end{aligned} \tag{4.13}$$

From (4.3), we know that the expression for potential energy is given by:

$$V(x) = \|\hat{\nabla} E\|^2 - \hat{\Delta} E.$$

Now we will calculate  $\|\hat{\nabla} E_{\text{eff}}\|^2$  and  $\hat{\Delta} E_{\text{eff}}$  for our case (both up to second order in  $(1 - \hat{g}_{k+1} \cdot \hat{g}_k)$ ). Since the configuration space is embedded in  $\mathbb{R}^n$ , we also note that the operators are defined as:

$$\begin{aligned}
\hat{\nabla} &:= \sum_j \frac{\partial}{\partial x_j} \hat{\mathbf{j}} \\
\hat{\Delta} &:= \sum_j \frac{\partial^2}{\partial x_j^2}
\end{aligned} \tag{4.14}$$

#### 4.4.1 $\|\hat{\nabla} E_{\text{eff}}\|^2$

The first derivative of  $E_{\text{eff}}$  w.r.t. a coordinate  $\theta_\rho$  is:

$$\begin{aligned}
\frac{\partial E_{\text{eff}}}{\partial \theta_\rho} &= \frac{\partial \phi}{\partial \theta_\rho} + \frac{\partial \omega}{\partial \theta_\rho} \\
\implies \|\hat{\nabla} E_{\text{eff}}\|^2 &\sim \sum_\rho \left( \frac{\partial \phi}{\partial \theta_\rho} \right)^2 \\
&= \sum_\rho \left( \frac{\partial}{\partial \theta_\rho} \left[ -\epsilon \sum_k (1 - \hat{g}_{k+1} \cdot \hat{g}_k) \right] \right)^2 \\
&= \sum_\rho \left( \frac{\partial}{\partial \theta_\rho} \left[ -\epsilon \sum_k (1 - \cos (\theta_{k+1} - \theta_k)) \right] \right)^2 \quad \text{using (4.13)} \\
&= \sum_\rho \epsilon^2 \{ \sin(\theta_{\rho+1} - \theta_\rho) - \sin(\theta_\rho - \theta_{\rho-1}) \}^2
\end{aligned}$$



Therefore,

$$\boxed{\|\hat{\nabla} E_{\text{eff}}\|^2 = \sum_{\rho} \epsilon^2 \{\sin(\theta_{\rho+1} - \theta_{\rho}) - \sin(\theta_{\rho} - \theta_{\rho-1})\}^2} \quad (4.15)$$

#### 4.4.2 $\hat{\Delta} E_{\text{eff}}$

The second derivative of  $E_{\text{eff}}$  w.r.t. a coordinate  $\theta_{\rho}$  is:

$$\frac{\partial^2 E_{\text{eff}}}{\partial \theta_{\rho}^2} = \frac{\partial^2 \phi}{\partial \theta_{\rho}^2} + \frac{\partial^2 \omega}{\partial \theta_{\rho}^2}$$

Let us further simplify the expression for  $E_{\text{eff}}$ :

$$\begin{aligned} E_{\text{eff}} &= -\sigma - \epsilon \sum_k (1 - \hat{g}_{k+1} \cdot \hat{g}_k) - \sum_{k,k'} f_{k,k'} (1 - \hat{g}_{k+1} \cdot \hat{g}_k) (1 - \hat{g}_{k'+1} \cdot \hat{g}_{k'}) \\ &= -\sigma - N\epsilon - \underbrace{\sum_{k,k'} f_{k,k'}}_A + \underbrace{\left( \epsilon + 2 \sum_{k'} f_{k,k'} \right)}_B \sum_k (\hat{g}_{k+1} \cdot \hat{g}_k) - \sum_{k,k'} f_{k,k'} (\hat{g}_{k+1} \cdot \hat{g}_k) (\hat{g}_{k'+1} \cdot \hat{g}_{k'}) \\ &= A + B \sum_k \cos(\theta_{k+1} - \theta_k) - \sum_{k,k'} f_{k,k'} \cos(\theta_{k+1} - \theta_k) \cos(\theta_{k'+1} - \theta_{k'}) \end{aligned} \quad (4.16)$$

