Manifold Learning and Convergence of Laplacian Eigenmaps

Mashbat Suzuki

Master of Science

Department of Mathematics and Statistics

McGill University

Montreal, Quebec

July 10, 2015

A thesis submitted to McGill University
in partial fulfilment of the requirements of the degree of Master of Science.

Acknowledgements

This thesis is dedicated to my loving mother Nara, thank you for your love, encouragement and constant support.

I am extremely grateful for both of my supervisors Prof. Dmitry Jakobson and Prof. Gantumur Tsogtgerel, without their patience, guidance and encouragement none of this would have been possible. Thank you for believing in me.

I have been very fortunate to be surrounded by great group of friends. I am thankful for their philosophical discussions and funny anecdotes for making me think and laugh.

To my loving mother Narantungalag

Abstract

In this thesis, we investigate the problem of obtaining meaningful low dimensional representation of high dimensional data, often referred to as manifold learning. We examine classical methods for manifold learning such as PCA and cMDS as well as some modern techniques of manifold learning namely Isomap, Locally Linear Embedding and Laplacian Eigenmaps. The algorithms for these individual methods are presented in mathematically consistent, concise and easy to understand fashion so that people with no computer science background can use the methods presented in their own research. Motivations and justifications of these manifold learning methods are provided. Finally we prove the convergence of Laplacian Eigenmaps method in a self contained and compact fashion following the work of Mikhail Belkin and Partha Niyogi.

Abrégé

Dans cette thèse, on aborde le problème de l'apprentissage de variété afin de réduire la dimensionalité d'ensembles de données pour obtenir une représentation significative de basse dimension. On examine les méthodes classiques d'apprentissage de variété telles que l'analyse en composantes principales (PCA) et le positionnement multidimensionnel classique (cMDS). On présente de plus trois techniques modernes pour résoudre ce problème: les méthodes Isomap, Locally Linear Embedding et Laplacian Eigenmaps. On expose les détails mathématiques de ces quelques algorithmes en termes concis et simples tout en préservant leur cohérence mathématique. Ces développements aideront sans doute d'autres chercheurs sans expérience en informatique à utiliser ces méthodes. Par la suite, on justifie l'applicabilité de ces techniques pour résoudre le problème de réduction de dimension. Finalement, on montre la convergence de la méthode Laplacian Eigenmaps d'une façon compacte en suivant l'approche de Mikhail Belkin et de Partha Nyogi.

TABLE OF CONTENTS

CHAPTER 1
Introduction

In the last few decades we have witnessed remarkable developments in computing power and data storage technology. With these developments came a tremendous growth of available data which opened new horizons in all aspects of our lives, including technology, government, business, and sciences. However today we are often faced with the challenges that come with big data. We are constantly flooded with large amounts of data from which we are tasked to extract meaningful information. These recent challenges led to the creation of new disciplines such as data mining and machine learning, which are flourished by researchers in applied mathematics, statistics and computer science.

Nowadays there are many different approaches to learning from data, and the study of these learning algorithms is called machine learning. Real world problems require deriving function estimates from a large number of data sets with many variables. Although having access to abundance of examples with many features is beneficial to an algorithm attempting to generalize from data, handling large number of variables or dimensions often adds to the complexity of the problem making it difficult to perform useful inference.

It is well known that high dimensionality presents obstacle to efficient processing of data, which phenomenon is often referred as *curse of dimensionality* as coined by Bellman in [5]. The heart of the challenge is that as dimensionality D increases the volume of the space increase too fast so that the available data becomes sparse. Here are some manifestations of curse of dimensionality:

- When D is large not all norms are numerically equivalent in $\mathbb{R}^D$. Notably the same function will have different degrees of smoothness under different norms.
- In order to estimate multivariate functions with the same accuracy as functions in low dimensions, we require that the sample size n grow exponentially with D.
- Many algorithms that are fast in low dimensions become extremely slow in high dimensions. In particular most nearest neighbour search algorithms scale exponentially in complexity when vertex of a graph has on average D edges.

Although there are some techniques to partially fix these problems in specific tasks such as in [27], curse is still difficult to overcome in general situations when only working in high dimensions. Thus there is a demand for techniques that are designed to reduce dimensions of the data, hence effectively avoiding the curse of dimensionality.

In many cases high dimensionality is an artefact of the choice of representation of data which has nothing to do with underlying complexity of the mechanisms that generated the data. Often the variables involved in the representation are correlated through some functional dependence, hence the number of independent variables necessary to efficiently describe the data is small. In such scenario it is possible to represent the data in much fewer dimensions than the dimensions of the original data, and this is referred as dimensionality reduction.

For instance, consider 128×128 gray scale image of a rectangle. Now consider data set of images where the same rectangle is translated around. Each image is represented as a point in $\mathbb{R}^{128\times 128}$, hence the data set is a subset of very high dimensional space but the underlying parameter governing the data is only two dimensional namely translation in the two axis (A slightly sophisticated example is given in Figure 1–1). As such under the assumption of low internal degrees of freedom, it is reasonable to transform the representation of the data to more efficient description by reducing the dimensionality. Developing such methods of dimensionality reduction and studying their properties is the main goal of the field *Manifold Learning*.

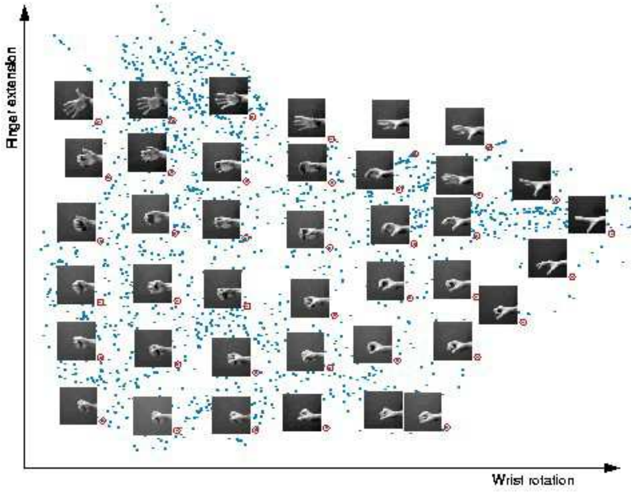

Figure 1–1: Example of dimension reduction from $\mathbb{R}^{64\times64}$ to $\mathbb{R}^2$. Sequence of pictures of opening and closing movements of the hand at different wrist orientations, each picture (64 pixel by 64 pixel) is treated as point in $\mathbb{R}^{64\times64}$. There there are 2000 pictures. When manifold learning method(Isomap[35]) is applied to the data, the algorithm detects two main degrees of freedom "finger extension" and "wrist rotation". Furthermore it correctly parametrizes the data according to these parameters

Recently many manifold learning algorithms have been put forward and implemented for real world problems successfully. These methods include Locally Linear Embedding (LLE), Isomap, and Laplacian Eigenmaps. However many of the aforementioned techniques are still in their adolescence, waiting for the theory to catch up. The purpose of this thesis is to give a self contained overview of manifold learning and dimensionalilty reduction from a mathematically rigorous viewpoint.

1.1 Outline of Thesis

In Chapter 2, we discuss classical methods for manifold learning such as PCA and cMDS, as well as some modern techniques of manifold learning namely Isomap, Locally Linear Embedding, and Laplacian Eigenmaps. For each method, the algorithm is given in compact, concise fashion. In addition, an intuitive justification for the algorithm is provided. For Laplacian Eigenmaps algorithm, we justify in detail the weights used for constructing the corresponding graph Laplacian.

In Chapter 3, we explore the relationships between the manifold learning methods presented in Chapter 2. Advantages and disadvantages of local and global approaches to manifold learning is discussed. We show that under certain assumptions, Laplacian Eigenmaps and Locally Linear Embedding are equivalent. Computational complexity of the manifold learning methods are studied. We also introduce new algorithms such as c-Isomap, L-Isomap, and hLLE, which are variants of the methods surveyed in Chapter 2. Finally, we explore some important issues facing manifold learning.

Both Chapter 4 and Chapter 5 are devoted to understanding the Laplace operator and the heat equation on manifolds so that we are equipped to understand the proof of convergence of Laplacian Eigenmaps algorithm in the last chapter. Standard techniques in differential geometry such as exponential mapping and computation of curvature and useful properties of heat operator are stated. We also give the short time asymptotic expansion of the heat kernel on manifolds.

Finally, in Chapter 6, we give a proof of convergence of Laplacian Eigenmaps method in a self contained and compact fashion, following the work of Mikhail Belkin and Partha Niyogi.

CHAPTER 2
Manifold Learning Algorithms

Manifold Learning is considered as subfield of machine learning. Hence we start this chapter with basic ideas and goals of machine learning. Excellent introduction to machine learning is given in [1].

Machine learning emerged out of a branch of computer science known as *artificial intelligence*(AI). As the name suggests the aim of the discipline is to make machines intelligent, in the sense that making rational decisions and solve problems autonomously. Since intelligence depends largely on the ability to learn, one of the central focus of AI is to make machines learn from previous experience or recognize patterns and structures in data.

In a broad sense machine is said to learn from experience with respect to some task if its performance on specific task improves with experience. However problem of learning is very general and its difficult to pinpoint a general definition, but almost any question in statistics has an anologue in machine learning.

Machine learning algorithms are data-driven. In other words the data itself reveals the proper answer. For instance if we are given a problem of distinguishing if an email is a spam or not, machine learning algorithm will go through examples of spam and non-spam messages and extract a patterns from these examples. These patterns and information that is extracted through machine learning is then used to predict whether or not an unseen email is spam. Also what can be considered spam changes in time, culture and from person to person, the coded definition has to be changed according to these circumstances, but machine learning algorithms can be applied regardless of circumstances as long as there are training data.

The idea demonstrated above of learning from experience (data) is fundamental to many types of problems. There are many ways of learning from data, so there are various categories of machine learning. Two of most common categories is *supervised learning* and *unsupervised learning*. The particular example discussed of spam filtering is an example of supervised learning. This means that the examples are labelled. In our case we explicitly know which messages are spam and which are not in the example data set.

In an unsupervised learning algorithm the example data are not labelled. The goal of these types of algorithms is cluster the examples into different groups or find structure in these examples.

SUPERVISED LEARNING: Learning algorithm receives set of input variables and their corresponding output variable. The problem is to find a function of the input variables that approximates the known output variable.

UNSUPERVISED LEARNING: Learning algorithm receives set of unlabelled data. The problem is to find hidden structure in this data.

In an essence supervised learning is the study of the relationship between input and output variables. On the other hand unsupervised learning is the study the particular characteristics of the input variables only.

There are algorithms that are in-between the two types mentioned above such as semi-supervised learning and active learning. In this thesis we will mainly consider subclass of unsupervised learning which is dimension reduction or manifold learning. The aim is to uncover low dimensional structures in a high dimensional unlabelled data.

In this chapter we will explore some existing manifold learning techniques and their descriptions. One of the first and most common method of dimension reduction is principle component analysis (PCA). PCA is a manifold learning algorithm where the underlying structure of data is linear. Since manifolds are locally linear it is important to be able to do linear dimensional reduction. In fact many of manifold learning methods can be thought of as non-linear versions of PCA. Detailed introduction to PCA and related methods are introduced in [22]

2.1 Principle Component Analysis

Suppose that the data consist of points $\{x_1, \cdots, x_n\} \in \mathbb{R}^D$. Without loss of generality assume the empirical mean of the data set is zero. We are looking for d-dimensional subspace along which the data has the maximum variance. If the data points lie exactly on some d-dimensional linear subspace then PCA will recover exactly the subspace, otherwise there would be some error. We would like to find a direction where the variance is maximized, first we construct the so-called data matrix $X = (x_1, \cdots, x_n)^T \in \mathbb{R}^{n \times D}$. The rows of X are the D-dimensional data points. Then for $d = 1$ maximizing variance is equivalent the following:

$$\begin{aligned}
\max_{\|w\|=1} Var(Xw) &= \max_{\|w\|=1} \mathbb{E}[(Xw)^2] - \mathbb{E}[Xw]^2 \\
&= \max_{\|w\|=1} \mathbb{E}[(Xw)^2] \\
&= \max_{\|w\|=1} \frac{1}{n} \sum_{i=1}^{n} (x_i^T w)^2 \\
&= \max_{\|w\|=1} \frac{1}{n} \sum_{i=1}^{n} w^T x_i x_i^T w \\
&= \max_{\|w\|=1} \frac{1}{n} w^T X^T X w
\end{aligned}$$

Hence the direction which maximizes the variance can be found by the optimization formula.

$$v_1 = \underset{\|w\|=1}{\text{argmax}}\{w^T X^T X w\}$$

Note that above quantity maximised is the Rayleigh quotient. The maximum is the largest eigenvalue which occurs when w is the corresponding eigenvector. Similar approach can be taken for the higher order eigenvectors.

PCA dimension reduction from $\mathbb{R}^D$ to $\mathbb{R}^d$:

- **Step 1:** Given arbitrary data set $S = \{x_i\}_{i=1}^n$ in $\mathbb{R}^D$, construct data matrix $\hat{X} = (x_1, \cdots, x_n)^T \in \mathbb{R}^{n \times D}$

- **Step 2:** Center the data to have empirical mean zero, in other words transform $\hat{X}$ so that columns add up to zero. This can be accomplished by multiplying the data matrix by centralizing matrix $H = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$. The new data matrix is then $X = H\hat{X}$

- **Step 3:** Compute eigenvalues and eigenvectors of sample covariance matrix $X^T X$. Order first d-eigenvectors v_i so that corresponding eigenvalues are in descending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$

- **Step 4:** Construct $V = (v_1, \cdots, v_d)$, and compute $XV \in \mathbb{R}^{n \times d}$. The i-th row vector of XV corresponds to the d-dimensional representation of $x_i \in \mathbb{R}^D$.

There is another way to interpret the d-dimensional representation of data points in S through the algorithm is via the following map:

$$\Phi \colon S \subset \mathbb{R}^D \longrightarrow \mathbb{R}^d$$

$$x_i \mapsto (x_i \cdot v_1, \cdots, x_i \cdot v_d).$$

PCA has been applied to variety of applied problems such as image processing, statistics, text mining and facial recognition. However there are obvious drawbacks to the method, clear one being that the centralized data is required to lie on a linear subspace or something very close to it. This is very strong assumption since it assumes that the variables of the data are correlated in a linear fashion which is not true in many applications. We will introduce other manifold learning algorithms that lifts this linearity assumption such as Isomap, Locally Linear Embedding and Laplacian Eigenmaps. Next we will examine another manifold learning method similar to PCA called Multidimensional Scaling(MDS), this method is used later on for non-linear dimensional reduction technique called Isomap.

2.2 Multidimensional Scaling

Multidimensional Scaling aims to preserve pairwise distances between data points while reducing dimension. For instance, given data set $S = \{x_i\}_{i=1}^n$ in $\mathbb{R}^D$ we would like find to similar data points $\{y_i\}_{i=1}^n \in \mathbb{R}^d$ such that the pairwise distances are preserved. There are different kinds of multidimensional scaling techniques in this section we will examine classical multidimensional scaling which was first introduced by Torgesen in [36]. More modern treatment of the subject and application are can be found in [7].

Classical Multidimensional Scaling(cMDS) from $\mathbb{R}^D$ to $\mathbb{R}^d$:

- **Step 1:** Given data set $\{x_i\}_{i=1}^n$ in $\mathbb{R}^D$, construct matrix of dissimilarities $D_{ij} = \|x_i - x_j\|^2$
- **Step 2:** Set $B = -\frac{1}{2}HDH$ where $H = I - \frac{1}{n}\mathbf{1}_n\ \mathbf{1}_n^T$, this step centralizes D.

- **Step 3:** Find spectral decomposition $B = U\Lambda U^T$. Which by spectral theorem $\Lambda = \textbf{diag}(\lambda_1, \cdots, \lambda_n)$ and $U = (v_1, \cdots, v_n)$, where v_i is the eigenvector corresponding to the i-th largest eigenvalue λ_i.

- **Step 4:** The d-dimensional spatial configuration of the data set is $Y = U_d \Lambda_d^{\frac{1}{2}} = (\sqrt{\lambda_1} v_1, \cdots, \sqrt{\lambda_d} v_d) = (y_1, \cdots, y_n)^T$. The d-dimensional representation of data x_i is then given by y_i.

The MDS algorithm described here is used in wide variety of applications such as in surface matching [9], marketing theory [20], and in psychometrics [34]. Major draw back of cMDS is in its sensitivity to noise and it doesn't work well when the underlying structure of the data is nonlinear. However cMDS has inspired nonlinear manifold learning technique Isomap which we will cover next.

2.3 Isomap

A useful point of view towards manifold learning is to treat the observations $\{x_i\}_{i=1}^n \in \mathbb{R}^D$ as image of a point y_i randomly sampled from a domain $\mathcal{M}$ in $\mathbb{R}^d$ under some map $\Psi : \mathcal{M} \subset \mathbb{R}^d \to \mathbb{R}^D$ such that $\Psi(y_i) = x_i$. Objective of manifold learning is then to recover $\mathcal{M}$ and Ψ. However as stated the problem is ill posed since for any observed data and for any $\mathcal{M}$ we can find Ψ that satisfies the condition, thus there needs to be restrictions for Ψ. Two main possibilities for the restrictions are to require Ψ to be either isometric embedding or conformal embedding.

Isometric Feature Mapping(Isomap) exploits Ψ to be isometric embedding. The method was first introduced in [35], the algorithm uses cMDS as subroutine. Unlike previously mentioned methods such as PCA and cMDS, Isomap can discover nonlinear degrees of freedom of the underlying data.

The effectiveness of Isomap is due to the fact that some important geometric aspects of $\mathcal{M}$ are preserved under Ψ. For instance geodesic distance $d_g(a,b)$ between two points $a, b \in \mathcal{M}$ is same as geodesic distance between $\Psi(a), \Psi(b)$ on $\Psi(\mathcal{M})$. Hence the geodesic distances are preserved, which means that $\Psi(\mathcal{M})$ and $\mathcal{M}$ are isometric when viewed as metric space under geodesic distance. Thus if we can approximate the geodesic metric for $\Psi(\mathcal{M})$, then we can use cMDS to get a lower dimensional point configuration with the same metric structure. The dimension reduced point configuration can be thought of as random samples from $\mathcal{M}$. Hence Isomap approximates geodesic metric for $\Psi(\mathcal{M})$. The approximation is done through constructing k-nearest neighbour graph on the data and utilize local linearity of the manifold $\Psi(\mathcal{M})$. The geodesic distance between neighbouring points are well approximated by the observed euclidean distance, however for far away points euclidean distance is no longer good approximation of geodesic. To overcome this we will add each small geodesics in the k-neighbourhood graph until the final point is reached. Such path with minimal distance on the k-nearest neighbourhood graph is an approximation for the geodesic as shown in [6]. With such approximation we can capture the intrinsic geometry of $\Psi(\mathcal{M})$.

Isomap nonlinear dimension reduction from $\mathbb{R}^D$ to $\mathbb{R}^d$:

- **Step 1:** Given data set $\{x_i\}_{i=1}^n$ in $\mathbb{R}^D$, form the weighted k-nearest neighbour graph G. That is put edge between vertices x_i and x_j if x_j is one of the k nearest neighbours of x_i or vice versa in the Euclidean distance. Assign each edge (x_i, x_j) weight $\|x_i - x_j\|$.

- **Step 2:** Compute the shortest path distances between all pair of vertices x_i and x_j store it in S_{ij}. This can be done using Floyd-Warshall's or

Dijkstra's algorithm. Store the square of the shortest path into distance matrix $D_{ij} = S_{ij}^2$

- **Step 3:** Apply cMDS algorithm with dissimilarity matrix D from previous step.

The algorithm is shown experimentally to perform well under some assumptions on $\mathcal{M}$. In the example shown in Figure 2–1 manifold $\mathcal{M}$ is flat two dimensional and the set of observations has three dimensions where Ψ is non-linear isometric embedding. In some special types of manifolds Isomap is guaranteed to recover $\mathcal{M}$.

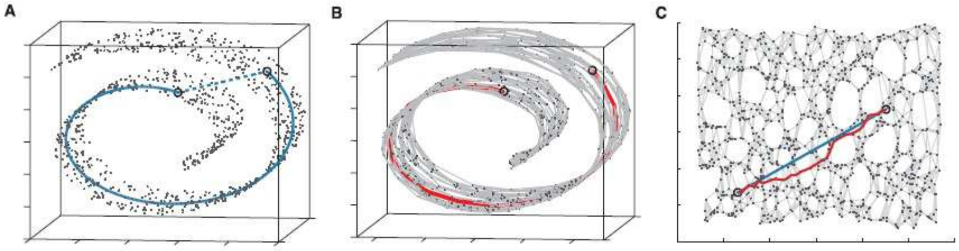

Figure 2–1: Illustration of Isomap for "Swiss roll" data set where $k = 7$ and $n = 1000$ as in [35]. Blue line is geodesic distance on $\Psi(\mathcal{M})$ while red line is approximation of the geodesic distance using shortest path on neighbourhood graph

Note in Figure 2–1A,B how the geodesic is approximated by the shortest path on graph G. In order to perform the geodesic length approximation we need to apply either Floyd-Warshall's or Dijkstra's algorithm. In practice for moderate amount of dataset Floyd-Warshall's algorithm is used due to its simple nature.

Floyd-Warshall algorithm shown above is known to be $\Theta(n^3)$ details can be found on [18]. The constant in $\Theta(n^3)$ is small since there is only one operation

Algorithm 1 Floyd-Warshall's

```
Let $S_{ij}$ be $n \times n$ matrix of minimal path distances initialized to $\infty$
for each vertex $x_i$
$S_{ii} \leftarrow 0$
for each edge $(x_i, x_j)$
$S_{ij} \leftarrow \|x_i - x_j\|$
for k from 1 to n
    for i from 1 to n
        for j from 1 to n
            if $S_{ij} > S_{ik} + S_{kj}$ then
                $S_{ij} \leftarrow S_{ik} + S_{kj}$
return $S_{ij}$
```

in the inner most loop and the algorithm is easy to implement thus it is commonly used. However for large data set the Floyd-Warshall is too slow due to its cubic growth. For large data it is advised to use Dijkstra's algorithm with Fibonacci heap since it has lower asymptotic time complexity [12].

The main obstacle of Isomap is the estimation of geodesic as stated before. The following theorem provides insight into when such difficulty can be overcame through Isomap.

Theorem 2.1 ([13])**.** *Let $\mathcal{M}$ be sampled from a bounded convex region in $\mathbb{R}^d$, with respect to some density function α. Let Ψ be C^2 isometric embedding of the region in $\mathbb{R}^D$. Given $\epsilon, \mu > 0$, for a suitable choice of neighbourhood size k, we have*

$$1 - \epsilon \leq \frac{\textit{recovered distance}}{\textit{original distance}} \leq 1 + \epsilon$$

with probability at least $1 - \mu$, provided that the sample size is sufficiently large.

The theorem 2.1 provides a theoretical guarantee for Isomap when $\mathcal{M}$ is sufficiently regular, convex and compact. Isomap has been applied for computer vision in video analysis [30], and also in face recognition [38].

2.4 Locally Linear Embedding

Locally Linear Embedding(LLE) was first introduced in [32] around the same time as Isomap. The main idea of LLE is to treat the manifold as a collection of overlapping coordinate patches each one being nearly linear. Intuition is then to encode the local information using the k-nearest neighbour as samples from linear patch, and characterize the geometry of these patches. In order to do this we express the vertex as a weighted combination of its neighbours. We find a matrix W which satisfies following optimality condition:

$$W = \operatorname*{argmin}_{\hat{W}} \sum_{i=1}^{n} \|x_i - \sum_{j \in N(i)} \hat{W}_{ij} x_j\|^2 \tag{2.1}$$

subject to invariance constraint $\sum_j \hat{W}_{ij} = 1$ for each $1 \leq i \leq n$, and sparseness constraint $\hat{W}_{ij} = 0$ if $x_i \notin N(j)$. The first constraint $\sum_j \hat{W}_{ij} = 1$ makes the weights invariant to global translations, global rotations and scaling. For instance to see invariance under global translation observe the following:

$$\begin{aligned}
\|(x_i + c) - \sum_{j \in N(i)} \hat{W}_{ij}(x_j + c)\| &= \|x_i + c - \sum_{j \in N(i)} \hat{W}_{ij} x_j - \sum_{j \in N(i)} \hat{W}_{ij} c\| \\
&= \|x_i + c - \sum_{j \in N(i)} \hat{W}_{ij} x_j - c\| \\
&= \|x_i - \sum_{j \in N(i)} \hat{W}_{ij} x_j\|
\end{aligned}$$

The weight matrix W reveals the local geometry of the embedded manifold. In order to use this in practice, we would like to find an algorithmic way of finding optimal solution to the optimization problem 2.1. For fixed x_i and for given $\hat{W}$ we may write

$$\|x_i - \sum_{j \in N_k(i)} \hat{W}_{ij} x_j\|^2 = \| \sum_{j \in N(i)} \hat{W}_{ij}(x_i - x_j)\|^2 = \hat{W}_i^T \Sigma^{(i)} \hat{W}_i$$

where $\hat{W}_i = (\hat{W}_{i1}, \cdots \hat{W}_{in})^T$ and $\Sigma^{(i)}_{\mu\nu} = (x_i - x_\mu)^T(x_i - x_\nu)$ which is called local covariance matrix. Note that $\Sigma^{(i)}$ is symmetric non-negative definite for each data point x_i. We can solve the optimization problem using Lagrange multiplier method which gives

$$f(\hat{W}) = \hat{W}_i^T \Sigma^{(i)} \hat{W}_i - \lambda(\mathbf{1}_n^T \hat{W}_i - 1)$$

Differentiating above respect to $\hat{W}$ and setting it equal to zero solves the optimization problem 2.1 with optimal weight

$$W_i = \frac{(\Sigma^{(i)})^{-1}\mathbf{1}_n}{\mathbf{1}_n^T(\Sigma^{(i)})^{-1}\mathbf{1}_n}$$

.

The local geometry at each data point x_i is captured by W_i. We saw that W_i was invariant under rotations, scaling and translations. The combination of the two facts imply that any linear mapping of the neighbourhood of x_i is also characterized by W_i. To reduce dimensions we need to find point configuration $\{y_i\}_{i=1}^n \in \mathbb{R}^d$ with the same local geometry as observations $\{x_i\}_{i=1}^n$, meaning that configuration that is best characterized by W. Hence d-dimensional configuration $\{y_i\}_{i=1}^n$ is obtained by minimizing

$$\sum_{i=1}^n \|y_i - \sum_{j=1}^n W_{ij} y_j\|$$

Thus we need to solve for $Y = (y_1, \cdots, y_n)$ from the following optimization problem

$$Y = \underset{\hat{Y}}{\operatorname{argmin}} \sum_{i=1}^n \|\hat{y}_i - \sum_{j=1}^n W_{ij} \hat{y}_j\| \tag{2.2}$$

since W is invariant under translations we may add constraint that the mean is zero $\sum_{i=1}^n y_i = 0$ and has unit covariance i.e $\frac{1}{n} Y Y^T = I_d$. It is shown in [33] that the optimization problem 2.2 is solved when $Y = (v_1, \cdots, v_d)^T$

where v_i is the eigenvector corresponding to the $(i+1)$-st smallest eigenvalue of $(I-W)^T(I-W)$. Summarizing the steps mentioned above the algorithm is presented below.

Locally Linear Embedding(LLE) dimensional reduction from $\mathbb{R}^D$ to $\mathbb{R}^d$:

- **Step 1:** Given data set $\{x_i\}_{i=1}^n$ in $\mathbb{R}^D$, form the weighted k-nearest neighbour graph. Compute local covariance matrix for each data x_i denoted $\Sigma_{\mu\nu}^{(i)} = (x_i - x_\mu)^T(x_i - x_\nu)$ for $x_\mu, x_\nu \in N_k(x_i)$.
- **Step 2:** Find optimal reconstruction weights are each x_i.

  $$W_i = \frac{(\Sigma^{(i)})^{-1}\mathbf{1}_n}{\mathbf{1}_n^T(\Sigma^{(i)})^{-1}\mathbf{1}_n}$$

  further set $W_{ij} = 0$ if $x_j \notin N_k(x_i)$
- **Step 3:** Compute eigenvalues and the eigenvectors of $(I-W)^T(I-W)$, denote v_t as the eigenvector corresponding to the $(t+1)$-st smallest eigenvalue. The d-dimensional spatial configuration of the data set is obtained by $Y = (y_1, \cdots, y_n) = (v_1, \cdots, v_d)^T$. Dimension reduced form of data point x_i is given by y_i

Note that LLE solves sparse eigenvalue problem. LLE is similar to another method called Laplacian Eigenmaps which we will introduce in next section. The LLE method is known to have applications to facial recognition [39], hand gesture recognition [19], speech and music analysis [23]. The method is used often due to its speed and simplicity.

2.5 Laplacian Eigenmaps

Laplacian Eigenmaps introduced in [2], relies on ideas from spectral geometry and spectral graph theory. In spirit Laplacian Eigenmaps is close to LLE, it also tries to capture information about the local geometry and reconstruct global geometry from the local information. We first introduce the algorithm and give justification later on.

Laplacian Eigenmaps dimensionality reduction from $\mathbb{R}^D$ to $\mathbb{R}^d$:

- **Step 1**: Given data set $\{x_i\}_{i=1}^n$ in $\mathbb{R}^D$ put edge between data x_i and x_j if they are near in the sense of one of the following two choices:

  1. **k-nearest neighbour**: Vertices x_i and x_j are connected by an edge if either x_j is among the k closest neighbours of x_i or x_i is among the k closest neighbours of x_j.