where we used the property shown in (4.13). Thus, the first and consequently the second derivative of  $E_{\text{eff}}$  can be calculated as follows

$$\begin{aligned} \frac{\partial E_{\text{eff}}}{\partial \theta_{\rho}} &= B[\sin(\theta_{\rho+1} - \theta_{\rho}) - \sin(\theta_{\rho} - \theta_{\rho-1})] - 2 \sin(\theta_{\rho+1} - \theta_{\rho}) \sum_{k'} f_{\rho k'} \cos(\theta_{k'+1} - \theta_{k'}) \\ &\quad + 2 \sin(\theta_{\rho} - \theta_{\rho-1}) \sum_{k'} f_{(\rho-1)k'} \cos(\theta_{k'+1} - \theta_{k'}) \\ \implies \frac{\partial^2 E_{\text{eff}}}{\partial \theta_{\rho}^2} &= -B[\cos(\theta_{\rho+1} - \theta_{\rho}) + \cos(\theta_{\rho} - \theta_{\rho-1})] + 2 \cos(\theta_{\rho+1} - \theta_{\rho}) \sum_{k'} f_{\rho k'} \cos(\theta_{k'+1} - \theta_{k'}) \\ &\quad + 2 \cos(\theta_{\rho} - \theta_{\rho-1}) \sum_{k'} f_{(\rho-1)k'} \cos(\theta_{k'+1} - \theta_{k'}) \\ &\quad - 2 \sin(\theta_{\rho+1} - \theta_{\rho}) \left\{ f_{\rho\rho} \sin(\theta_{\rho+1} - \theta_{\rho}) - f_{\rho(\rho-1)} \sin(\theta_{\rho} - \theta_{\rho-1}) \right\} \\ &\quad + 2 \sin(\theta_{\rho} - \theta_{\rho-1}) \left\{ f_{(\rho-1)\rho} \sin(\theta_{\rho+1} - \theta_{\rho}) - f_{(\rho-1)(\rho-1)} \sin(\theta_{\rho} - \theta_{\rho-1}) \right\} \end{aligned}$$

Using the above result, after grouping the terms appropriately, we get

$$\hat{\Delta}E_{\text{eff}} = \sum_{\rho} \left[ -B[\cos(\theta_{\rho+1} - \theta_{\rho}) + \cos(\theta_{\rho} - \theta_{\rho-1})] + 2 \cos(\theta_{\rho+1} - \theta_{\rho}) \sum_{k'} f_{\rho k'} \cos(\theta_{k'+1} - \theta_{k'}) \right. \\ \left. + 2 \cos(\theta_{\rho} - \theta_{\rho-1}) \sum_{k'} f_{(\rho-1)k'} \cos(\theta_{k'+1} - \theta_{k'}) \right. \\ \left. - 2 \sin(\theta_{\rho+1} - \theta_{\rho}) \left\{ f_{\rho\rho} \sin(\theta_{\rho+1} - \theta_{\rho}) - f_{\rho(\rho-1)} \sin(\theta_{\rho} - \theta_{\rho-1}) \right\} \right. \\ \left. + 2 \sin(\theta_{\rho} - \theta_{\rho-1}) \left\{ f_{(\rho-1)\rho} \sin(\theta_{\rho+1} - \theta_{\rho}) - f_{(\rho-1)(\rho-1)} \sin(\theta_{\rho} - \theta_{\rho-1}) \right\} \right]$$

Here, terms like  $B[\cos(\theta_{\rho+1} - \theta_{\rho}) + \cos(\theta_{\rho} - \theta_{\rho-1})]$ , once acted upon by summation over the index  $\rho$ , become equal to each other. Therefore, for simplicity, we can write:

$$\sum_{\rho} B[\cos(\theta_{\rho+1} - \theta_{\rho}) + \cos(\theta_{\rho} - \theta_{\rho-1})] = 2B \cos(\theta_{\rho+1} - \theta_{\rho}) \quad (4.17)$$

On using (4.17), we get the following expression:

$$\hat{\Delta}E_{\text{eff}} = \sum_{\rho} \left[ -2B \cos(\theta_{\rho+1} - \theta_{\rho}) + 4 \cos(\theta_{\rho+1} - \theta_{\rho}) \sum_{k'} f_{\rho k'} \cos(\theta_{k'+1} - \theta_{k'}) \right. \\ \left. - 4f_{\rho\rho} \sin^2(\theta_{\rho+1} - \theta_{\rho}) + 4f_{(\rho-1)\rho} \sin(\theta_{\rho+1} - \theta_{\rho}) \sin(\theta_{\rho} - \theta_{\rho-1}) \right] \quad (4.18)$$

Thus, after using (4.15) and (4.18), the expression for  $V$  is as follows:

$$V = \sum_{\rho} \left[ \epsilon^2 \{ \sin(\theta_{\rho+1} - \theta_{\rho}) - \sin(\theta_{\rho} - \theta_{\rho-1}) \}^2 + 2B \cos(\theta_{\rho+1} - \theta_{\rho}) \right. \\ \left. + 4f_{\rho\rho} \sin^2(\theta_{\rho+1} - \theta_{\rho}) - 4f_{(\rho-1)\rho} \sin(\theta_{\rho+1} - \theta_{\rho}) \sin(\theta_{\rho} - \theta_{\rho-1}) \right. \\ \left. - 4 \cos(\theta_{\rho+1} - \theta_{\rho}) \sum_{k'} f_{\rho k'} \cos(\theta_{k'+1} - \theta_{k'}) \right]$$

Once again, on using the property shown in (4.17), we get the following expression for  $V$ :

$$V = \sum_{\rho} \left[ (2\epsilon^2 + 4f_{\rho\rho}) \sin^2(\theta_{\rho+1} - \theta_{\rho}) - (4f_{(\rho-1)\rho} + 2\epsilon^2) \sin(\theta_{\rho+1} - \theta_{\rho}) \sin(\theta_{\rho} - \theta_{\rho-1}) \right. \\ \left. + 2 \cos(\theta_{\rho+1} - \theta_{\rho}) \left( B - 2 \sum_{k'} f_{\rho k'} \cos(\theta_{k'+1} - \theta_{k'}) \right) \right] \quad (4.19)$$

## 4.5 Expression for $\mathcal{H}_{\text{eff}}$

Now that we have the expression of  $V$ , it is straightforward to write down the final Hamiltonian according to (4.2). From (4.14), we know that  $\hat{\Delta} = \partial_{\theta_{\rho}}^2$ . Thus, using (4.19) and (4.7), we get the following expression for the Hamiltonian:

$$\mathcal{H}_{\text{eff}} = \sum_{\rho} \left[ -\partial_{\theta_{\rho}}^2 + (2\epsilon^2 + 4f_0) \sin^2(\theta_{\rho+1} - \theta_{\rho}) \right. \\ \left. - (4f_1 + 2\epsilon^2) \sin(\theta_{\rho+1} - \theta_{\rho}) \sin(\theta_{\rho} - \theta_{\rho-1}) \right. \\ \left. + 2 \cos(\theta_{\rho+1} - \theta_{\rho}) \left( B - 2 \sum_{k'} f_{\rho k'} \cos(\theta_{k'+1} - \theta_{k'}) \right) \right] \quad (4.20)$$

where,

$$B = \left( \epsilon + 2 \sum_{k'} f_{k,k'} \right)$$

## 4.6 Ground state check for $\mathcal{H}_{\text{eff}}$

Since there were some approximations involved, we also perform a consistency check for the Hamiltonian that we have obtained. Let us take  $\psi_m = \rho$  and apply the Hamiltonian on this

state. Since  $V = \frac{\nabla\rho}{\rho}$ ,  $H\rho = 0$  and hence,  $\rho$  is the ground state of the Hamiltonian. We can also see this by noting the following from (2.21):

$$\lambda_n \sim 1 - \epsilon E_n, \quad \psi_n \sim \frac{\Psi_n}{\rho} \quad (4.21)$$

Since  $\lambda_0 = 1$ ,  $E_0 = 0$  and therefore  $\rho$  is the ground state.