  2. **ϵ-neighbourhoods**: Vertices x_i and x_j are connected by and edge if $\|x_i - x_j\| < \epsilon$ where the norm is the usual euclidean norm in $\mathbb{R}^D$.

- **Step 2**: We have two options for choosing the edge weights

  1. **Heat Kernel:** If x_i and x_j are connected by an edge then set the edge weight as

  $$W_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}}$$

  otherwise, set $W_{ij} = 0$.

  2. **Combinatorial:** $W_{ij} = 1$ if the vertex i and j are connected by an edge and $W_{ij} = 0$ otherwise. This choice of weights avoids the need to choose a parameter t, making it more convenient to apply.

- **Step 3**: Let G be the graph constructed according to previous two steps. Furthermore assume G is connected graph otherwise apply the current step to each connected component of G. Compute eigenvalues and eigenvectors of the generalized eigenvalue problem

  $$Lf = \lambda D f$$

where D is diagonal weight matrix called degree matrix, and its entries are $D_{ii} = \sum_{j \in N(x_i)} W_{ji}$. We call $L = D - W$ graph Laplacian matrix. By the spectral theorem we know that the eigenvalues are real. We order the eigenvalues in an increasing order $\lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_d$ let ϕ^i be corresponding eigenvectors such that $L\phi^i = \lambda_i D\phi^i$. We leave out the zeroth eigenvector (since it is constant) and proceed with the embedding using the following map

$$\begin{aligned} \Phi \colon \mathcal{M} &\longrightarrow \mathbb{R}^D \\ x_i &\mapsto (\phi_i^1, \cdots, \phi_i^d). \end{aligned}$$

where ϕ_i^j stands for i-th component of j-th eigenvector.

The idea behind the method is to map close together points to close together points in the new dimension reduced space. For instance if we construct the graph G according to the algorithm description and looking to reduce dimension to d, then reasonable approach is to find points $\{y_i\}_{i=1}^n \in \mathbb{R}^d$ such that the following functional is minimized

$$\sum_{ij} \|y_i - y_j\|^2 W_{ij}$$

with some restrictions on $\{y_i\}_{i=1}^n$ to avoid trivial solutions. Let $Y = (y_1, \cdots, y_n)$, then the functional can be written as the following:

$$\sum_{ij} \|y_i - y_j\|^2 W_{ij} = \mathbf{tr}(YLY^T)$$

As a result the problem reduces to constrained optimization problem where the constraint is $YDY^T = I$, this prevents a collapse onto subspace of dimension smaller than d. The standard methods show that the optimal solution

$$Y = \underset{\hat{Y}D\hat{Y}^T = I}{\operatorname{argmin}} \mathbf{tr}(\hat{Y}L\hat{Y}^T)$$

can be found through and eigenvalue problem $L\phi_i = \lambda_i D\phi_i$ and the solution is $Y = (\phi_1, \cdots, \phi_d)^T$ where ϕ_i is the eigenvector corresponding to $(i+1)$-st smallest eigenvalue. The procedure explained here gives justification to the algorithm except the choice of the weight W_{ij}, we will explain the reasoning behind the choice of weight W_{ij}.

Weight matrix W and Laplace Beltrami Operator

The weights are chosen so that the operator L approximates the Laplace Beltrami operator on manifold.The operator $D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$ is also called the normalized discrete Laplace operator and it has been extensively studied in [10].

Heat equation is closely related to the Laplace Beltrami operator. Let $f : \mathcal{M} \to \mathbb{R}$ be the initial heat distribution and $u(x,t)$ be the heat distribution at time t. The heat equation is written $(\partial_t + \Delta_{\mathcal{M}})u = 0$. The solution to the heat equation is obtained by the following

$$u(x,t) = \int_{\mathcal{M}} H_t(x,y) f(y) dV_y$$

Where H_t is the heat kernel on the manifold $\mathcal{M}$. When x and y are close enough , and for small time the manifold heat kernel is approximated by the Euclidean heat kernel

$$H_t(x,y) \approx \frac{1}{(4\pi t)^{\frac{d}{2}}} e^{-\frac{\|x-y\|^2}{4t}}$$

For small time heat kernel becomes increasingly localized tends to Dirac δ-function, in other words

$$\lim_{t\to 0} \int_{\mathcal{M}} H_t(x,y) f(y) = \delta * f(x) = f(x)$$

Using the definition of heat kernel and heat equation we may write Laplace Beltrami operator as

$$\Delta_M f = -\left[\partial_t \int_M H_t(x,y) f(y)\right]_{t=0}$$

We may approximate derivative above as the difference

$$\Delta_{\mathcal{M}} f(x) \approx \frac{1}{t}\left[f(x)\frac{1}{(4\pi t)^{\frac{d}{2}}}\int_{\mathcal{M}} e^{-\frac{\|x-y\|^2}{4t}} dV_y - \frac{1}{(4\pi t)^{\frac{d}{2}}}\int_{\mathcal{M}} e^{-\frac{\|x-y\|^2}{4t}} f(y) dV_y\right]$$

If we assume that the points $\{x_i\}$ are distributed dense enough, then we may replace the integral as the empirical estimate hence

$$\Delta_{\mathcal{M}} f(x_i) \approx \frac{1}{t}\frac{1}{(4\pi t)^{\frac{d}{2}}}\frac{1}{|V|}\left[f(x_i)\sum_{x_j \in N_\epsilon(x_i)} e^{-\frac{\|x_i-x_j\|^2}{4t}} - \sum_{x_j \in N_\epsilon(x_i)} e^{-\frac{\|x_i-x_j\|^2}{4t}} f(x_j)\right]$$

Note that above in the bracket can be identified as the i'th component Lf with

$$W_{ij} = \begin{cases} e^{-\frac{\|x_i-x_j\|^2}{4t}} & \text{if} \quad \|x_i - x_j\| < \epsilon \\ 0 & \text{otherwise} \end{cases}$$

This is exactly the weight formula for Laplacian Eigenmaps which justifies the choice of W. Hence we have the following approximation involving graph Laplacian

$$\Delta_{\mathcal{M}} f(x_i) \approx \frac{1}{t}\frac{1}{(4\pi t)^{\frac{d}{2}}}\frac{1}{|V|}[Lf]_i$$

note that the constant in front of the discrete Laplacian only scales the eigenvectors but the important information about the embedding resides in the relationship between the eigenvectors. In 6 we will discuss convergence of discrete Laplacian to Laplace-Beltrami operator in providing theoretical justification to the Laplacian Eigenmaps algorithm.

CHAPTER 3
Relationships Between Learning Algorithms

3.1 Global versus Local methods

The manifold learning methods we studied so far can be divided into two big categories *global* and *local* methods. Usually the local method is equivalent to solving sparse eigenvalue problem while global method is to solve dense eigenvalue problem. The local methods include Laplacian Eigenmaps, LLE where the method uses the local information to construct embedding. On the other hand global method uses information between all pairs of data to construct embedding. The Isomap algorithm is an example of global method [13]. There are methods that belong to both of these categories such as Semidefinite embedding [37] . Semidefinite embedding is considered local in the sense that neighbourhood distances are equated with a geodesic distances but also considered global method since the objective function considers distances between all pair of data points. There are advantages and disadvantages to both types of methods, choosing which method to use depends heavily on the underlying geometry of the problem.

Local methods work well for characterizing the local geometry of manifolds accurately, however the method is not very effective at the global scale giving inaccurate description of geometry at far away points. The reason is simple, the constraints from the local method do not apply to far away points. On the other hand global methods such as Isomap gives correct description of distances of far away data points but provides inaccurate intra-neighbourhood distances. However local methods in general tend to handle sharp curvatures better than global methods.

Methods in each category share some common characteristics. For instance local methods LLE and Laplacian Eigenmaps are considered same under certain circumstances which we will explore next section.

3.2 Connection Between LLE and Laplacian Eigenmaps

In [2] the authors showed that under certain assumptions LLE reduces to Laplacian Eigenmaps. Recall from 2.4 that performing LLE is equivalent to finding eigenvalues and eigenvectors of matrix $E = (I - W)^T(I - W)$ where W satisfies

$$W = \underset{\hat{W}}{\operatorname{argmin}} \sum_{i=1}^{n} \|x_i - \sum_{j \in N(i)} \hat{W}_{ij} x_j\|^2$$

subject to invariance constraint $\sum_j W_{ij} = 1$ for each $1 \leq i \leq n$ and sparseness constraint $W_{ij} = 0$ if $x_i \notin N(j)$. We will show that under some conditions the following hold

$$Ef \approx \frac{1}{2} L^2 f \tag{3.1}$$

where L is the Laplacian matrix mentioned in Laplacian Eigenmaps method.

To demonstrate 3.1 we fix a data point x_i, we show that

$$[(I - W)f]_i \approx \sum_j W_{ij}(x_i - x_j)^T H(x_i - x_j) \tag{3.2}$$

where H is the Hessian matrix of f at x_i. To demonstrate 3.2 we change coordinate system to one one tangent plane at x_i. Hence we set $x_i = 0$ and $v_j = x_j - x_i$, where x_j's are assumed to be $x_j \in N(x_i)$. In this sense we see that v_j are vectors in tangent plane at x_i. Noting that x_i belongs to affine span of its neighbours and by the property of W we have

$$\begin{aligned}
0 &= x_i - \sum_j W_{ij} x_j \\
&= x_i \sum_j W_{ij} - \sum_j W_{ij} x_j \\
&= \sum_j W_{ij}(x_i - x_j) = \sum_j W_{ij} v_j
\end{aligned}$$

Assuming f is smooth we may use Taylor expansion to get the desired identity as follows:

$$\begin{aligned}
[(I-W)f]_i &= f(0) - \sum_j W_{ij} f(v_j) \\
&= f(0) - \sum_j W_{ij} \left(f(0) + v_j^T \nabla f + \frac{1}{2} v_j^T H v_j + o(\|v_j\|^2) \right) \\
&\approx \left(f(0) - \sum_j W_{ij} f(0) \right) - \sum_j W_{ij} v_j^T \nabla f - \frac{1}{2} \sum_j W_{ij} v_j^T H v_j \\
&= \frac{1}{2} \sum_j W_{ij} v_j^T H v_j \\
&= \sum_j W_{ij} (x_i - x_j)^T H (x_i - x_j)
\end{aligned}$$

Observe that if $\sqrt{W_{ij}} v_j$ form an orthonormal basis then

$$[(I-W)f]_i \approx \sum_j W_{ij} v_j^T H v_j = tr(H) = [Lf]_i \tag{3.3}$$

Of course the assumption $\sqrt{W_{ij}} v_j$ form an orthonormal basis is not always true, but when it is true we have equivalence between LLE and Laplacian Eigenmaps by applying 3.3.

$$(I-W)^T (I-W) f \approx \frac{1}{2} L^2 f$$

the eigenvectors of L^2 is same as that of L hence the LLE reduces to finding eigenvectors of graph Laplacian when the assumption is satisfied. Although equivalence is shown in this specific setting, the two methods differ in general.

3.3 Complexity of Dimensional Reduction Methods

Computational complexity plays important role in the discussion of manifold learning since if the method scales poorly with amount of data then the method is difficult to be used for large data making it less useful for real life applications.

Among the manifold learning methods introduced the spectral decomposition is the primary computational bottleneck. For local methods the eigenvalue problem is sparse and for global methods the eigenvalue problem to be solved is dense. Spectral decomposition for dense matrix is $O(n^3)$, this makes it difficult to apply the algorithm to large data sets.Hence global methods such as Isomap is only suited for small to medium amount of data which in practice means less than roughly 2000 data points on regular desktop computer. On the other hand local methods only require solving sparse eigenvalue problems where there are some algorithms that perform much better than $O(n^3)$. When the algorithm involves k-nearest neighbours, the sparse eigenvalue problems using standard method have complexity $O((d+k)n^2)$ where d is dimension which the data is being reduced.

The local methods LLE, Laplacian eigenmaps and their variants can scalable to large data sets. Global methods such as cMDS and Isomap on the other hand needs modification in order for it to scale up to large data due to its complexity. The algorithm Landmark Isomap introduced in [14] is designed to make Isomap scalable. Landmark Isomap (L-Isomap) designates $r < n$ data points called landmark points. Instead of finding shortest distance between each pair of points L-Isomap finds shortest distances from each data point to the landmark point. We can use Dijkstra's algorithm, the resulting time complexity is $O(krn\log(n))$. Given the shortest distances from all data points to landmark data point, we find Euclidean with the provided distances. Hence overall complexity of L-Isomap to $O(r^2 n)$. Although L-Isomap is fast compared to Isomap, the solution obtained through L-Isomap is only and approximate one compared to Isomap.

3.4 Supplementary Algorithms

There are many other manifold learning algorithms that are variants of the manifold learning methods discussed. We have already introduced L-Isomap which is a variant of Isomap. There is another version of Isomap called Conformal Isomap(c-Isomap) discovered in [13], this method replaces the assumption that the embedding is isometric embedding to conformal embedding. The method c-Isomap tries to capture the conformal geometry through the graph. Related work has been done in [24].

Conformal Isomap nonlinear dimension reduction from $\mathbb{R}^D$ to $\mathbb{R}^d$:

- **Step 1:** Given data set $\{x_i\}_{i=1}^n$ in $\mathbb{R}^D$, form the weighted k-nearest neighbour graph G. Assign each edge (x_i, x_j) weight $\frac{\|x_i - x_j\|}{\sqrt{M(x_i)M(x_j)}}$. Where $M(x_i) = \frac{1}{k}\sum_{x_j \in N(x_i)} \|x_i - x_j\|$ which is the mean distance between the neighbours.

- **Step 2:** Compute the shortest path distances between all pair of vertices with the edge weights described above and store it in S_{ij}. Use Dijkstra's algorithm with Fibonacci Heap. Store the square of the shortest path into distance matrix $D_{ij} = S_{ij}^2$

- **Step 3:** Apply cMDS algorithm with dissimilarity matrix D.

Due to application of cMDS i.e dense eigenvalue problem, the complexity of c-Isomap is same as that of Isomap which is $O(n^3)$. However we may approximate the exact solution using the landmark idea introduced for L-Isomap. There are other methods such as Hessian Locally Linear Embedding(HLLE) [17] and Local Tangent Space Alignment(LTSA) introduced in [40] . Similar to LLE the HLLE looks at locally linear patches and map them to lower dimensional space. The HLLE maps the tangent spaces estimates obtained by PCA on the neighbourhood of x_i to linear patches. The mapping is obtained by solving a sparse eigenvalue problem. The HLLE algorithm has an optimality guarantee as shown in [17]. Unlike Isomap the HLLE algorithm does not require the underlying parameter space to be convex, and embedding to be globally isometric. The method HLLE requires the parameter space to be connected and locally isometric to the manifold which is a weaker assumption than what Isomap requires. It has been shown in [17] that HLLE can outperform Isomap on some non-convex manifolds. Even though HLLE has some properties which shows its superiority to Isomap, due to its difficulty to implement the method is not as popular as Isomap.

There are other algorithms that came more recently such as Local Tangent Space Alignment(LTSA) introduced in [40] and Manifold Charting first appeared in [8]. These methods are variants of the methods we have already introduced but with different cost functions and parametrizations.

A close relative to Laplacian Eigenmaps is Diffusion Maps a detailed introduction given in [11][29]. The diffusion maps originates from ideas of dynamical systems by defining Markov chain on the graph of the data [28]. There are plethora of algorithms developed recently but their relations to other methods and usefulness to real life data remains elusive.

3.5 Common Issues Facing Manifold Learning

Although manifold learning methods have been used with some success to selected real life data sets, there are challenges to applying these methods to general data. One of the clear challenges is to identify the underlying dimensionality of the data living in the ambient space. The problem of intrinsic dimensionality can be resolved in PCA and Isomap through analysis of variance of dimension reduced data. However finding intrinsic dimensionality of the manifold when using methods such as LLE and Laplacian Eigenanmaps is difficult.

Recently there have been explosion of techniques in manifold learning , however most of them are experimented only on synthetic data sets which are carefully chosen. It is difficult to gauge their performances, since the idea of useful low dimensional representation varies from problem to problem. The correct performance measure for these algorithms are hard to define. The main evaluation method for most of these manifold learning algorithm is to run the algorithm on some artificial data set and observe whether the result is intuitively pleasing. This evaluation method is neither precise nor objective. Such lack of precise evaluation method could be one of the reasons why there are many similar algorithms present. Many of these new algorithms do not have theoretical guarantee, hence it is not clear that they are correct even in principle.

For methods with theoretical guarantee, not much is known for the convergence rates of manifold learning methods. We expect that the convergence rate to depend on the geometry of the underlying manifold.

In some problems one could choose distribution which to sample the data on the manifold. It is clear that for efficient representation of the manifold, we need to sample considering curvature into account. Such method of sampling on manifold leads to the idea of adaptive manifold learning methods. Although some preliminary work has been done in [41], there are still lots of room for improvement.

Many real life data suffer from having to deal with high noise but manifold learning methods such as Isomap and cMDS are extremely sensitive to noise making it difficult to use for realistic data. Even though Laplacian Eigenmaps and LLE are less sensitive to noise they are extremely parameter sensitive, hence choosing the right parameter is often challenging. Also given data points checking when the manifold assumption is reasonable is a difficult question yet to be fully answered.

In summary, for manifold learning methods to perform well and to be more useful for real life data the issues of performance measure, noise sensitivity and parameter selections must be addressed.

CHAPTER 4
Laplacian on Manifolds

To understand manifold learning methods well, one needs to understand basic ideas of geometric analysis. In this chapter we will discuss essential differential geometry to be used later on the for the proof of convergence of Laplacian Eigenmaps algorithm in chapter 6.

4.1 Riemannian Metric

We would like to generalize Laplacian on $\mathbb{R}^n$ to one on smooth compact Riemannian manifolds without boundary. Detailed introduction is given in [31]. Recall that the Euclidean Laplacian is

$$\Delta_{\mathbb{R}^n} = \sum_{i=1}^{n} \left(\frac{\partial}{\partial x^i} \right)^2$$

In order to generalize Euclidean Laplace operator to Riemannian manifolds, we need to define Riemannian metric. Riemannian metric gives notion of distance on the manifold. By giving distance we should we able to measure lengths of curves on the manifold $\mathcal{M}$. For instance given a surface $\mathcal{M} \subset \mathbb{R}^3$, we measure the length of the curve $\gamma : [0,1] \to \mathcal{M}$ as

$$l_\gamma = \int_0^1 |\gamma'(t)| dt$$

Observe that the key ingredient here used to measure the length is the modulus of the tangent vector $\gamma'(t) \in T_{\gamma(t)}\mathcal{M}$. In the same way we need to introduce structure on the tangent space in order to measure length, area and volume.

Definition 4.1 (Riemannian Manifold). *A Riemannian Manifold is a smooth manifold $\mathcal{M}$ with a family of smoothly varying positive definite inner products g_p on $T_p\mathcal{M}$ for each $p \in \mathcal{M}$. The family g is called Riemannian metric. Two Riemannian manifolds $(\mathcal{M}, g)$ and $(\mathcal{N}, h)$ are called isometric if there exists a smooth diffeomorphism $f : \mathcal{M} \to \mathcal{N}$ such that*

$$g_p(X, Y) = h_{f(x)}(f_*X, f_*Y)$$

for all $X, Y \in T_p\mathcal{M}$, for all $p \in \mathcal{M}$.

By above definition we may identify g_p as bilinear form on $T_p\mathcal{M}$ i.e an element of $T_p^*\mathcal{M} \otimes T_p^*\mathcal{M}$ hence g is a smooth section of $T^*\mathcal{M} \otimes T^*\mathcal{M}$. So similar to what we have demonstrated before, the length of a curve inside a manifold is

$$l(\gamma) = \int_0^1 \sqrt{g_{\gamma(t)}(\gamma'(t), \gamma'(t))}dt$$

It is useful to be able to compute the metric in local coordinates. Let $X, Y \in T_p\mathcal{M}$ and $(x^1, \cdots, x^n)$ are be local coordinate near $p \in \mathcal{M}$. From linearity there exists $\alpha^i, \beta^i \in \mathbb{R}$ such that

$$X = \sum_{i=1}^{n} \alpha^i \frac{\partial}{\partial x^i} \quad Y = \sum_{i=1}^{n} \beta^i \frac{\partial}{\partial x^i}$$

By bi-linearity of g_p we have

$$\begin{aligned} g_p(X, Y) &= g_p\left(\sum_{i=1}^{n} \alpha^i \frac{\partial}{\partial x^i}, \sum_{j=1}^{n} \beta^j \frac{\partial}{\partial x^i}\right) \\ &= \sum_{i,j} \alpha^i \beta^i g_p\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right) \end{aligned}$$

Hence g_p can be identified with symmetric, positive definite matrix $g_{ij}(p) = g_p\left(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j}\right)$. Thus we may write the metric as

$$g = \sum_{i,j} g_{ij} dx^i \otimes dx^j$$

note that this form makes it much more convenient to perform explicit computations in Riemannian Geometry.

As we can see being embedded in $\mathbb{R}^n$ is not necessary to define Riemannian manifolds. The question is then can any compact Riemannian manifold be isometrically embedded in $\mathbb{R}^n$. The answer has been given by Nash embedding theorem, which says it is always possible to isometrically embed Riemannian manifolds into Euclidean spaces as long as the dimension of the Euclidean space is high enough.

Theorem 4.2 (Nash)**.** *Any compact Riemannian C^k manifold with $3 \leq k \leq \infty$ of dimension d, has a C^k isometric embedding in $\mathbb{R}^D$ where $D = \frac{n(3n+11)}{2}$*

Manifold learning methods such as Isomap assumes that the underlying Riemannian manifold is isometrically embedded in space of observations which is Euclidean space. The theorem 4.2 gives validity to this assumption.

4.2 Sobelev Spaces on Manifold

Having the right notion for the space of functions is essential to analysis of differential equations. We will first define L^2 space on manifolds. This notion of L^2 space should be compatible with the usual notion of L^2 space on Euclidean space. We define volume form in the following way

Definition 4.3. *Volume form of a Riemannian manifold is the top dimensional form dV with local coordinates*

$$dV = \sqrt{\det g}\ dx^1 \wedge \cdots \wedge dx^n$$

where $(\frac{\partial}{\partial x^1}, \cdots, \frac{\partial}{\partial x^n})$ is a positively oriented basis of $T_p\mathcal{M}$. We set the volume of (M, g) to be

$$Vol(\mathcal{M}) = \int_{\mathcal{M}} dV$$

Above allows us to integrate any function $f \in C^\infty(\mathcal{M})$, hence we can now define the Hilbert space L^2 as follows:

Definition 4.4. *Given Riemannian manifold $(\mathcal{M}, g)$, we define Hilbert space $L^2(\mathcal{M})$ as completion of $C_c^\infty(\mathcal{M})$ with respect to the inner product*

$$\langle f(x), g(x) \rangle = \int_{\mathcal{M}} f(x)g(x)dV$$

One can verify the above definition is indeed a Hilbert space. It is also possible to define L^2 space of k-forms, however that is not useful in our discussion since it will not come up in further discussions. We may in a similar way define $L^p(\mathcal{M})$ as completion of $C_c^\infty(\mathcal{M})$ with respect to norm $\|f\|_p = \int_{\mathcal{M}} |f|^p dV$. We now extend the definition of Sobolev space to Riemannian manifolds.

Definition 4.5. *Let $\{x_i\}_{i=1}^N \subset \mathcal{A}$ such that $\mathcal{M} = \bigcup_{i=1}^N U_i$ and let $\{\phi_i\}_{i=1}^N$ be partition of unity subordinate to cover $\{U_i\}_{i=1}^N$. We now define $f \in W^{k,p}(\mathcal{M})$*

$$\|f\|_{W^{k,p}(\mathcal{M})} = \sum_{i=1}^N \|(\phi_i f) \circ x_i^{-1}\|_{W^{k,p}}$$

Indeed this definition is well posed in the sense that it only depends on the metric itself. Similar to Euclidean space we denote $H^k(\mathcal{M}) := W^{k,2}(\mathcal{M})$. Note that $H^k(\mathcal{M})$ is a Hilbert space. Most results that are true about Sobolev space on $\mathbb{R}^n$ are also true for compact Riemannian manifolds.

4.3 Levi-Civita Connection and Curvature

In order to compare tangent spaces at different points we need to define a notion of affine connection.

Definition 4.6. *Let $(\mathcal{M}, g)$ be a smooth manifold and let $\Gamma^\infty(T\mathcal{M})$ be space of smooth section on the tangent bundle $T\mathcal{M}$. Then an affine connection is a* *bilinear map*

$$\nabla : \Gamma^\infty(T\mathcal{M}) \times \Gamma^\infty(T\mathcal{M}) \longrightarrow \Gamma^\infty(T\mathcal{M})$$
$$(X, Y) \mapsto \nabla_X Y$$

such that for all f in $C^\infty(\mathcal{M})$ and $X, Y \in \Gamma^\infty(T\mathcal{M})$

- $\nabla_{fX} Y = f \nabla_X Y$
- $\nabla_X (fY) = df(X)Y + f\nabla_X Y$

As defined there are infinitely many affine connections on a manifold. However on every Riemannian manifold $(\mathcal{M}, g)$ there is a unique affine connection ∇ called Levi-Civita connection such that:

- $\nabla_X Y - \nabla_Y X = [X, Y]$ for all $X, Y \in \Gamma^\infty(\mathcal{M})$
- Parallel transport is an isometry, meaning that the inner products defined using g between tangent vectors are preserved.

From now on if otherwise stated all connections will be Levi-Civita connections. We will now define Riemann curvature tensor which is the most common method for describing curvature of Riemannian manifold.

Definition 4.7. *The curvature R of a Riemannian manifold $(\mathcal{M}, g)$ is a correspondence that associates to every pair $X, Y \in \Gamma^\infty(\mathcal{M})$ a mapping $R(X, Y) : \Gamma^\infty(T\mathcal{M}) \to \Gamma^\infty(T\mathcal{M})$ given by*

$$R(X, Y)Z = \nabla_Y \nabla_X Z - \nabla_X \nabla_Y Z + \nabla_{[X,Y]} Z \qquad Z \in \Gamma^\infty(T\mathcal{M}) \tag{4.1}$$

One can write the curvature tensor as

$$R^s_{ijk} = dx^s R(\partial_i, \partial_j)\partial_k = dx^s(\nabla_{\partial_j}\nabla_{\partial_i} - \nabla_{\partial_i}\nabla_{\partial_j})\partial_k$$

Note that R is a measure of noncommutativity of covariant derivative. Often people write it in the form $R_{ijks} = g_{i\rho} R^{\rho}_{jks}$.

We have the following identities, the first of which is known as *Bianchi identity*:

$$R_{ijks} + R_{jkis} + R_{kijs} = 0$$

$$R_{ijks} = -R_{jiks}$$

$$R_{ijks} = -R_{ijsk}$$

$$R_{ijks} = R_{ksij}$$

Explicit calculation of Riemann curvature tensor is often difficult, hence it is common to use *Christoffel symbol.* Which is defined as $\nabla_{\partial_i} \partial_j := \Gamma^k_{ij} \partial_k$ in local coordinates

$$\Gamma^k_{ij} = \frac{1}{2} g^{k\mu} \left(\frac{\partial g_{\mu i}}{\partial x^j} + \frac{\partial g_{\mu j}}{\partial x^i} - \frac{\partial g_{ij}}{\partial x^\mu} \right)$$

Hence we may write the Riemann curvature tensor in terms of the Christoffel symbol

$$R^i_{j\mu\nu} = \partial_\mu \Gamma^i_{\nu j} - \partial_\nu \Gamma^i_{\mu j} + \Gamma^i_{\mu\lambda} \Gamma^\lambda_{\nu j} - \Gamma^i_{\nu\lambda} \Gamma^\lambda_{\mu j}$$

It is useful to consider traces of the Riemann curvature tensor. *Ricci curvature* tensor is defined as trace $R_{ij} := R^{\rho}_{i\rho j}$. Another useful geometric quantity is *scalar curvature* which is defined as trace of Ricci curvature tensor with respect to the metric $R := g^{ij} R_{ij}$.

4.4 Laplacian on Functions

Much of the thesis uses properties of Laplace operator. One might wonder why there is a big emphasize on Laplacian and not some other differential operator. This is because Laplacian commutes with isometries of the manifold.

For instance on $\mathbb{R}^n$ the isometries are translation and rotation, so given an isometry T the operator satisfies $\Delta(\phi \circ T) = (\Delta\phi) \circ T$. In fact much more is true if L is any operator that commutes with translations and rotations then there exists $a_i \in \mathbb{R}$ such that $L = \sum_{j=1}^{N} a_j \Delta^j$. Hence Laplacian is the building block of operators that are invariant under isometries.

Given a Riemannian manifold $(\mathcal{M}, g)$ we can define analogues of divergence and gradient.

Let $\mathcal{M}$ be a Riemannian manifold(with or without boundary) of dimension greater than two with metric g. The metric induces the analogue of **div** and **grad** on manifold.

Definition 4.8. *Let $f \in C^1(\mathcal{M})$, then* $\mathbf{grad}f$ *is the vector field satisfying*

$$\langle \mathbf{grad}f, X \rangle_g = Xf \quad \forall X \in \Gamma^\infty(TM)$$

In a similar way we can define divergence on manifold.

Definition 4.9. *Let X be differentiable vector field on $\mathcal{M}$, define the real valued function* $\mathbf{div}X$ *by*

$$\mathbf{div}X(p) = Tr(\xi \to \nabla_\xi X)$$

where ξ ranges over $T_p\mathcal{M}$

Just as in on Euclidean space divergence on manifold satisfies following properties.

$$\begin{aligned} \mathbf{div}(X+Y) &= \mathbf{div}X + \mathbf{div}Y \\ \mathbf{div}(fX) &= f\mathbf{div}X + Xf \end{aligned}$$

Now we can define Laplacian such that it coincide with our intuition on Euclidean space.