Let us check if  $P(\{\hat{g}_k\}) = e^{-E_{\text{eff}}}$  calculated in (4.9), is the ground state of  $\mathcal{H}_{\text{eff}}$ . We will only keep terms up to first order in  $(1 - \hat{g}_{k+1} \cdot \hat{g}_k)$  for the expression of  $E_{\text{eff}}$ . Thus we have:

$$\begin{aligned} E_{\text{eff}} &\sim -\sigma - \epsilon \sum_k \{1 - \cos(\theta_{k+1} - \theta_k)\} \\ \Psi_0 &\sim 1 + \sigma + (\epsilon + \sigma\epsilon) \sum_k \{1 - \cos(\theta_{k+1} - \theta_k)\} + \frac{\sigma^2}{2} \\ &\quad + \frac{\epsilon^2}{2} \sum_{k,k'} \{1 - \cos(\theta_{k+1} - \theta_k)\} \{1 - \cos(\theta_{k'+1} - \theta_{k'})\} \\ &= 1 + \sigma + \epsilon + \sigma\epsilon + \frac{\sigma^2}{2} + \frac{\epsilon^2}{2} - (\epsilon + \sigma\epsilon + \epsilon^2) \sum_k \cos(\theta_{k+1} - \theta_k) \\ &\quad + \frac{\epsilon^2}{2} \sum_{k,k'} \cos(\theta_{k+1} - \theta_k) \cos(\theta_{k'+1} - \theta_{k'}) \\ \mathcal{H}_{\text{eff}} &\sim \sum_\rho \left[ -\partial_{\theta_\rho}^2 + 2\epsilon \cos(\theta_{\rho+1} - \theta_\rho) + 2\epsilon^2 \sin^2(\theta_{\rho+1} - \theta_\rho) - 2\epsilon^2 \sin(\theta_{\rho+1} - \theta_\rho) \sin(\theta_\rho - \theta_{\rho-1}) \right] \end{aligned}$$

Acting the Laplacian on the ground state gives:

$$\begin{aligned} -\sum_\rho \partial_{\theta_\rho}^2 \Psi_0 &= \sum_\rho \left[ -2 \cos(\theta_{\rho+1} - \theta_\rho) (\epsilon + \sigma\epsilon + \epsilon^2) + \epsilon^2 \{ \sin(\theta_{\rho+1} - \theta_\rho) - \sin(\theta_\rho - \theta_{\rho-1}) \}^2 \right. \\ &\quad \left. + 2\epsilon^2 \cos(\theta_{\rho+1} - \theta_\rho) \sum_k \cos(\theta_{k+1} - \theta_k) \right] \end{aligned} \quad (4.22)$$

Multiplying the potential energy term the ground state gives (up to second order in the

coefficients):

$$\sum_{\rho} V_{\theta_{\rho}} \Psi_0 = \sum_{\rho} \left[ \cancel{2\epsilon \cos(\theta_{\rho+1} - \theta_{\rho})} + 2\epsilon^2 \sin^2(\theta_{\rho+1} - \theta_{\rho}) - \cancel{2\epsilon^2 \sin(\theta_{\rho+1} - \theta_{\rho}) \sin(\theta_{\rho} - \theta_{\rho-1})} + \cancel{2\epsilon\sigma \cos(\theta_{\rho+1} - \theta_{\rho})} \right] \quad (4.23)$$

Since the Hamiltonian is the addition of the terms written in (4.22) and (4.23), we can cancel out the terms that cancel out. Terms struck with same colour cancel each other out. (Note that for one such cancellation, we used the property mentioned in (4.17). Thus, we see that  $\mathcal{H}_{\text{eff}}\Psi_0 \sim 0$ , as expected. As we keep on including higher order terms in (4.23), the terms get cancelled among themselves.

## 4.7 Hamiltonian in the Continuum Limit

Now we calculate the expression for the Hamiltonian in the continuum limit, i.e.  $N_k \rightarrow \infty$ . The discrete Hamiltonian was:

$$\mathcal{H}_{\text{eff}} = \sum_{\rho} \left[ -\partial_{\theta_{\rho}}^2 + (2\epsilon^2 + 4f_0) \sin^2(\theta_{\rho+1} - \theta_{\rho}) - (4f_1 + 2\epsilon^2) \sin(\theta_{\rho+1} - \theta_{\rho}) \sin(\theta_{\rho} - \theta_{\rho-1}) + 2 \cos(\theta_{\rho+1} - \theta_{\rho}) \left( B - 2 \sum_{k'} f_{\rho k'} \cos(\theta_{k'+1} - \theta_{k'}) \right) \right] \quad (4.24)$$