Definition 4.10. *Let $f \in C^2(\mathcal{M})$, we define Laplacian of f,Δf by*

$$\Delta f = \mathbf{div}(\mathbf{grad} f)$$

In local coordinates we may write the Laplacian as

$$\Delta f = \frac{1}{\sqrt{\det g}} \frac{\partial}{\partial x^j}(g^{ij}\sqrt{\det g}\frac{\partial}{\partial x^i} f)$$

Where $g^{ij} = (g_{ij})^{-1}$. Observe that in Euclidean case where $g_{ij} = \delta_{ij}$ then one retains the usual Laplacian from the coordinate definition of Laplacian on functions since $\det(g_{ij}) = 1$ in this case.

$$\Delta_{\mathbb{R}^n} = \sum_{i=1}^{n} (\frac{\partial}{\partial x^i})^2$$

The following are important properties of the Laplace-Beltrami operators.

Theorem 4.11. *Let $\mathcal{M}$ be a closed compact manifold, and consider Laplace eigenvalue problem. Then:*

- *The set of eigenvalue consists of infinite sequence $0 < \lambda_1 \leq \lambda_2 \leq ... \rightarrow \infty$*
- *Each eigenvalue has finite multiplicity and the eigenspaces corresponding to distinct eigenvalues are $L^2(\mathcal{M})$ orthogonal*
- *Each eigenfunction is smooth analytic.*

Theorem 4.12. *The Laplace operator depends only on the given Riemannian metric. If*

$$F : (\mathcal{M}, g) \rightarrow (\mathcal{N}, h)$$

is an isometry, then $(\mathcal{M}, g)$ and $(\mathcal{N}, h)$ have the same spectrum. Also if ϕ is an eigenfunction on $(\mathcal{N}, h)$, then $\phi \circ F$ is eigenfunction on $(\mathcal{M}, g)$

As we have seen in chapter 2 finding geodesics play an important role in manifold learning.

The geodesics on Riemannian manifold are given by the following Euler-Lagrange equation

$$\frac{d^2\gamma^i}{dt^2} + \Gamma^i_{jk}\frac{d\gamma^j}{dt}\frac{d\gamma^k}{dt} = 0 \qquad i = 1, \cdots, n$$

where $\gamma^i = x^i \circ \gamma$. By the local flatness there is a neighbourhood of x that is diffeomorphic to $\mathbb{R}^n$, the existence and uniqueness theory of ODEs guarantees that for any $x \in \mathcal{M}$ and $v \in T_x\mathcal{M}$, there exists an $\epsilon > 0$ and a unique geodesic $\gamma_v(t), t \in (-\epsilon, \epsilon)$, with $\gamma(0) = x, \gamma'(0) = v$.

Definition 4.13. *Let $\mathcal{M}$ be compact. For $x \in \mathcal{M}$ the exponential map $\exp_x : T_x\mathcal{M} \to \mathcal{M}$ is defined by*

$$\exp_x(v) = \gamma_v(1)$$

Many objects in Riemannian geometry are expressed much more conveniently in geodesic normal coordinate.

Theorem 4.14. *For $\epsilon > 0$ small enough, $\exp_x$ restricted to $B_\epsilon(0) \subset T_x\mathcal{M}$ is a diffeomorphism onto its image.*

Corollary 4.15. *For any $p \in \mathcal{M}$, then there exists an $\epsilon > 0$ so that any $q \in \mathcal{M}$ with $d_g(p, q) < \epsilon$ such that there is a unique geodesic connecting p to q whose length is less than ϵ.*

The exponential map takes a given tangent vector on the manifold, runs along the geodesic starting at that point and going in that direction. In order to construct the coordinate chart we use the natural isomorphism

$$E : \mathbb{R}^n \to T_p\mathcal{M}$$

Where the natural isomorphism to $\mathbb{R}^n$ is given by the mapping between the basis vectors. Finally the *geodesic normal coordinates* are constructed as follows

$$\phi := E^{-1} \circ \exp_p^{-1} : U \to \mathbb{R}^n$$

The coordinates have the following properties:

- The coordinates of p are $(0, \cdots, 0)$.
- In this coordinate the Riemannian metric g is equal to δ_{ij}
- The Christoffel symbols vanish at p, also $\partial_k g_{ij} = 0$ for all k.

These properties allow simpler formulation of many differential operators on manifold making it easier to use. The normal coordinate chart can be used to give information about the metric. For instance knowing the normal coordinate chart and the curvature tensor in a neighbourhood of a point allows one to find the metric tensor in that neighbourhood.

CHAPTER 5
Heat Operator on Manifold

We have seen that in chapter 2 that heat equation was used to derive weights for Laplacian Eigenmaps.

Given compact Riemannian manifold without boundary the operator $\partial_t + \Delta_g$ is called the heat operator. Heat operator contains crucial informations about the geometry of manifold. For instance one could deduce information about curvature and topology from the solutions of the heat equation [31]. In this chapter we will investigate important properties of the heat operator which we will use later on for the proof of convergence of Laplacian Eigenmaps.

The heat operator acts on $C(\mathcal{M} \times \mathbb{R}_+)$ which are C^2 in space and C^1 on time. The homogeneous heat equation is written as

$$\begin{cases} \partial_t u(x,t) + \Delta_g u(x,t) = 0 & (x,t) \in \mathcal{M} \times \mathbb{R}_+ \\ u(x,0) = f(x) & x \in \mathcal{M} \end{cases} \tag{5.1}$$

where $f(x) \in L_2(\mathcal{M})$ which should be thought of as the initial heat distribution on the manifold. The solution $u(x,t)$ has non-increasing L_2 norm in time. This can easily be verified observing that the time derivative of $\frac{\partial}{\partial t}\|u(x,t)\|_2 \leq 0$. This information can be used to prove the uniqueness of solutions to heat operator.

In order to analyze the solutions in more generality one uses the notion of *fundamental solution* to the heat equation which is also called the *heat kernel*.

Heat kernel is $p(t,x,y) \in C^\infty(\mathbb{R}_+ \times \mathcal{M} \times \mathcal{M})$ such that

$$\begin{cases} (\partial_t + \Delta_x)p(t,x,y) = 0 \\ \lim_{t\to 0} \int_{\mathcal{M}} p(t,x,y)f(y)dV_y = f(x) \end{cases} \tag{5.2}$$

where Δ_x denotes the Laplacian acting in the x variable. The existence of such function $p(t,x,y)$ needs proof, the details are provided in [31]. The reason why $p(t,x,y)$ is called the fundamental solution is perhaps given any initial heat distribution $f(x) \in L_2(\mathcal{M})$ one could construct the solution to Equation 5.1 by setting $u(x,t) = \int_{\mathcal{M}} p(t,x,y)f(y)dV_y$. Indeed one can check

$$\begin{aligned} \partial_t u(x,t) &= \int_{\mathcal{M}} \partial_t p(t,x,y)f(y)dV_y \\ &= -\int_{\mathcal{M}} \Delta_x p(t,x,y)f(y)dV_y \\ &= -\Delta_x \int_M p(t,x,y)f(y)dV_y \\ &= -\Delta_x u(x,t) \end{aligned}$$

Furthermore $\lim_{t\to 0} u(x,t) = \lim_{t\to 0} \int_{\mathcal{M}} p(t,x,y)f(y)dV_y = f(x)$ thus $u(x,t)$ also satisfies the initial condition.

Theorem 5.1 (Sturm-Liouville decomposition)**.** *For $\mathcal{M}$ compact Riemannian manifold without boundary, there exists orthonormal basis $\{\phi_0, \phi_1, \cdots\}$ of $L^2(\mathcal{M})$ consisting of eigenfunctions of Δ_g where ϕ_j having eigenvalue λ_i satisfying*

$$\lambda_0 \leq \lambda_1 \leq \cdots \to \infty$$

the heat kernel is given by

$$p(x,y,t) = \sum_{j=0}^{\infty} e^{-\lambda_j t}\phi_j(x)\phi_j(y)$$

Note that the decomposition holds in the sense of point-wise convergence. To illustrate an example of how the heat kernel contains information about the geometry of the manifold, one can study the trace of heat kernel $Tr(e^{-\Delta t}) = \int_{\mathcal{M}} p(t,x,x)dV = \sum_{j=0}^{\infty} e^{-\lambda_j t}\|\phi_j\|_2 = \sum_{j=0}^{\infty} e^{-\lambda_j t}$. There is an asymptotic formula for the trace given in [26]

$$\sum_{j=0}^{\infty} e^{-\lambda_j t} = \frac{1}{(4\pi t)^{\frac{n}{2}}}\left(Vol(M) + \frac{t}{6}\int_{\mathcal{M}} R_g dV + O(t^2)\right)$$

this shows that the heat kernel contains information about the volume and curvature of the manifold.

We define heat propagator as follows

Definition 5.2 (Heat Propagator)**.** *Given $t > 0$, the heat propagator $e^{-t\Delta_g} : L^2(\mathcal{M}) \to L^2(\mathcal{M})$ is defined as*

$$e^{-t\Delta_g} f(x) = \int_{\mathcal{M}} p(t,x,y)f(y)dV_y$$

The heat propagator has the following properties:

Proposition 5.3. *The heat operator satisfies the following properties*

- $e^{-t\Delta_g} \circ e^{-s\Delta_g} = e^{-(t+s)}\Delta_g$
- $e^{-t\Delta_g}$ *is self-adjoint and positive*
- $e^{-t\Delta_g}$ *is compact operator*

Heat kernel gives rich information about the geometry of the manifold. As we have seen the heat kernel also satisfies some nice properties.

In $\mathbb{R}^n$ the heat kernel is given by $p(t,x,y) = \frac{1}{(4\pi t)^{\frac{n}{2}}} e^{\|x-y\|^2/4t}$ however on general Riemannian manifolds such explicit formula does not exist. Hence one relies on asymptotic expansions of heat kernel.

Theorem 5.4 ([31]). *There exists $\epsilon > 0$ such that for all $x, y \in M$ with $d_g(x, y) < \epsilon$*

$$p(x, y, t) = \frac{e^{-d_g^2(x,y)/4t}}{(4\pi t)^{\frac{n}{2}}} \left(\sum_{i=0}^{N} t^i u_i(x, y) + O(t^{N+1}) \right)$$

Where

$$u_i(x, y) = -d_g^{-i}(x, y) \det(y)^{-\frac{1}{2}} \int_0^r \det(x(s))^{\frac{1}{2}} \Delta_y u_{i-1}(x(s), y) s^{i-1} ds$$

denote $x(s)$ as the geodesic from x to $y = x(d(x, y))$. Note

$$u_0(x, y) = \frac{1}{\sqrt{\det(g_y)}}$$

Above theorem implies that for small enough time and small enough neighbourhood the heat kernel on manifold is approximated by the heat kernel on Euclidean space. In next chapter we will use techniques mentioned from this chapter for the proof of convergence Laplacian Eigenmaps.

CHAPTER 6
Convergence of Laplacian EigenMaps

In this chapter we aim to establish convergence of eigenvectors of graph Laplacian as mentioned in chapter 2 associated to point cloud data set to eigenfunctions of Laplace-Beltrami operator when the data set is sampled from a uniform probability distribution on the embedded manifold. The result presented in this chapter are due to [3]. Generalization of this result to arbitrary probability distribution is given by Lafon in [25]. Similar attempts of showing convergence of discrete Laplace spectrum to continuous one is given in [4][15]. In what follows $\mathcal{M}$ is a d dimensional submanifold of $\mathbb{R}^D$ with the corresponding induced volume form dV. The data points denoted $\{x_i\}_{i=1}^n \in \mathcal{M} \subset \mathbb{R}^D$, the Laplacian Eigenmaps algorithm is to uncover eigenfunctions and eigenvalues of Laplace-Beltrami operator on $\mathcal{M}$ from the data set.

Recall from chapter 2 for small time $t > 0$ the discrete Laplacian approximates the Laplace-Beltrami operator

$$\Delta_{\mathcal{M}} f(x_i) \approx \frac{1}{t} \frac{1}{(4\pi t)^{\frac{d}{2}}} \frac{1}{|V|} \left[f(x_i) \sum_{x_j \in N_\epsilon(x_i)} e^{-\frac{\|x_i - x_j\|^2}{4t}} - \sum_{x_j \in N_\epsilon(x_i)} e^{-\frac{\|x_i - x_j\|^2}{4t}} f(x_j) \right]$$

We may rewrite above using discrete Laplace operator $L = D - W$ from Laplacian Eigenmaps algorithm

$$\Delta_{\mathcal{M}} f(x_i) \approx \frac{1}{t} \frac{1}{(4\pi t)^{\frac{d}{2}}} \frac{1}{|V|} [Lf]_i$$

Note that the we derived the above approximation through heat equation $\Delta_{\mathcal{M}}u(x,t) = -\frac{\partial}{\partial t}u(x,t)$ with initial condition $u(x,0) = f$. Hence the Laplace-Beltrami operator satisfies

$$\begin{aligned}
\Delta_{\mathcal{M}}f &= -\lim_{t\to 0}\frac{1}{t}(u(x,t) - f(x)) && (6.1)\\
&= \lim_{t\to 0}\frac{1}{t}\left(f(x) - \int_{\mathcal{M}} H_t(x,y)f(y)dV_y\right) && (6.2)\\
&= \lim_{t\to 0}\left(\frac{1-e^{-\Delta_{\mathcal{M}}}}{t}\right)f && (6.3)
\end{aligned}$$

We know that from asymptotic expansion of heat kernel

$$\left(\frac{1-e^{-\Delta_{\mathcal{M}}}}{t}\right)f \approx \frac{1}{t}\frac{1}{(4\pi t)^{\frac{d}{2}}}\left[f(x)\int_{\mathcal{M}} e^{-\frac{\|x-y\|^2}{4t}}dV_y - \int_{\mathcal{M}} e^{-\frac{\|x-y\|^2}{4t}}f(y)dV_y\right]$$

Observe the empirical version of the right hand side is the discrete Laplace operator as we have seen in chapter 2. We will extend the definition of Laplace operator to include points outside the sample points. We first define the operator $\mathcal{L}_t : L^2(\mathcal{M}) \to L^2(\mathcal{M})$ as shown below

$$\mathcal{L}_t(f)(p) = \frac{1}{t(4\pi t)^{d/2}}\left(\int_{\mathcal{M}} e^{-\frac{\|p-y\|^2}{4t}}f(p)dV_y - \int_{\mathcal{M}} e^{-\frac{\|p-y\|^2}{4t}}f(y)dV_y\right)$$

the empirical version of above operator is

$$L_{t,n}(f)(p) = \frac{1}{t(4\pi t)^{d/2}}\frac{1}{|V|}\left(\sum_{i=1}^{n} e^{-\frac{\|p-x_i\|^2}{4t}}f(p) - \sum_{i=1}^{n} e^{-\frac{\|p-x_i\|^2}{4t}}f(x_i)\right)$$

The empirical operator $L_{t,n}$ is a discrete Laplace operator on a point cloud. To see this consider the graph whose vertices $V = \{x_1, ...x_n\}$ and edge weight matrix $W_{ij} = \frac{1}{n}\frac{1}{t(4\pi 4)^{d/2}}e^{-\frac{\|x_i-x_j\|^2}{4t}}$. Note that for any $f : \mathcal{M} \to \mathbb{R}$ if one considers the restriction of f to the vertex set V which we denote f_V, then $L_{t,n}(f)|_V = L_{t,n}(f_V)$. One can see that $L_{t,n}f|_V = (D-W)f_V$, where D is a diagonal matrix with entries $D_{ii} = \sum_{j=1}^{n} W_{ij}$ which is a sum of all entries in the i-th row of W. Hence the eigenvectors of graph Laplacian is same as eigenfunctions $L_{t,n}$ when restricted to the point cloud.

The main result of this chapter is due to [3].

Theorem 6.1. *Let $\lambda^t_{n,i}$ be the i-th eigenvalue of $L_{t,n}$ and $\phi^t_{n,i}(x)$ be the corresponding eigenfunction. Let λ_i and $\phi_i(x)$ be the corresponding eigenvalue and eigenfunction of $\Delta_{\mathcal{M}}$ respectively. Then there exists a sequence $t_n \to 0$ such that*

$$\lim_{n\to\infty} \lambda^{t_n}_{n,i} = \lambda_i$$

$$\lim_{n\to\infty} \|\phi^{t_n}_{n,i}(x) - \phi_i(x)\|_2 = 0$$

where the limits are taken in probability.

The proof requires two steps first showing convergence of eigenfunctions and eigenvalues of $L_{t,n}$ to that of $\mathcal{L}_t$ and then convergence of $\mathcal{L}_t$ to $\Delta_{\mathcal{M}}$. Hence 6.1 can be divided into two separate theorems.

Theorem 6.2. *Let $\lambda_i, \lambda^t_i, \phi_i, \phi^t_i$ be the i-th smallest eigenvalues and the corresponding eigenfunctions of Δ and $\mathcal{L}_t$ respectively. Then*

$$\lim_{t\to 0} |\lambda_i - \lambda^t_i| = 0$$
$$\lim_{t\to 0} \|\phi_i - \phi^t_i\| = 0$$

Theorem 6.3. *Let $\lambda^t_{n,i}$ and λ^t_i be the i-th eigenvalue of $L_{t,n}$ and $\mathcal{L}_t$ respectfully. Let $\phi^t_{n,i}$ and ϕ^t_i be the corresponding eigenfunctions. Then there exist $t > 0$ small enough such that*

$$\lim_{n\to\infty} \lambda^t_{n,i} = \lambda^t_i$$
$$\lim_{n\to\infty} \|\phi^t_{n,i} - \phi^t_i\| = 0$$

whenever $\lambda^t_i < \frac{1}{2t}$ where the convergence is almost sure.

Between the two theorems, theorem 6.2 is more difficult. To demonstrate the convergence we use different functional approximation $\frac{1-e^{-t\Delta_{\mathcal{M}}}}{t}$. Although $\frac{1-e^{-t\Delta_{\mathcal{M}}}}{t}$ does not converge uniformly to $\Delta_{\mathcal{M}}$ they share eigenfunctions and eigenvalues in the limit of small time. To show operators $\frac{1-e^{-t\Delta_{\mathcal{M}}}}{t}$ and $\mathcal{L}_t$ have approximately the same spectrum we use the following proposition.

Proposition 6.4. *Let A, B be positive, self-adjoint operators in $L^2(\mathcal{M})$. Let $R = A - B$, and $\lambda_1(A) \le \lambda_2(A) \le \cdots$ and $\lambda_1(B) \le \lambda_2(B) \le \cdots$ denote the eigenvalues of A and B respectively. Assume there exist $\epsilon > 0$ such that for all $f \in L^2(\mathcal{M})$ following holds*

$$\frac{|\langle Rf, f\rangle|}{|\langle Af, f\rangle|} \le \epsilon \tag{6.4}$$

Then for all k, we have $1 - \epsilon \le \frac{\lambda_k(B)}{\lambda_k(A)} \le 1 + \epsilon$

Proof. For any $f \in L^2$ using 6.4 we have

$$\begin{aligned}|\langle Af, f\rangle| &= |\langle (A - B + B)f, f\rangle| \\ &\le |\langle Rf, f\rangle + \langle Bf, f\rangle| \\ &\le |\langle Bf, f\rangle| + |\langle Rf, f\rangle| \le |\langle Bf, f\rangle| + \epsilon|\langle Af, f\rangle|\end{aligned}$$

Similarly,

$$|\langle Af, f\rangle| \ge |\langle Bf, f\rangle| - |\langle Rf, f\rangle| \ge |\langle Bf, f\rangle| - \epsilon|\langle Af, f\rangle|$$

Together they imply

$$(1 - \epsilon)|\langle Af, f\rangle| \le |\langle Bf, f\rangle| \le (1 + \epsilon)|\langle Af, f\rangle|$$

Now let H be arbitrary k-dimensional subspace of $L^2(\mathcal{M})$. Then we have

$$(1 - \epsilon) \max_H \min_{f \in H^\perp} |\langle Af, f\rangle| \le \max_H \min_{f \in H^\perp} |\langle Bf, f\rangle| \le (1 + \epsilon) \max_H \min_{f \in H^\perp} |\langle Af, f\rangle|$$

where $H^{\perp}$ is the orthogonal complement of H . Using the Courant-Fischer min-max theorem we get

$$(1-\epsilon)\lambda_k(A) \leq \lambda_k(B) \leq (1+\epsilon)\lambda_k(A)$$

the desired result follows. □

Hence if we can show that $R_t = \frac{1-e^{-t\Delta_{\mathcal{M}}}}{t} - \mathcal{L}_t$ satisfies

$$\frac{|\langle R_t f, f\rangle|}{|\langle \frac{1-e^{-t\Delta_{\mathcal{M}}}}{t} f, f\rangle|} \leq \epsilon \quad \text{for any} \quad f \in L^2(\mathcal{M}) \tag{6.5}$$

for some small enough t then by proposition 6.4 we would be able to deduce that

$$1-\epsilon \leq \frac{\lambda_k(\mathcal{L}_t)}{\lambda_k(\frac{1-e^{-t\Delta_{\mathcal{M}}}}{t})} \leq 1+\epsilon$$

note that eigenvalue

$$\lambda_k\left(\frac{1-e^{-t\Delta_{\mathcal{M}}}}{t}\right) = \frac{1-e^{-t\lambda_k}}{t} = \lambda_k + O(t)$$

where λ_k is the eigenvalue of the Laplace-Beltrami operator. This shows that

$$\lim_{t\to 0}\left|\lambda_k\left(\frac{1-e^{-t\Delta_{\mathcal{M}}}}{t}\right) - \lambda_k(\mathcal{L}_t)\right| = \lim_{t\to 0}|\lambda_k - \lambda_k^t| = 0$$

where the k-th eigenvalue $\lambda_k(\mathcal{L}_t)$ is denoted as λ_k^t. Also from proposition 6.4 we have convergence of eigenfunctions. Hence it is enough to show equation (6.5) in order to prove theorem 6.2, showing the result requires two estimates on R_t which we call remainder estimates.

Proposition 6.5. *Let $f \in L^2(\mathcal{M})$, there exists $C \in \mathbb{R}$ such that for all sufficiently small values of* t > 0 *the following holds:*

$$\|R_t f\| \leq C\|f\|$$

Proposition 6.6. *Let $f \in H^{\frac{d}{2}+1}(\mathcal{M})$, there exists $C \in \mathbb{R}$ such for small enough values of $t > 0$ following holds:*

$$\|R_t f\| \leq C\sqrt{t}\|f\|_{H^{\frac{d}{2}+1}}$$

Theorem 6.7. *For $t > 0$ sufficiently small, there exists a constant $C > 0$ that is independent of t such that the following is satisfied*

$$\sup_{f \in L^2} \frac{|\langle R_t f, f\rangle|}{|\langle \frac{1-e^{-t\Delta}}{t} f, f\rangle|} \leq Ct^{\frac{2}{k+6}}$$

furthermore,

$$\lim_{t\to 0} \sup_{f \in L^2} \frac{|\langle R_t f, f\rangle|}{|\langle \frac{1-e^{-t\Delta}}{t} f, f\rangle|} = 0$$

and hence R_t is dominated by $\frac{1-e^{-t\Delta_{\mathcal{M}}}}{t}$.

Proof. Recall from previous discussions

$$\frac{1-e^{-t\Delta_{\mathcal{M}}}}{t}\phi_i = \frac{1-e^{-\lambda_i t}}{t}\phi_i$$

We would like to have a lower bound on the eigenvalue of above operator in terms of t and λ_k. Hence consider a function $\varphi(x) = \frac{1-e^{-xt}}{t}$ for $x \geq 0$ where t is fixed. Note that φ is concave monotone increasing positive function of x. Let $x_0 = \frac{1}{\sqrt{t}}$. Then $\varphi(x_0) = \frac{1-e^{-\sqrt{t}}}{t}$ which implies $\frac{\varphi(x_0)}{x_0} = \frac{1-e^{-\sqrt{t}}}{\sqrt{t}}$. We can then split the real line in two intervals $\mathbb{R} = [0, x_0] \cup [x_0, \infty)$ and using concavity and monotonicity of φ we obtain

$$\varphi(x) \geq \min(\frac{1-e^{-\sqrt{t}}}{\sqrt{t}}x, \frac{1-e^{-\sqrt{t}}}{t})$$

Hence for sufficiently small $0 < t < 1/10$

$$\varphi(x) \geq \frac{1}{2}\min\left(x, \frac{1}{\sqrt{t}}\right) \tag{6.6}$$

Let $\phi_i(x)$ be the i-th eigenfunction of $\Delta_{\mathcal{M}}$ and let λ_i be the corresponding eigenvalue. By spectral theorem that the eigenfunctions form orthonormal basis of $L^2(\mathcal{M})$. Hence any $f \in L^2(\mathcal{M})$ can be written in terms of the eigenfunctions as

$$f(x) = \sum_{k=1}^{\infty} a_k \phi_k(x)$$

where $a_k \in \mathbb{R}$ are such that $\sum_{k=1}^{\infty} a_k^2 < \infty$. We may assume wlog $\|f\| = 1$ and f is orthogonal to the constant functions. Thus we get the following

$$\left\langle \frac{1 - e^{-t\Delta_{\mathcal{M}}}}{t} \phi_i, \phi_i \right\rangle = \frac{1 - e^{-\lambda_i t}}{t} = \varphi(\lambda_i) \geq \frac{1}{2} \min\left(\lambda_i, \frac{1}{\sqrt{t}}\right) \tag{6.7}$$

For $\alpha > 0$, we can split f as a sum of f_1 and f_2 as follows

$$f_1 = \sum_{\lambda_k \leq \alpha} a_k \phi_k \quad f_2 = \sum_{\lambda_k > \alpha} a_k \phi_k$$

We have then $\|f\|^2 = \|f_1\|^2 + \|f_2\|^2$. Observe the following

$$\begin{aligned} \left\langle \frac{1 - e^{-t\Delta}}{t} f, f \right\rangle &= \left\langle \sum_{k=1}^{\infty} \frac{1 - e^{-t\lambda_k}}{t} a_k \phi_k, \sum_{k=1}^{\infty} a_k \phi_k \right\rangle \\ &= \sum_{k=1}^{\infty} a_k^2 \left(\frac{1 - e^{-t\lambda_k}}{t} \right) \end{aligned}$$

By the inequality 6.7 we see

$$\sum_{k=1}^{\infty} a_k^2 \left(\frac{1 - e^{-t\lambda_k}}{t} \right) \geq \frac{1}{2} \sum_{k=1}^{\infty} a_k^2 \min\left(\lambda_k, \frac{1}{\sqrt{t}}\right) \geq \frac{1}{2} \sum_{k=1}^{\infty} a_k^2 \min\left(\lambda_1, \frac{1}{\sqrt{t}}\right)$$

Hence if we choose $0 < t < \min(\frac{1}{\lambda_1^2}, \frac{1}{10})$ and noting by assumption $\|f\| = 1$ i.e $\sum_{k=1}^{\infty} a_k^2 = 1$ we get

$$\left\langle \frac{1 - e^{-t\Delta}}{t} f, f \right\rangle \geq \frac{1}{2} \lambda_1 \sum_{k=1}^{\infty} a_k^2 = \frac{\lambda_1}{2} \tag{6.8}$$

Since we have obtained a lower bound on the denominator, we need upper bound on the numerator of the quotient. Using self adjointness of R_t we analyse $R_t f_1$ and $R_t f_2$ separately

$$\begin{aligned} |\langle R_t f, f\rangle| &= |\langle R_t f_1, f_1\rangle + 2\langle R_t f_1, f_2\rangle + \langle R_t f_2, f_2\rangle| \qquad (6.9)\\ &\leq \|R_t f_1\|\|f_1\| + 2\|R_t f_1\|\|f_2\| + \|R_t f_2\|\|f_2\| \qquad (6.10)\\ &\leq \|R_t f_1\|(\|f_1\| + 2\|f_2\|) + \|R_t f_2\|\|f_2\| \qquad (6.11)\\ &\leq 3\|R_t f_1\| + \|R_t f_2\|\|f_2\| \qquad (6.12)\end{aligned}$$

Where we used Cauchy-Schwarz and triangle inequalities for above. By proposition 6.5 we have a following bound

$$\|R_t f_1\| < C\sqrt{t}\|f_1\|_{H^{\frac{d}{2}+1}}$$

Recall from basic property of eigenfunctions of Laplacian $\|\phi_i\|_{H^{\frac{d}{2}+1}} \leq \Omega\lambda_i^{\frac{k+2}{4}}$ where Ω is some universal constant. Since f_1 is a band limited by α , we get

$$|\ R_t f_1\| \leq C\sqrt{t}\left\|\sum_{\lambda_k\leq\alpha} a_k\phi_k\right\|_{H^{\frac{d}{2}+1}} < C_1\sqrt{t}\alpha^{\frac{k+2}{4}}$$

Thus combining above and inequality 6.8 we get

$$\frac{\|R_t f_1\|}{\langle\frac{1-e^{-t\Delta}}{t}f, f\rangle} < \frac{2C_1}{\lambda_1}\sqrt{t}\alpha^{\frac{k+2}{4}} \qquad (6.13)$$

We will now bound $\frac{\|R_t f_2\|\|f_2\|}{\langle\frac{1-e^{-t\Delta}}{t}f,f\rangle}$. By applying proposition 6.7 we deduce that

$$\|R_t f_2\|\|f_2\| \leq C\|f_2\|^2$$

The denominator on the other hand satisfy

$$\begin{aligned}\left\langle\frac{1-e^{-t\Delta}}{t}f, f\right\rangle &\geq \left\langle\frac{1-e^{-t\Delta}}{t}f_2, f_2\right\rangle \qquad (6.14)\\ &\geq \frac{1}{2}\sum_{\lambda_k>\alpha} a_k^2 \min\left(\alpha, \frac{1}{\sqrt{t}}\right) \qquad (6.15)\end{aligned}$$

$$\geq \frac{1}{2} \min\left(\alpha, \frac{1}{\sqrt{t}}\right) \|f_2\|^2 \tag{6.16}$$

Which means we have

$$\frac{\|R_t f_2\| \|f_2\|}{\langle \frac{1-e^{-t\Delta}}{t} f, f\rangle} \leq \frac{C}{\min(\alpha, \frac{1}{\sqrt{t}})} \leq C \max(\frac{1}{\alpha}, \sqrt{t}) \tag{6.17}$$

Finally we use both inequalities 6.13 and 6.17 to obtain the following

$$\frac{|\langle R_t f, f\rangle|}{|\langle \frac{1-e^{-t\Delta}}{t} f, f\rangle|} \leq \frac{3\|R_t f_1\| + \|R_t f_2\| \|f_2\|}{|\langle \frac{1-e^{-t\Delta}}{t} f, f\rangle|} \tag{6.18}$$

$$\leq C\left(\sqrt{t}\alpha^{\frac{k+2}{4}} + \max\left(\frac{1}{\alpha}, \sqrt{t}\right)\right) \tag{6.19}$$

Thus setting $\alpha = t^{-\frac{2}{k+6}}$, we reach the desired result. □

Note that this the theorem 6.7 is enough to prove theorem 6.2. The proof of the theorem depended heavily on propositions 6.5 and 6.6 which we called the remainder estimates. Next section we will give proofs of remainder estimates 6.5 and 6.6.