where,

$$B = \left( \epsilon + 2 \sum_{k'} f_{k,k'} \right)$$

Now we calculate the expression for the Hamiltonian in the continuum limit. Definition of  $B$  now becomes:

$$B := \left( \epsilon + 2 \int dk f_{k,k'} \right)$$

To take the continuum limit, we replace  $\theta_{k+1}$  with  $\left(\theta_k + \partial_k \theta_k + \frac{\partial_k^2 \theta_k}{2}\right)$ . Thus the Hamiltonian now becomes:

$$\begin{aligned} \mathcal{H}_{\text{eff}} = \int dk \left[ -\delta_\theta^2 + (2\epsilon^2 + 4f_0) \sin^2\left(\partial_k \theta_k + \frac{\partial_k^2 \theta_k}{2}\right) \right. \\ \left. - (4f_1 + 2\epsilon^2) \sin\left(\partial_k \theta_k + \frac{\partial_k^2 \theta_k}{2}\right) \sin\left(\partial_k \theta_k + \frac{\partial_k^2 \theta_k}{2}\right) \right. \\ \left. + 2 \cos\left(\partial_k \theta_k + \frac{\partial_k^2 \theta_k}{2}\right) \left( B - 2 \int dk' f_{\rho k'} \cos\left(\partial_{k'} \theta_{k'} + \frac{\partial_{k'}^2 \theta_{k'}}{2}\right) \right) \right] \end{aligned} \quad (4.25)$$

Now we expand the sine and cosine terms such that every term in the whole expression is up to fourth order in  $\partial_k \theta_k$ .

$$\begin{aligned} \mathcal{H}_{\text{eff}} = \int dk \left[ -\delta_\theta^2 + 4(f_0 - f_1) \left\{ (\partial_k \theta_k)^2 + \partial_k \theta_k \partial_k^2 \theta_k + \frac{(\partial_k^2 \theta_k)^2}{4} \right\} \right. \\ \left. + 2 \left\{ 1 - \left\{ \frac{(\partial_k \theta_k)^2}{2} + \frac{\partial_k \theta_k \partial_k^2 \theta_k}{2} + \frac{(\partial_k^2 \theta_k)^2}{8} + \frac{(\partial_k \theta_k)^4}{4!} \right\} \right\} \left( B - 2 \int dk' f_{kk'} \right) \right. \\ \left. - 4 \left\{ \frac{(\partial_k \theta_k)^2}{2} + \frac{\partial_k \theta_k \partial_k^2 \theta_k}{2} + \frac{(\partial_k^2 \theta_k)^2}{8} \right\} \int dk' f_{kk'} \left\{ \frac{(\partial_{k'} \theta_{k'})^2}{2} + \frac{\partial_{k'} \theta_{k'} \partial_{k'}^2 \theta_{k'}}{2} + \frac{(\partial_{k'}^2 \theta_{k'})^2}{8} \right\} \right. \\ \left. + 4 \int dk' f_{kk'} \left\{ \frac{(\partial_{k'} \theta_{k'})^2}{2} + \frac{\partial_{k'} \theta_{k'} \partial_{k'}^2 \theta_{k'}}{2} + \frac{(\partial_{k'}^2 \theta_{k'})^2}{8} + \frac{(\partial_k \theta_k)^4}{4!} \right\} \right] \end{aligned} \quad (4.26)$$

Since  $\partial_k \theta_k \partial_k^2 \theta_k = \frac{\partial_k (\partial_k \theta_k)^2}{2}$ , it is a total derivative and it thus vanishes.

$$\begin{aligned} \mathcal{H}_{\text{eff}} = \int dk \left[ -\delta_\theta^2 + 4(f_0 - f_1) \left\{ (\partial_k \theta_k)^2 + \frac{(\partial_k^2 \theta_k)^2}{4} \right\} \right. \\ \left. + 2 \left\{ 1 - \left\{ \frac{(\partial_k \theta_k)^2}{2} + \frac{(\partial_k^2 \theta_k)^2}{8} + \frac{(\partial_k \theta_k)^4}{4!} \right\} \right\} \left( B - 2 \int dk' f_{kk'} \right) \right. \\ \left. - 4 \left\{ \frac{(\partial_k \theta_k)^2}{2} + \frac{(\partial_k^2 \theta_k)^2}{8} \right\} \int dk' f_{kk'} \left\{ \frac{(\partial_{k'} \theta_{k'})^2}{2} + \frac{(\partial_{k'}^2 \theta_{k'})^2}{8} \right\} \right. \\ \left. + 4 \int dk' f_{kk'} \left\{ \frac{(\partial_{k'} \theta_{k'})^2}{2} + \frac{(\partial_{k'}^2 \theta_{k'})^2}{8} + \frac{(\partial_k \theta_k)^4}{4!} \right\} \right] \end{aligned} \quad (4.27)$$

Now we simplify the expression further:

$$\begin{aligned}
\mathcal{H}_{\text{eff}} = \int dk & \left[ -\delta_\theta^2 + \left( 4(f_0 - f_1) - B + 2 \int dk' f_{kk'} \right) \left\{ (\partial_k \theta_k)^2 + \frac{(\partial_k^2 \theta_k)^2}{4} \right\} \right. \\
& + 2 \left( B - 2 \int dk' f_{kk'} \right) \left( 1 - \frac{(\partial_k \theta_k)^4}{4!} \right) + 4 \int dk' f_{kk'} \frac{(\partial_k \theta_k)^4}{4!} \\
& \left. + 4 \left\{ 1 - \left\{ \frac{(\partial_k \theta_k)^2}{2} + \frac{(\partial_k^2 \theta_k)^2}{8} \right\} \right\} \int dk' f_{kk'} \left\{ \frac{(\partial_{k'} \theta_{k'})^2}{2} + \frac{(\partial_{k'}^2 \theta_{k'})^2}{8} \right\} \right]
\end{aligned} \tag{4.28}$$

Thus, we have the following expression for  $\mathcal{H}_{\text{eff}}$  in the continuum limit:

$$\begin{aligned}
\mathcal{H}_{\text{eff}} = \int dk & \left[ -\delta_\theta^2 + 2\epsilon + (4(f_0 - f_1) - \epsilon) \left( (\partial_k \theta_k)^2 + \frac{(\partial_k^2 \theta_k)^2}{4} \right) - 2\epsilon \frac{(\partial_k \theta_k)^4}{4!} + \int dk' f_{kk'} \frac{(\partial_k \theta_k)^4}{3!} \right. \\
& \left. + \left\{ 2 - \left( (\partial_k \theta_k)^2 + \frac{(\partial_k^2 \theta_k)^2}{4} \right) \right\} \left\{ \int dk' f_{kk'} \left( (\partial_{k'} \theta_{k'})^2 + \frac{(\partial_{k'}^2 \theta_{k'})^2}{4} \right) \right\} \right]
\end{aligned} \tag{4.29}$$

We can further simplify this to (after introducing the angular momentum field  $L_k = -i \frac{\delta}{\delta \theta_k}$ )

$$\begin{aligned}
\mathcal{H}_{\text{eff}} = \int dk & \left[ L_k^2 + 2\epsilon + c_1 \left[ (\partial \theta_k)^2 + \frac{(\partial^2 \theta_k)^2}{4} \right] + c_2 \frac{(\partial_k \theta_k)^4}{4!} \right] \\
& - \int dk \int dk' f_{k-k'} (\partial \theta_k)^2 (\partial \theta_{k'})^2 + \mathcal{O}(\partial^6)
\end{aligned} \tag{4.30}$$

where  $c_1 = 4(f_0 - f_1) - \epsilon + 2 \int dk' f_{kk'}$  and  $c_2 = -2\epsilon + 4 \int dk' f_{kk'}$ . Defining the two-component unit field  $\hat{n}_k = (\cos \theta_k, \sin \theta_k)^T$ , we can also write this more explicitly as a 1D, non-local, non-linear sigma model

$$\begin{aligned}
\mathcal{H}_{\text{eff}} = \int dk & \left[ L_k^2 + 2\epsilon + c_1 \left[ (\partial \hat{n}_k)^2 + \frac{(\partial^2 \hat{n}_k) \times \hat{n}_k}{4} \right] + c_2 \frac{(\partial \hat{n}_k)^4}{4!} \right] \\
& - \int dk \int dk' f_{k-k'} (\partial \hat{n}_k)^2 (\partial \hat{n}_{k'})^2 + \mathcal{O}(\partial^6)
\end{aligned} \tag{4.31}$$

## 4.8 Discussion and Outlook

Thus, we obtain the auxiliary quantum many-body Hamiltonian for the Diffusion process defined on the data set of Monte-carlo sampled 1-D 2 band Hamiltonians with chiral symmetry. Such cases of duality are interesting because they often illuminate deep connections between entities that look unrelated.