6.1 Remainder Estimates

The goal of this section is to give qualitative proof for both propositions 6.5 and 6.6. First note the following two different kernels for setting up the notations. We call ambient Gaussian kernel as

$$G_t(p,q) = \frac{1}{(4\pi t)^{\frac{d}{2}}} e^{-\frac{\|p-q\|^2}{4t}}$$

and we denote geodesic Guassian kernel as

$$E_t(p,q) = \frac{1}{(4\pi t)^{\frac{d}{2}}} e^{-\frac{d_g(p,q)^2}{4t}}$$

furthermore we denote their associated integral operators as $\mathcal{G}_t$ and $\mathcal{E}_t$ respectively i.e

$$\mathcal{G}_t f(p) = \int_{\mathcal{M}} f(q) G_t(p,q) dV_q \quad \mathcal{E}_t f(p) = \int_{\mathcal{M}} f(p) E_t(p,q) dV_q$$

First we prove the remainder estimate with the Sobolev norm which is proposition 6.6.

Proposition ($H^{\frac{d}{2}+1}$ estimate)**.** *Let $f \in H^{\frac{d}{2}+1}(\mathcal{M})$, there exists $C \in \mathbb{R}$ such for small enough values of $t > 0$ following holds:*

$$\|R_t f\| \leq C\sqrt{t}\|f\|_{H^{\frac{d}{2}+1}}$$

Proof. Observe that

$$\begin{aligned}
R_t f(p) &= \left(\frac{1 - e^{-t\Delta_{\mathcal{M}}}}{t}\right) f(p) - \mathcal{L}_t f(p) \\
&= \frac{1}{t}\left(f(p) - \int_{\mathcal{M}} H_t(p,q) f(q) dV_q\right) \\
&\quad - \frac{1}{t}\left(f(p)\int_{\mathcal{M}} G(p,q) dV_q - \int_{\mathcal{M}} G(p,q) f(q) dV_q\right) \\
&= \frac{1}{t}\int_{\mathcal{M}} (H_t(p,q) - G_t(p,q))(f(p) - f(q)) dV_q
\end{aligned}$$

Where the last step was obtained by noting constant function is an eigenfunction of $e^{-t\Delta_{\mathcal{M}}}$, which means $\int_{\mathcal{M}} H_t(p,q) dV_q = 1$. In order to bound $R_t f$ we consider geodesic ball $B_\epsilon(p) = \{q \in \mathcal{M} \mid d_g(p,q) < \epsilon\}$ for some $\epsilon > 0$ and split the integral according as follows

$$\begin{aligned}
R_t f(p) &= \frac{1}{t}(I + II) \\
I &= \int_{B_\epsilon(p)} (H_t(p,q) - G_t(p,q))(f(p) - f(q)) dV_q \\
II &= \int_{\mathcal{M}\setminus B_\epsilon(p)} (H_t(p,q) - G_t(p,q))(f(p) - f(q)) dV_q
\end{aligned}$$

Use the exponential map exp : $T_p\mathcal{M} \to \mathcal{M}$ as studied in chapter 4, write $q = \exp(x)$. In this coordinate we have

$$I = \int_{B_\epsilon} (H_t - G_t)(f(0) - f(x))\sqrt{\det(g)}dx$$

Recall that from theorem 5.4 for small $\|x\|$ and small t, we have

$$H_t = \frac{1}{(4\pi t)^{\frac{d}{2}}} e^{-\frac{\|x\|}{4t}} (u_0 + tu_1 + t^2 u_2) + O(t^2)$$

also that $u_0 = \det^{-\frac{1}{2}}(g)$. The Taylor expansion of the determinant of metric in geodesic coordinates is given $\det(g) = 1 - \frac{1}{6}R_{ij}x^i x^j - \frac{1}{12}(\nabla_\rho R_{ij})x^\rho x^i x^j + O(\|x\|^4)$. Hence we see that for small $\|x\|$ and sufficiently small t, we have

$$H_t = E_t + O(E_t(\|x\|^2 + t) + t^2)$$

Using lemma 6.8, we know that $G_t = E_t + O(tE_{2t})$ Thus we have bound on the integrand

$$|H_t - G_t| \le C(E_t(\|x\|^2 + t) + t^2 + tE_{2t}) \tag{6.20}$$

We need to have control on $|f(0) - f(x)|$, since $f \in H^{\frac{d}{2}+1}$ by Sobolev inequality we know that the Lipshitz constant is bounded by $\gamma\|f\|_{H^{\frac{d}{2}+1}}$ where γ is some universal constant depending on geometry of $\mathcal{M}$. Hence we get

$$|f(0) - f(x)| \le \gamma\|f\|_{H^{\frac{d}{2}+1}}\|x\| \tag{6.21}$$

combining both inequalities 6.20 and 6.21 we see that

$$|I| \le C\|f\|_{H^{\frac{d}{2}+1}} \int_{B_\epsilon} \left(E_t(\|x\|^2 + t) + t^2 + tE_{2t}\right) \|x\|\sqrt{\det(g)}dx \le Ct^{\frac{3}{2}}\|f\|_{H^{\frac{d}{2}+1}}$$

In order to bound II we use Theorem 1.1 of A.Grigoryan's paper [21]. The theorem says for smooth compact manifold $\mathcal{M}$ and for $\epsilon > 0$ small enough

$$H_t(p,q) \le Ct^{\frac{3}{2}} \quad \forall p,q \in \mathcal{M} \text{ such that } \quad d_g(p,q) \ge \epsilon \tag{6.22}$$

where the constant depends on ϵ and geometry of the manifold, the same is true for $G_t(p,q)$. Hence we see that

$$|H(p,q) - G(p,q)| \leq O(t^{\frac{3}{2}}) \ \ \forall q \in \mathcal{M} \setminus B_\epsilon(p) \tag{6.23}$$

and also by the previous argument with Sobolev inequality we have $|f(p) - f(q)| = O(\|f\|_{H^{\frac{d}{2}+1}})$. Combination with 6.23 give us

$$|II| \leq Ct^{\frac{3}{2}} \|f\|_{H^{\frac{d}{2}+1}}$$

Putting these together, we obtain the result as desired

$$|R_t f(p)| = \frac{1}{t}(|I| + |II|) \leq \frac{1}{t}(Ct^{\frac{3}{2}} \|f\|_{H^{\frac{d}{2}+1}}) \leq Ct^{\frac{1}{2}} \|f\|_{H^{\frac{d}{2}+1}}$$

hence indeed we get the desired result

$$\begin{aligned} \|R_t f\|^2 &= \left(\int_{\mathcal{M}} |R_t f(p)|^2 dV_p \right) \\ &\leq C\text{Vol}(\mathcal{M}) t \|f\|^2_{H^{\frac{d}{2}+1}} \end{aligned}$$

take square root on both sides to get $\|R_t f\| \leq Ct^{\frac{1}{2}} \|f\|_{H^{\frac{d}{2}+1}}$ as needed to show. □

Above proof of the proposition 6.6 uses the following lemma.

Lemma 6.8. *For any point $p \in \mathcal{M}$ and $q \in B_\epsilon(p)$ and small enough t, we have*

$$|E_t(p,q) - G_t(p,q)| \leq CE_{2t}(p,q)$$

Proof. We use the exponential map as studied in chapter 4, recall $\exp : B_\epsilon(0) \subset T_p\mathcal{M} \to \mathcal{M}$ we set $q = \exp(x)$. Thus from the property of exponential map we get $d_g(p,q) = \|x\|$. Observe that

$$E_t(p,q) = \frac{1}{(4\pi t)^{\frac{d}{2}}} e^{\frac{-\|x\|^2}{4t}}$$

and by Lemma 7 of [4], we have $\|p-q\|^2 \geq \|x\|^2 - a\|x\|^4$ where $a > 0$. Hence we obtain

$$G_t(p,q) = \frac{1}{(4\pi t)^{\frac{d}{2}}} e^{-\frac{\|p-q\|^2}{4t}} \geq \frac{1}{(4\pi t)^{\frac{d}{2}}} e^{\frac{\|x\|^2 - a\|x\|^4}{4t}}$$

We then observe

$$|E_t(p,q) - G_t(p,q)| \leq \frac{1}{(4\pi t)^{\frac{d}{2}}} \left(e^{-\frac{\|x\|^2 - a\|x\|^4}{4t}} - e^{\frac{-\|x\|^2}{4t}} \right)$$

Thus it is enough to show the following

$$\frac{1}{(4\pi t)^{\frac{d}{2}}} \left(e^{-\frac{\|x\|^2 - a\|x\|^4}{4t}} - e^{\frac{-\|x\|^2}{4t}} \right) \leq Ct \frac{1}{(4\pi t)^{\frac{d}{2}}} \left(e^{\frac{-\|x\|^2}{8t}} \right) = CtE_{2t}(p,q)$$

Re-arrenging terms above shows that the lemma is equivalent to showing

$$e^{\frac{-\|x\|^2}{8t}} \left(e^{\frac{a\|x\|^4}{4t}} - 1 \right) \leq Ct \tag{6.24}$$

the inequality 6.24 follows from Taylor series expansion for small t. □

We have left to show the L^2 estimate of the remainder. However in order to prove proposition 6.5 we need following results. The proof of lemma 6.9 is presented in [3].

Lemma 6.9. *There is a constant C independent of t with*

$$\|\mathcal{E}_t - \mathcal{G}_t\| \leq Ct$$

where the norm is the usual L^2 operator norm i.e $\|\mathcal{E}_t\| = \sup_{\|f\|=1} \|\mathcal{E}_t f\|$.

Another useful lemma relating heat kernel and geodesic Gaussian kernel is given below.

Lemma 6.10. *For any p, q close enough and $t > 0$ sufficiently small, then the following holds*

$$|H_t(p,q) - E_t(p,q)| \leq Ct(H_{2t}(p,q) + H_t(p,q) + 1)$$

Proof. Through expansion of heat kernel [31], it is known that for p, q close enough there exists continuous functions u_0, u_1 such that

$$|H_t(p,q) - E_t(p,q)(u_0(p,q) + tu_1(p,q))| < Ct \tag{6.25}$$

hence it follows that

$$|H_t(p,q) - E_t(p,q)| \leq E_t|u_0(p,q) - 1| + tE_t(p,q)|u_1(p,q)| + Ct$$

Since $\mathcal{M}$ is compact, we have $K = \sup_{p,q\in\mathcal{M}} |u_1(p,q)| < \infty$. Thus, we have

$$|H_t(p,q) - E_t(p,q)| \leq E_t(p,q)|u_0(p,q) - 1| + KtE_t(p,q) + Ct \tag{6.26}$$

As noted in theorem 5.4, $u_0(p,q) = \frac{1}{\sqrt{\det g_{ij}(q)}}$. The asymptotic expansion in normal coordinates with $q = \exp(x)$ shows

$$\det(g_q) = 1 - \frac{1}{6}R_{ij}x^i x^j - \frac{1}{12}(\nabla_\rho R_{ij})x^\rho x^i x^j + O(\|x\|^5)$$

Thus we get $u_0 = 1 + O(\|x\|^2)$. Hence we have

$$E_t|u_0 - 1| \leq K' \frac{1}{(4\pi t)^{\frac{d}{2}}} e^{-\frac{\|x\|^2}{4t}} \|x\|^2$$

Note that if we set $z = \frac{\|x\|}{\sqrt{t}}$

$$\begin{aligned} \frac{1}{(4\pi t)^{\frac{d}{2}}} e^{-\frac{\|x\|^2}{4t}} \|x\|^2 &= t\frac{1}{(4\pi t)^{\frac{d}{2}}} e^{\frac{-z^2}{4}} z^2 \\ &\leq Ct\frac{1}{(4\pi t)^{\frac{d}{2}}} e^{\frac{-z^2}{8}} \\ &= C'tE_{2t} \end{aligned}$$

Thus we get

$$|H_t(p,q) - E_t(p,q)| \leq Ct(E_{2t} + E_t + 1) \tag{6.27}$$

since $E_t \leq CH_t$ as seen in [31] we have the desired result

$$|H_t(p,q) - E_t(p,q)| \leq Ct(H_{2t} + H_t + 1) \tag{6.28}$$

□

Following proposition is essential ingredient for the L^2 remainder estimate (proposition 6.5).