This kind of treatment also has the potential to uncover interesting things in both the diffusion map algorithm as well the quantum many-body problem. In [4], the quantum problem corresponding to the diffusion map on the data set of  $XY$  model could reveal important properties of the  $XY$  model.

In our case, the low energy spectrum and eigenstates will give us the  $\lambda_n$  and  $\psi_n$  that we obtained after running the Diffusion Map algorithm for classifying the Hamiltonians. In future, we plan to solve the auxiliary quantum Hamiltonian and show that it indeed captures the winding numbers, and that the problem of classifying the sampled Hamiltonians is equivalent to solving this auxiliary Hamiltonian.





# Bibliography

- [1] X.-L. Qi and S.-C. Zhang, *Reviews of Modern Physics* **83**, 1057 (2011).
- [2] A. P. Schnyder, S. Ryu, A. Furusaki, and A. W. Ludwig, *Physical Review B* **78**, 195125 (2008).
- [3] J. Carrasquilla and R. G. Melko, *Nature Physics* **13**, 431 (2017).
- [4] J. F. Rodriguez-Nieva and M. S. Scheurer, *Nature Physics* **15**, 790 (2019).
- [5] M. S. Scheurer and R.-J. Slager, *Physical Review Letters* **124**, 226401 (2020).
- [6] P. W. Anderson, *Science* **177**, 393 (1972).
- [7] K. Binder, *Reports on Progress in Physics* **50**, 783 (1987).
- [8] P. Hohenberg and A. Krekhov, *Physics Reports* **572**, 1 (2015).
- [9] L. LANDAU, *Nature* **138**, 840 (1936).
- [10] M. Z. Hasan and C. L. Kane, *Reviews of modern physics* **82**, 3045 (2010).
- [11] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, *Reviews of Modern Physics* **91**, 045002 (2019).
- [12] L. Wang, *Physical Review B* **94**, 195105 (2016).
- [13] E. P. Van Nieuwenburg, Y.-H. Liu, and S. D. Huber, *Nature Physics* **13**, 435 (2017).
- [14] N. Batra and G. Sheet, *Resonance* **25**, 765 (2020).
- [15] J. K. Asbóth, L. Oroszlány, and A. Pályi, *Lecture notes in physics* **919**, 166 (2016).
- [16] W. P. Su, J. R. Schrieffer, and A. J. Heeger, *Phys. Rev. Lett.* **42**, 1698 (1979).

- [17] R. R. Coifman and S. Lafon, *Applied and Computational Harmonic Analysis* **21**, 5 (2006).
- [18] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, *Applied and Computational Harmonic Analysis* **21**, 113 (2006).



# Appendix

## K-means clustering

Here, we present a brief explanation of the working of K-means clustering algorithm. It is an unsupervised machine learning algorithm which is used to find clusters/groups in a data set.

To see how it works in principle, let's assume we have a data set  $X = \{x_i\}$ ,  $i = \{1, 2, \dots, N\}$ . The K-Means algorithm divides the data set into  $M$  groups ( $M < N$ ),  $G_j$ ,  $j = \{1, 2, \dots, M\}$ . It assigns data points from  $X$  into the groups  $G_j$  in such a way that the assignment minimises the following function:

$$E = \sum_{j=1}^M \sum_{x_i \in G_j} \|x_i - g_j\|^2$$

where  $g_j$  (known as the centroid of the group) is the mean of the points in group  $G_j$ .

Typically, one needs to provide the number of clusters ( $M$ ) one is expecting within the data set. Based on this, the algorithm makes the clusters and assigns a centroid (corresponding the group) to every data point.