Proposition 6.11. *For $t > 0$ small enough*

$$\|e^{-t\Delta_{\mathcal{M}}} - \mathcal{E}_t\| \leq Ct$$

where C is constant depends only on the manifold

Proof. We write operator explicitly

$$\begin{aligned}(e^{-t\Delta_{\mathcal{M}}} - \mathcal{E}_t)(f)(p) &= \int_{\mathcal{M}} (H_t(p,q) - E_t(p,q))f(q)dV_q = I + II \\ I &= \int_{B_\epsilon(p)} (H_t(p,q) - E_t(p,q))f(q)dV_q \\ II &= \int_{\mathcal{M}\setminus B_\epsilon(p)} (H_t(p,q) - E_t(p,q))f(q)dV_q\end{aligned}$$

By lemma 6.10, we know that on $B_\epsilon(p)$ there exists a constant $C > 0$ such that

$$|H_t(p,q) - E_t(p,q)| \leq Ct(H_{2t} + H_t + 1)$$

Therefore we have

$$\begin{aligned}|I| &\leq \int_{B_\epsilon(p)} |H_t(p,q) - E_t(p,q)||f(q)|dV(q) && (6.29) \\ &\leq Ct\int_{B_\epsilon(p)} (H_{2t} + H_t + 1)|f(q)|dV(q) && (6.30) \\ &\leq Ct(e^{-2t\Delta_{\mathcal{M}}}(|f|) + e^{-t\Delta_{\mathcal{M}}}(|f|) + \|f\|) && (6.31) \\ &\leq Ct\|f\| && (6.32)\end{aligned}$$

We also have

$$|II| \leq \int_{\mathcal{M}\backslash B_\epsilon(p)} H_t|f(q)|dV_q + \int_{\mathcal{M}\backslash B_\epsilon(p)} E_t|f(q)|dV_q$$

when $d_g(p,q) \geq \epsilon$ then $|E_t(p,q)|$ and $|H_t(p,q)|$ can be bounded by $O(t)$ shown in [21]. Hence we have

$$\begin{aligned} |II| &\leq C\int_{\mathcal{M}\backslash B_\epsilon(p)} t|f(q)|dV_q + C'\int_{\mathcal{M}\backslash B_\epsilon(p)} t|f(q)|dV_q && (6.33)\\ &\leq Ct\int_{\mathcal{M}} |f(q)|dV_q && (6.34)\\ &\leq Ct\|f\| && (6.35)\end{aligned}$$

Finally combining both 6.29 and 6.33, we obtain

$$\begin{aligned} \|(e^{-t\Delta_{\mathcal{M}}} - \mathcal{E}_t)f\| &\leq \int_{\mathcal{M}} |I|dV_p + \int_{\mathcal{M}} |II|dV_p\\ &\leq Ct\|f\|\end{aligned}$$

the desired result follows immediately. □

Finally we are in a position to prove the L^2 remainder estimate (proposition 6.5).

Proposition. *Let $f \in L^2(\mathcal{M})$, there exists $C \in \mathbb{R}$ such that for all sufficiently small values of $t > 0$ the following holds:*

$$\|R_t f\| \leq C\|f\|$$

Proof. Recall that

$$\mathcal{L}_t f(x) = \frac{1}{t}\left(f(x)\int_M G_t(x,y)dV_y - \int_{\mathcal{M}} f(y)G_t(x,y)dV_y\right)$$

Thus we have

$$R_t f = \frac{1}{t}\left(1 - \int_{\mathcal{M}} G_t(x,y)dV_y\right) f - \frac{1}{t}(e^{-t\Delta_{\mathcal{M}}} - \mathcal{G}_t)f$$

Note the following

$$\|e^{-t\Delta_{\mathcal{M}}} - \mathcal{G}_t\| \leq \|e^{-t\Delta_{\mathcal{M}}} - \mathcal{E}_t\| + \|\mathcal{E}_t - \mathcal{G}_t\| \leq Ct \tag{6.36}$$

Where the bound on the first term came from proposition 6.11 and bound on the second one was obtained from lemma 6.9.

As for the term with Gaussian kernel we have the following bound

$$\left(1 - \int_{\mathcal{M}} G_t(x,y)dV_y\right) \leq Ct \tag{6.37}$$

as shown in [3]. Combining the inequalities 6.37 and 6.36 we indeed have

$$\|R_t f\| \leq C\|f\|$$

as desired. □

6.2 Proofs of Main Convergence Theorems

We have proven the remainder estimates which were used to show theorem 6.7. As we have discussed earlier that theorem 6.7 is enough to prove theorem 6.2.

Theorem. *6.2 Let $\lambda_i, \lambda_i^t, \phi_i, \phi_i^t$ be the i-th smallest eigenvalues and the corresponding eigenfunctions of Δ and $\mathcal{L}_t$ respectively. Then*

$$\lim_{t\to 0} |\lambda_i - \lambda_i^t| = 0$$

$$\lim_{t\to 0} \|\phi_i - \phi_i^t\| = 0$$

Proof. Note that i-th eigenvalue of $\frac{1-e^{-t\Delta}}{t}$ is equal to $\frac{1-e^{-\lambda_i}}{t}$ where λ_i is an eigenvalue of Laplace- Beltrami operator. Combining both 6.7 and 6.4 we know that

$$\lim_{t \to 0} \frac{\lambda_i^t}{\lambda_i} = 1$$

We also get convergence of eigenfunctions since eigenfunctions of $\Delta_{\mathcal{M}}$ coincide with those of $e^{-t\Delta_{\mathcal{M}}}$ and thus by proposition 6.7 eigenfunctions of $\mathcal{L}_t$ converge to eigenfunctions of $\Delta_{\mathcal{M}}$. □

Important ingredient for the final main convergence theorem is the convergence of eigenvalues and eigenfunctions of $L_{t,n}$ to that of $\mathcal{L}_t$. The proof of theorem 6.3 is provided in [3].

Theorem. *6.3 Let $\lambda_{n,i}^t$ and λ_i^t be the i-th eigenvalue of $L_{t,n}$ and $\mathcal{L}_t$ respectfully. Let $\phi_{n,i}^t$ and ϕ_i^t be the corresponding eigenfunctions. Then there exist $t > 0$ small enough such that*

$$\lim_{n \to \infty} \lambda_{n,i}^t = \lambda_i^t$$

$$\lim_{n \to \infty} \|\phi_{n,i}^t - \phi_i^t\| = 0$$

whenever $\lambda_i^t < \frac{1}{2t}$. The convergence is almost surely.

Finally we prove theorem 6.1 using both theorems 6.2 and 6.3. This establishes the convergence of eigenvalues and eigenfunctions of Laplacian Eigenmaps algorithm.

Theorem (Main Convergence)**.** *Let $\lambda_{n,i}^t$ be the i-th eigenvalue of $L_{t,n}$ and $\phi_{n,i}^t(x)$ be the corresponding eigenfunction. Let λ_i and $\phi_i(x)$ be the corresponding eigenvalue and eigenfunction of $\Delta_{\mathcal{M}}$ respectively. Then there exists a sequence $t_n \to 0$ such that*

$$\lim_{n \to \infty} \lambda_{n,i}^{t_n} = \lambda_i$$

$$\lim_{n\to\infty} \|\phi_{n,i}^{t_n}(x) - \phi_i(x)\|_2 = 0$$

where the limits are taken in probability.

Proof. We know from 6.2 and 6.3 the following convergence properties

$$\mathbf{Eig} L_{t,n} \xrightarrow{n\to\infty} \mathbf{Eig}\mathcal{L}_t \xrightarrow{t\to 0} \mathbf{Eig}\Delta_{\mathcal{M}}$$

We immediately see from 6.2 and 6.3 that for any sequence $t_n \to 0$ as $n \to \infty$

$$\lim_{n\to\infty} \lambda_{n,i}^{t_n} = \lambda_i$$

As for eigenfunctions note that for any $i \in \mathbb{N}$ and any $\epsilon > 0$ by theorem 6.2 we may choose any $\bar{t}$ small enough such that $\forall t < \bar{t}$ following holds

$$\|\phi_i - \phi_i^t\| < \frac{\epsilon}{2} \quad \forall t < \bar{t} \tag{6.38}$$

similarly by theorem 6.3, given any $i \in \mathbb{N}$ and $\epsilon > 0$ we can choose $\bar{t} < \frac{2}{\lambda_i}$ such that

$$\lim_{n\to\infty} \mathbb{P}\left\{\|\phi_{n,i}^t - \phi_i^t\| \geq \frac{\epsilon}{2}\right\} = 0 \qquad \forall t \leq \bar{t} \tag{6.39}$$

and hence from identities 6.38 and 6.39 we observe that for any $0 < t < \bar{t}$ and any probability $p > 0$ there exist $N \in \mathbb{N}$ such that $\forall n > N$ following holds

$$\begin{aligned}\mathbb{P}\{\|\phi_{n,i}^t - \phi_i\| > \epsilon\} &\leq \mathbb{P}\{\|\phi_{n,i}^t - \phi_i^t\| + \|\phi_i^t - \phi_i\| > \epsilon\} \\ &\leq \mathbb{P}\left\{\|\phi_{n,i}^t - \phi_i^t\| \geq \frac{\epsilon}{2}\right\} \\ &\leq p\end{aligned}$$

Hence inverting this relationship gives that for any N and for any probability p_N there exists corresponding $t_N > 0$ with

$$\mathbb{P}\{\|\phi_{n,i}^{t_N} - \phi_i\| > \epsilon\} < p_N \quad \forall n > N$$

We obtain the convergence in probability by taking a sequence p_N tending to zero. □

The basic convergence theorem presented in this chapter shows that at least in principle the Laplacian Eigenmaps algorithm is correct. The proof suggests ways of performing computational harmonic analysis and probabilistic approach to solving some classes of PDEs using random point samples. Natural extension of this work is to extend the current result to include convergence of differential forms in the spirit of [16]. Also if possible obtain a version of Hodge theory from the random samples, but difficulty lies in that discrete differential geometry is still undergoing development and its not clear how one should define differential forms on point clouds in a consistent way.

CHAPTER 7
Conclusion

We have investigated the problem of obtaining meaningful low dimensional representation of high dimensional data. We examined different methods of manifold learning from classical methods such as PCA and cMDS to more modern methods such as Isomap, Locally Linear Embedding and Laplacian Eigenmaps. The algorithms for these methods were presented in mathematically consistent, concise and easy to understand fashion so that it may serve as a reference for future researchers. For each algorithm we have presented motivations and justifications behind the algorithm as well as their drawbacks and advantages in terms of their complexity and theoretical guarantee.

We have also explored the relationships and similarities between the manifold learning algorithms. The methods Isomap and Laplacian Eigenmaps have not been used extensively to real life applications. We have investigated the reasoning behind this issue, one possible reason being the lack of performance measure.

Furthermore we have given short necessary background behind Laplacian on Riemannian manifold as well as heat operator on manifolds in order to prove the convergence of Laplacian Eigenmaps algorithm. Finally we present convergence of Laplacian Eigenmaps method [3] in a self contained and compact fashion following the work of Mikhail Belkin and Partha Niyogi.

References

[1] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2014.

[2] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2002.

[3] Mikhail Belkin and Partha Niyogi. Convergence of laplacian eigenmaps. In *Advances in Neural Information Processing Systems 19*, volume 129, 2007.

[4] Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. 74(8):1289–1308, 2008.

[5] R.E. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[6] Mira Bernstein, Vin De Silva, John C Langford, and Joshua B Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, 2000.

[7] I. Borg and P.J.F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2005.

[8] Matthew Brand. Charting a manifold. In *Advances in neural information processing systems*, pages 961–968, 2002.

[9] Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences of the United States of America*, 103(5):1168–1172, 2006.

[10] F.R.K. Chung. *Spectral Graph Theory*. Number no. 92 in CBMS Regional Conference Series. Conference Board of the Mathematical Sciences.

[11] Ronald R. Coifman and Stphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5 – 30, 2006. Special Issue: Diffusion Maps and Wavelets.

[12] Thomas H Cormen, Charles E Leiserson, et al. *Introduction to algorithms*, volume 2. 2001.

[13] V. de Silva and J. B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. *Advances in Neural Information Processing Systems 15.*, 2002.

[14] Vin De Silva and Joshua B Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, Technical report, Stanford University, 2004.

[15] Tamal K Dey, Pawas Ranjan, and Yusu Wang. Convergence, stability, and discrete approximation of laplace spectra. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 650–663. Society for Industrial and Applied Mathematics, 2010.

[16] Jozef Dodziuk and Vijay Kumar Patodi. Riemannian structures and triangulations of manifolds. *J. Indian Math. Soc.(NS)*, 40(1-4):1–52, 1976.

[17] David L Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

[18] Robert W Floyd. Algorithm 97: shortest path. *Communications of the ACM*, 5(6):345, 1962.

[19] Shuzhi Sam Ge, Y Yang, and Tong Heng Lee. Hand gesture recognition and tracking based on distributed locally linear embedding. *Image and Vision Computing*, 26(12):1607–1620, 2008.

[20] Paul E Green. Marketing applications of mds: Assessment and outlook. *The Journal of Marketing*, pages 24–31, 1975.

[21] Alexander Grigor'yan. Gaussian upper bounds for the heat kernel on arbitrary manifolds. *history*, 4(2exp):4t, 1997.

[22] A.J. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* Springer Texts in Statistics. Springer New York, 2009.

[23] Viren Jain and Lawrence K Saul. Exploratory analysis and visualization of speech and music by locally linear embedding. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 3, pages iii–984. IEEE, 2004.

[24] Dmitry Jakobson, Thomas Ng, Matthew Stevenson, and Mashbat Suzuki. Conformally covariant operators and conformal invariants on weighted graphs. *Geometriae Dedicata*, 174(1):339–357, 2015.

[25] Stéphane S Lafon. *Diffusion maps and geometric harmonics.* PhD thesis, Yale University, 2004.

[26] Subbaramiah Minakshisundaram. Eigenfunctions on riemannian manifolds. *J. Indian Math. Soc*, 17:159–165, 1953.

[27] Fionn Murtagh, Jean-Luc Starck, and Michael W Berry. Overcoming the curse of dimensionality in clustering by means of the wavelet transform. *The Computer Journal*, 43(2):107–120, 2000.

[28] Boaz Nadler, Stephane Lafon, Ronald R Coifman, and Ioannis G Kevrekidis. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. *arXiv preprint math/0506090*, 2005.

[29] Boaz Nadler, Stéphane Lafon, Ronald R Coifman, and Ioannis G Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

[30] Robert Pless. Image spaces and video trajectories: Using isomap to explore video sequences.

[31] Steven Rosenberg. *The Laplacian on a Riemannian manifold : an introduction to analysis on manifolds*. London Mathematical Society student texts. Cambridge University Press, Cambridge, U.K., New York, NY, USA, 1997.

[32] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[33] Lawrence K Saul and Sam T Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *The Journal of Machine Learning Research*, 4:119–155, 2003.

[34] Yoshio Takane. Applications of multidimensional scaling in psychometrics. 2006.

[35] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319, 2000.

[36] W.S. Torgerson. *Theory and methods of scaling*. Wiley, 1958.

[37] Kilian Q Weinberger, Fei Sha, and Lawrence K Saul. Learning a kernel matrix for nonlinear dimensionality reduction. In *Proceedings of the twenty-first international conference on Machine learning*, page 106. ACM, 2004.

[38] Ming-Hsuan Yang. Face recognition using extended isomap. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 2, pages II–117. IEEE, 2002.

[39] Junping Zhang, Huanxing Shen, and Zhi-Hua Zhou. Unified locally linear embedding and linear discriminant analysis algorithm (ullelda) for face recognition. In *Advances in Biometric Person Authentication*, pages 296–304. Springer, 2005.

[40] Tianhao Zhang, Jie Yang, Deli Zhao, and Xinliang Ge. Linear local tangent space alignment and application to face recognition. *Neurocomputing*, 70(7):1547–1553, 2007.

[41] Zhenyue Zhang, Jing Wang, and Hongyuan Zha. Adaptive manifold learning. *IEEE transactions on pattern analysis and machine intelligence*, 34(2):253–265, 2012.